TUM School of Computation, Information and Technology

Understanding the Legitimacy of Digital Socio-Technical Classification Systems

Severin Karl David Engelmann

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:	apl Prof. Dr. Georg Groh
Prüfer*innen der Dissertation:	1. Prof. Dr. Jens Grossklags
	2. Prof. Dr. Bettina Berendt
	3. Prof. Dr. Malte Ziewitz

Die Dissertation wurde am 28.09.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 15.06.2023 angenommen.

Acknowledgments

I am deeply grateful to my supervisor Prof. Jens Grossklags, Chair of the Professorship of Cyber Trust at the Technical University of Munich. I am very thankful to you for introducing me to science as a profession and for inviting me to discover the many different facets of what it means to be a scientist. Thank you for allowing me on board and for giving me the opportunity and support to pursue my doctorate in your research group. I am thankful for the countless discussions on our research ideas and for the many rounds of invaluable feedback you gave me. Your curiosity inspired me to think deeper and your encouragement motivated me to keep exploring. Thank you!

I am also thankful to Prof. Bettina Berendt, Director of the Weizenbaum Institute for the Networked Society - The German Internet Institute, Berlin. I am grateful for our interactions at several conferences, particularly, at the Privacy Forum Barcelona in 2018. I am equally thankful to Prof. Malte Ziewitz at Cornell University for giving me guidance on my initial research ideas at the 2019 Interdisciplinary Summer School on Privacy in Nijmegen, The Netherlands. To both of you: Thank you for giving me the sense of being part of a research community when I had just started my doctorate.

It takes a village to raise a Ph.D. and there are several collaborators and supporters that I would like to express my gratitude to. I am very grateful to Dr. Mo Chen. It has been such a pleasure working together with you in the past years and it has been great sharing an office with you. Likewise, thank you, Chiara Ullstein. It has been a wonderful journey developing and realizing our research on facial analysis AI together. Thank you, Dr. Orestis Papakyriakopoulos, for our cooperation on several projects, I have learnt a lot from you.

Thank you, Tamara Kastenmeier for all the support and help with the many different administrative hurdles throughout my doctorate.

I am very grateful to Prof. Deirdre Mulligan of the School of Information at the University of California, Berkeley, for hosting me as a visiting scholar in 2022. Thank you for an incredibly valuable experience, in particular, for our discussions on essentially contested concepts. Thank you for introducing me to the brilliant scholars in your group, I learnt a lot during my stay. Thank you Mariel, Halyna, and Sabrina, you made my visit fabulous.

None of this would have been possible without the support of my beloved family. To my mother Christiana and my father Bernd: Thank you for everything! I am deeply grateful to my brother Jan who has supported me on so many different levels along this journey. Thank

you! I would also like to thank all of my closest friends for giving me the necessary comfort and energy throughout this time. Thank you, Yassin, Bob, Henning, Ebru, Moritz, Selina, Niels, and Mistale.

Abstract

This doctoral thesis advances our understanding of the legitimacy of different digital socio-technical classification systems.

First, taking a *system-level perspective*, I introduce research on the implementation and legitimization of the Chinese social credit system (SCS, 社会信用体系 or shehui xinyong tixi). Based on a unique data set of reputational blacklists and redlists in 30 Chinese provincial-level administrative divisions (ADs), I present the first comprehensive empirical study of digital blacklists (classifying "bad" behavior) and redlists (classifying "good" behavior) in the Chinese SCS. An analysis of SCS role-model narratives demonstrates that the SCS adopts virtue ethical principles around honesty and dishonesty to legitimate one of the largest state-run digital classification systems in the world.

Second, in our work on social media profiling I investigate *procedural normative choices* in social media classification. Social media platforms enable advertisers to create and target user audiences based on the classification of several thousand user attributes such as likes, interests, beliefs, behaviors, relationships, moral convictions, and political leanings. I define such procedural normative choices based on an extensive engagement with theories of personal identity in philosophy. I then present an empirical study that explores how social media users evaluate social media's classifications with respect to their accuracy and transparency. While most studies have paid attention to the *consequences* of social media classifications, this research deepens our understanding of their *procedural* legitimacy.

Third, I present our research on the legitimacy of facial analysis AI classifications. The vast abundance of visual data with recent technological developments in computer vision AI have raised concerns about the kinds of conclusions AI should make about people based on their facial appearance. Some scholars speak of supposedly "common sense" facial inferences. Others see the return of an automated version of "physiognomic practices". Using the transformer-based language model roBERTa, our study analyzes participants' nearly 30.000 written justifications of specific facial analysis classifications. One key finding is that people legitimize visual classifications by both epistemic *and* pragmatic considerations. I argue that pragmatic considerations represent a "legitimacy pitfall". In a follow-up study, we investigate how people with AI-competence evaluate facial analysis AI. Overall, participants' reflections underline the normative complexity behind facial analysis AI classifications.

Finally, I argue that a comprehensive understanding of the legitimacy of digital socio-technical classification systems critically requires an in-depth engagement with essentially contested concepts.

Publications

Thesis Publications

Ordered by date of publication:

- Ullstein*, C., **Engelmann***, **S.**, Papakyriakopoulos, O., Hohendanner, M., & Grossklags, J. (2022). AI-competent individuals and laypeople tend to oppose facial analysis AI. *Proceedings of the Second ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*.

- **Engelmann, S.**, Scheibe, V., Battaglia, F., & Grossklags, J. (2022). Social media profiling continues to partake in the development of formalistic self-concepts. Social media users think so, too. *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- Chen, M., **Engelmann, S.**, & Grossklags, J. (2022). Ordinary people as moral heroes and foes: Digital role model narratives propagate social norms in China's Social Credit System. *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- **Engelmann***, **S.**, Ullstein*, C., Papakyriakopoulos, O., & Grossklags, J. (2022). What People Think AI Should Infer from Faces. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

- **Engelmann, S.**, Chen, M., Dang, L., & Grossklags, J. (2021). Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness. *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- **Engelmann, S.**, & Grossklags, J. (2019). Setting the Stage: Towards Principles for Reasonable Image Inferences. *Workshop on Fairness in User Modeling, Adaptation and Personalization (FairUMAP), 27th Conference on User Modeling, Adaptation and Personalization (ACM UMAP)*.

- **Engelmann***, **S.**, Chen*, M., Fischer, F., Kao, C. & Grossklags, J. (2019). Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

Other Publications

Ordered by date of publication:

- Cypris, N., **Engelmann, S.**, Sasse, J., Grossklags, J., & Baumert, A. (2022). Intervening Against Online Hate Speech: A Case for Automated Counterspeech. *IEAI Research Brief*, Technical University of Munich.
- **Engelmann, S.**, Grossklags, J., & Herzog, L. (2020). Should users participate in governing social media? Philosophical and technical considerations of democratic social media. *First Monday*.
- **Engelmann, S.**, Grossklags, J., & Papakyriakopoulos, O. (2018). A Democracy called Facebook? Participation as a Privacy Strategy on Social Media. *Proceedings of the Annual Privacy Forum 2018. Lecture Notes in Computer Sciences (LNCS)*.

*Denotes equal contribution.

Contents

1 Introduction

1.1 Classification as a fundamental human activity and as a driver of social progress

We live in a world of ubiquitous classification. Classification is the fundamental human activity of making an otherwise semantically ambiguous world legible, actionable and improvable [1, 2]. In today's modern age, classification and, in particular, digital classification has entered nearly all spheres of life. For example, classification is the key practice of the scientific paradigm of the modern era [3]. Defining objects and assigning them higher-order classifications are essential procedures of scientific activity. Much of science means arguing for and against the boundaries of classifying phenomena. Today, across many scientific fields, but in particular in the natural sciences, progress is intimately tied to *technological progress* that facilitates more fine-grained and efficient grouping and clustering of related phenomena. This forms the basis for a better representation, better explanation, and, eventually, a better prediction of the studied phenomena.

Up until the mid-nineteenth century, systematic classification had been reserved only for natural objects. Then, a radical social transformation occurred: society itself was conceptualized as an object of *scientific description by classification* and of *scientific analysis by statistical methodologies* [2]. Profiling of age distributions, literacy and crime rates, medical records, or property ownership documentation promised to enable accurate representations of social phenomena similar to the detailed descriptions of forests, agricultural spaces, and other classifications of the natural sciences [4, 5]. This conceptual transfer and application onto society resulted in enormous benefits and social progress [6, 3]. First, the ability to produce classifications of social welfare meant that such issues could be acted on and improved under the authority and legitimacy of scientific objectivity. Second, states benefited, among others, because a more healthy population resulted in more economic fitness and therefore higher taxes. This period of "high modernism" [2] laid the foundation to engineer society in desirable ways and to view it as a project of "nation-building" [5] by systematic classification.

The hidden power engines of modern states are technological infrastructures of classification that underlie their administrative and bureaucratic practices, legal formalisms, and economic activities. At the beginning of the twentieth century, Max Weber famously defined the characterizing principle of "Western rationalization" as the "mastery of all things by calculation" [7]. This mastery critically presupposes a shared system of units, standards, and metrics that powers classification and "serves to master fragmented and disconnected realities" while also creating "regularities of action" [8]. Classification systems form the basis

of a shared reality that is necessary for human cooperation. It is a modern notion that kinds, things, phenomena, documents and so on can be classified according to their fundamental essences.

1.2 The rise of *digital* classification systems across nations, political systems, and cultures

The classification of objects, people, relationships, activities and so on using large-scale digital systems has become one of the essential identities of life in the twenty-first century. In many societies around the world, classification systems have become digital. Accordingly, their governments – whether democratic or authoritarian – attempt to implement ambitious policy goals to make progress toward a "digital society" that critically depends on such digital classification systems. In a diverse set of nations, political systems, and cultures – take the United States, Germany, China, South Korea or Kenya (and many others) – social media platforms, socio-technical credit systems, search engines, or digital medical systems automatically classify objects, people, activities, and other social phenomena. Increasingly, policy plans and objectives are formulated in such a way that they can be implemented using digital classification systems [9, 10].

Novel digital classification systems promise social progress and are commonly met with enthusiasm. Such promises are far from unfounded and digital classification systems have clearly improved access to informational resources and to social networks. For example, social media platforms have been found to be of enormous benefit for users. They offer social connectivity and exchange [11], establishment and maintenance of social capital [12], as well as public and semi-public identity representations that are performative, liberating and, in particular, entertaining [13]. Social media platforms enable users to create standardized profiles that allow them to engage with other peers to form online companionship around interests that may not be shared with individuals within their vicinity. Platform operators automatically classify information conveyed through such profiles to facilitate the delivery of personalized advertisements by social media marketers. Social media platforms represent a dual-architecture classification system with standardized user profiles and the classification of users' identities for commercial purposes [14]. With their dual-architecture classification, social media classification systems anchor identity declarations of users around semantic affordances they have set up in the user interface and hence classify users around such semantics for commercial purposes [15]. Digital classification systems successfully implement some of the most profitable business models of our time as illustrated by the enormous economic power of social media platforms and search engines [16, 17, 18, 19, 20].

Search engines classify informational resources for users according to *their* classified search histories, demographics, or interests [21]. Such personalization or recommender systems solve one of the major challenges of the big data age: They make information retrieval usable in the first place as they pre-select content relevant to an individual user who would otherwise have to maneuver a sea of unordered and chaotic information [22].

Another prominent example of digital classification systems is digital credit scoring. Credit scoring classifies individuals that seek to borrow financial credit according to their predicted risk of defaulting on a loan [23]. "Traditionally", such predictions were made exclusively based on information deemed directly relevant for financial creditworthiness including sociodemographic data, previous defaulting, savings, and financial assets. Digital credit scoring, on the other hand, classifies borrowers into "creditworthy" or "not creditworthy" by incorporating non-financial information, in particular, digital footprints to run predictive models [24]. Digital footprints can be the device type and operating system, information taken from social media platforms as well as information "left behind" when visiting other websites [25]. Research studies have suggested that easily available digital footprints such as device type, operating system, or email host can match or even exceed the information value of traditional credit bureau scores. Such digital credit scoring facilitates the creditworthiness assessment of borrowers without any previous financial history [23]. Indeed, one reason for the support of digital credit scoring systems is that they supposedly minimize the transaction costs between borrower and lender, as well as to increase allocative efficiency, accuracy and distributive fairness in the loan application procedure [26]. The key promise of such alternative classification of financial creditworthiness lies in their inclusiveness. Digital credit scoring systems purportedly enable access to financial services for an estimated 2.5 billion individuals that are "unbanked", which means that they do not have any financial history or documentation [24]. Moreover, they are expected to make it easier for low-income borrowers and micro-enterprises to apply for financial loans. These are all advantages that traditional credit scoring does not offer.

1.3 Understanding the legitimacy of large digital socio-technical systems

Despite these advances and further promises of progress, there have been growing concerns regarding the legitimacy of a plethora of digital classification systems. For example, the informational asymmetry between data controllers and data subjects has been a breeding ground for numerous different privacy scandals such as the Cambridge Analytica Scandal [27], the Equifax Scandal [28] or the NSA files [29]. Narratives around the development of AI that is increasingly capable of performing human cognition tasks such as playing Go [30], recognizing human emotions from faces [31], engaging in human-like conversation [32] go hand in hand with increasing fears of human replacement by automated decision-making systems [33]. There are substantiated challenges regarding the automation and amplification of social biases through AI-based classification systems [34, 35].

The aim of this research thesis is to advance our understanding of the legitimacy of different digital socio-technical classification systems. Producing a uniform, comprehensive and conclusive understanding of the legitimacy of any large-scale digital classification system is a difficult task.

To account for the multidimensional nature of digital classification, my research approach on the legitimacy of digital classification systems consists of a *system-level perspective*, a perspective on *classification procedures*, and a perspective on *specific classifications.*

First, from a *system-level perspective*, large-scale digital classification systems depend on a multifaceted technical infrastructure that operates dynamically — responding to shifts in data input, for example — and at high speed. For many classification systems, such as search engines, their technological infrastructure is spread around the globe and therefore cannot be exactly pinpointed territorially. The scope of operation of large-scale digital classification systems covers millions and, in many cases, billions of different people across different states with different forms of governance and policy-making as well as nations with different cultures, communities, traditions and social norms.

Moreover, large-scale digital classification systems are, first and foremost, developed according to the incentives of governmental or commercial interests and, albeit in different forms, respond to the pressures of particular governmental and economic environments with various legal rights and obligations. A comprehensive and conclusive understanding of the legitimacy of a digital classification system necessarily needs to account for all of these different dimensions.

Second, the *process of classifying an object* is an inherently normative undertaking. What quality of evidence justifies a classification and how much evidence is needed to justify such a classification? Social media platforms, for example, classify social phenomena such as interests, social relationships, political leanings by essentially defining all *procedural elements* that govern the assignment of a semantic declaration to an object. Understanding such procedural elements of classification in digital systems is complicated by the fact that they usually operate under conditions of opaqueness. In many cases, their lack of transparency impedes an investigation and understanding of the epistemic quality of such classifications.

Early within the development of the system, designers negotiate what should constitute the essence of the phenomena that the system should classify. Once a system is up and running, preceding negotiations about the meaning of inherently ambiguous concepts move towards the background and their plurality and contextuality is typically forgotten [1, 2]. It tends to be difficult to go back and re-negotiate the essence of classifications, or at least re-examine the assumptions that underlie a system's procedural classification practices and characteristics. Bowker & Star have argued that the lack of transparency of large classification systems directly serves the *naturalization* of the system's classifications [1]. Classification systems thereby naturalize their own definitions of inherently vague phenomena once the system operationalizes these definitions under conditions of opaqueness. A system's classifications become natural, they lose their "anthropological strangeness" [1].

Third and finally, classifications are never simply given. They always take place in a cultural meaning giving structure and depend on the defining entity's beliefs and goals on what should constitute the classification of a phenomenon. Given their "situatedness", classifications are never value-free or value-neutral [36]. A particularly illustrating example

is the classification of contested social phenomena in social media platforms: relationships between people, interests of a particular individual, or moral and political convictions are, first and foremost, inherently underdetermined phenomena. Other examples abound. What factors determine whether a person can be classified as sufficiently trustworthy to justify the allocation of a loan? What qualifies a person for employment? Answering such questions presupposes a normative judgment. This is why in modern societies that depend on classification systems, those that have the (often technological) resources to classify people, objects, relationships, and activities are typically those that exercise power over society [2]. Digital socio-technical systems create, fixate, and operationalize a particular definition of otherwise semantically vague social phenomena. Digital socio-technical systems determine the fundamental meaning of inherently ambiguous concepts and they often do so for a global society.

In conclusion, this doctoral thesis takes on these three perspectives – a system-level, a procedural, and a classification-level perspective – to produce a more comprehensive understanding of the legitimacy of different large-scale classification systems. A summary of the main research questions, together with an outline of the research agenda, is provided in the next section.

1.4 Overall research agenda and guiding research questions

Overall research agenda:

The aim of this research thesis is to advance our understanding of the legitimacy of different digital socio-technical classification systems.

This doctoral dissertation applies a multi-methodological approach to understand the legitimacy of different digital socio-technical systems. For this purpose, it offers three different perspectives:

- First, a *system-level perspective* on the Chinese Social Credit System (SCS).
- Second, a perspective on the legitimacy of *classification procedures* in social media platforms.
- Third, a perspective on the legitimacy of *specific classifications* in facial analysis AI.

1.5 *Research Contribution 1: System-level analysis*

On the legitimacy of the Chinese Social Credit System (SCS)

Taking a system-level perspective, the first part of the doctoral thesis explores how the Chinese government implements and justifies the construction of a nation-wide digital social credit system with the aim to classify citizens, companies, and other organizations into "good" and "bad" categories via publicly accessible digital platforms. Here, so-called redlists showcase entities that have complied with social and legal norms while blacklists display those entities that have not complied with such norms. In the Chinese SCS, "good" behavior can result in material rewards and reputational gain while "bad" behavior can lead to the exclusion from material resources and reputational loss.

To understand the legitimacy of the Chinese SCS, we first investigated part of its core technical implementation: publicly accessible blacklists and redlists.

Given that China provides only restricted access to its digital platforms from outside China, very little is known about the actual *implementation* of the SCS. This fact is troubling since the Chinese SCS is currently the largest state-run digital social credit system in the world. It covers all Chinese citizens, Chinese businesses as well as all foreign businesses operating in China (among others). As we discuss in our research papers, the Chinese SCS has received

significant media coverage with varying information on the actual implementation of the system. We set out to explore one of the core technical implementations of the Chinese SCS: its nationwide redlists and blacklist infrastructure where entities are classified into "good" (redlist) and "bad" (blacklist) categories. Our empirical research on the Chinese SCS first focuses on the system's *implementation* and then on its *legitimization*.

Research Study 1:

RQ 1: Are there different degrees of transparency in blacklist and redlist records in the Beijing listing infrastructure?

We conducted a first empirical study on blacklists and redlist implementation in the municipality of Beijing at the end of 2018 (see Research Article 1 in chapter 3.1). We collected and analyzed the different types of blacklists and redlists to understand what sanctions and rewards they displayed and enforced, respectively. Moreover, we collected close to two hundred thousand blacklist and redlist records to investigate the level of explanation they provided as to what caused a particular entity to be placed on either blacklists or redlists.

Research Study 2:

RQ 2: How diverse, comprehensive, and flexible is the blacklist and redlist infrastructure across China?

In a second study on China's blacklists and redlists, we extended our analysis to the implementation of the listing infrastructure across 30 Chinese provincial-level administrative divisions (see Research Article 2 in chapter 3.2). This work focused on the the diversity, flexibility, and comprehensiveness of the nationwide listing infrastructure as of 2021. Specifically, this study aimed to provide an in-depth understanding on the types of classifications represented in the lists and their credit records, the information credit records contained, and the flexibility by which novel types of lists could be set up during the Coronavirus pandemic.

Research Study 3:

RQ 3: How does the Chinese government justify and legitimize the development of the SCS?

In a third study, we conducted an in-depth analysis of role model narratives published on the national SCS platform creditchina.gov.cn (see Research Article 3 in chapter 3.3). This research demonstrates how the Chinese government uses role model narratives on digital communication channels to advertise the SCS as a solution to many of society's ills as well as to inform the Chinese public about "good" and "bad" classifications. The use of such role model narratives is particularly interesting when viewed from an ethics perspective as Chinese ethics has had a long tradition of social norm propagation through reader-friendly narratives and stories on "good" and "bad" behaviors.

1.5.1 Summary of research contributions

The analysis of both the SCS listing infrastructure as well as the SCS role model narratives reveals how the Chinese government couples vague ethical principles, in particular, virtues of honesty and vices of dishonesty, with policy-making through a digital socio-technical classification system. Thus, our analysis of key SCS platforms provides a much more profound understanding of how the Chinese government justifies, motivates, and promotes the SCS to society. I will elaborate and clarify the implications of our observations in the discussion section.

I believe that the core contribution of our research on the SCS is to illustrate how an authoritarian system successfully legitimates the development and implementation of a nationwide digital classification system based on the ideals reflected in the virtue-ethical principles of Confucianism.

1.6 *Research Contribution 2: Procedural normativity of classifications*

On the procedural legitimacy of social media classifications

This second perspective will present work on the legitimacy of procedural classifications in social media platforms. This necessarily means a transition to an analysis of *commercial* digital socio-technical systems.

To understand the procedural legitimacy of social media classifications, we apply a part philosophical, part empirical methodology.

Social media platforms are among the most technologically advanced digital classification systems. They classify users into thousands of categories in user profiling procedures. These classifications are sold to advertisers to show users personalized advertisements. While previous classification systems typically stood outside looking in, social media platforms are in themselves "classification markets" that are able to "classify from within" [37]. Many research papers have focused on the *consequences* of social media classification. For example, research has shown how social media classifications can lead to discriminatory distribution of advertisement [38], political polarization [39], and amplification of hate speech [40]. This doctoral thesis focuses on social media classifications with regard to their *procedural* legitimacy.

This part philosophical, part empirical research investigates distinct procedural normative choices in social media classification for audience targeting.

Research Study:

RQ 1: What are key procedural normative trade-offs in social media user profiling?

We first analyzed theories of personal identity in philosophy to understand possible normative trade-offs in social media classifications (see Research Article 1 chapter 4.1). Philosophical theories of personal identity provided a fundamental perspective on the core normative challenges of social media user profiling. They allowed us to formulate two normative trade-offs inherent to the classification procedures of user profiling. We called the first normative trade-off the "accuracy vs. privacy" trade-off. If it were normative for social media user profiling to represent a person's self-concept as accurately as possible then this would result in significant privacy implications. The second normative trade-off is called the "transparency vs. autonomy" trade-off. Here, if social media classifications were transparent to users then this could result in a decrease of autonomy because transparency would enable social media user classifications to influence a person's self-concept.

RQ 2: How do social media users evaluate such normative trade-offs?

Zooming in on the procedural challenges of social media profiling lays a foundation to design an empirical vignette study to explore how social media users evaluate such normative trade-offs. Accordingly, we conducted an empirical vignette study to understand how individuals evaluate social media's identity claims with regard to accuracy, transparency, and control. The goal of the vignette study was to take a tangible step towards understanding whether social media users preferred accuracy of social media identity declarations over privacy (trade-off 1) and whether they believed that social media identity declarations would influence their self-concept (trade-off 2).

RQ 3: Do social media users believe that social media classifications can represent parts of their self-concept?

Moreover, our vignette study explored how social media user related to social media classifications. The vignette asked social media users whether they believed social media profiling could accurately infer elements of their self-concept, whether they considered accuracy of these identity declarations to be desirable, whether they had motivation to view and correct identity declarations, and whether they believed that social media identity declarations would influence their self-concept if they were made transparent to them.

1.6.1 Summary of research contributions

Our conceptual analysis of theories of personal identity in philosophy finds that philosophers generally agree that individuals have the capacity to *justify* and *control* essential elements of their self-concept. We argue that social media user profiling generates formalistic self-concepts when it determines the meaning of views, clicks, posts, relationships, or location data of social media users. The procedural ability to create formalistic self-concepts makes social media platforms powerful classification systems. Moreover, they do not offer any means for justification and control over formalistic self-concepts as philosophical theories of personal

identity suggest.

The process of generating formalistic self-concept is an inherently normative process. For example, platforms decide how much evidence is sufficient for determining that a user has a particular interest. They also decide whether the quality of the data is sufficient to justifiably make an inference about a user.

We find that people believe that social media can make accurate judgements about them but that they cannot represent their entire self-concept. Respondents thought that social media profiling is able to accurately infer whether they have changed as a person over time, but that it cannot tell an accurate story of their life. Respondents showed a strong preference for more transparency and stated that they would compare their own self-concept with a variety of social media identity declarations. We take it that social media users have some motivation to control essential aspects of their social media identity declarations. Finally, respondents strongly objected that viewing social media identity declarations would cause them to reevaluate their self-concept.

I believe that the core contribution of our research on the procedural legitimacy of social media classifications is to show how theories of personal identity can bring to the surface ethical challenges of social media classification that are independent of the *consequences* of social media classification. Together with the results of our vignette study, I believe this theoretical framework can meaningfully inform alternate platform designs that help individuals better negotiate and contest algorithmically constructed self-concepts.

1.7 *Research Contribution 3: The normativity of specific classifications*

On the legitimacy of facial analysis AI classifications

In a final research project, the thesis presents work that explores the legitimization of a specific type of AI classification: facial analysis AI. In developing computer vision AI, human faces are currently the most frequently occurring "object of analysis" in computer vision AI training sets [41].

To understand the legitimacy of facial analysis AI classifications, we apply computational, quantitative, and philosophical methodologies to compare ethical evaluations of non-experts with the evaluations of people with AI-competence.

Every day, more than 2 billion images are uploaded just across the Facebook services[1]. We live in an increasingly *visual data culture* and there has been an enormous push to develop computer vision AI that analyzes this sheer infinite amount of visual data. Online, visual data are popular vehicles to showcase an intelligible self-concept. Indeed, "showing rather

[1] `https://bit.ly/3QTf9R2`

than telling" has become the most common self-presentation strategy among users on social media platforms.

Moreover, faces play an enormously important role in human social interaction and, consequently, in human social psychology. When humans encounter each other for the first time, they make a variety of judgments about each other based on facial looks. In many cultures around the world, faces are taken to be "a window to a person's soul" and people can rapidly make judgments about a person's *apparent* trustworthiness or likability based on facial looks. Research in psychology has demonstrated that humans are prone to evaluate facial information to make consequential judgments across various different decision scenarios. For example, first facial impressions can determine hiring choices [42] or election outcomes [43].

Facial expressions are taken to be reliable indicators of emotional sensations. There has been a plethora of developments in facial emotion recognition AI for a variety of domains such as automated hiring, digital marketing, or surveillance of digital examinations during the coronavirus pandemic. As we discuss in our research papers, psychological studies on first impressions have produced overwhelming evidence that first facial impressions are largely inaccurate (e.g., [44]), however, another body of literature presents evidence that proposes such impressions to have some accuracy rendering them not entirely invalid (e.g., [45]).

Among the different semantics that visual data can contain, AI analysis has focused particularly on human faces [41]. This begs the question what kind of classifications computer vision AI should and should not perform about people based on their faces. This question is not just important in the context of image or video analysis that contains faces. Embodied, humanlike, social robotics that interact with humans already operate AI systems that recognize and classify human facial expressions [46]. What kind of inferences should such social robots draw from human faces? More generally, there is a growing need to argue for ethically justifiable automated facial inferences for a growing number of human-computer interactions. How should we demarcate permissible from impermissible facial inferences in these different interaction contexts? What conceptual basis should we apply in making an argument for or against specific facial inferences drawn by AI?

Research Study 1:

RQ 1: Based on an empiricist notion of reasonableness, what are initial principles for reasonable and unreasonable facial analysis inferences?

We "set the stage" for such an inquiry in a first conceptual attempt to define reasonable inferences based on an empiricist notion of reasonableness (see Research Article 1 in chapter 5.1). The core contribution of this work was to demonstrate that what may first appear to be a purely epistemic question – "*What does a face look like?*" – turns out to be a profound ethical challenge.

Research Study 2:

RQ 2: How do non-experts in AI ethically justify facial analysis AI?

In a second research article, this thesis presents a study on how non-experts in AI ethically evaluate specific AI facial inferences (see Research Article 2 in chapter 5.2). A growing body of literature has offered conceptual criticism of facial analysis AI with references to the historic projects of physiognomy and phrenology. We wanted to understand how non-experts justify what they believe differentiates permissible from impermissible facial AI inferences. This way, we hoped not just to allow non-experts to participate in the debate on ethical facial analysis AI. Centering our study on non-experts' written justifications of specific AI inferences helped us to identify potential justification pitfalls that support the legitimization of what appear to be physiognomic AI inferences.

Research Study 3:

RQ 3: Do ethical justifications of facial analysis AI differ between non-experts and people with AI competence?

In a final research study, we complemented our first investigation on non-experts' ethical evaluations on facial analysis with a study on the ethical evaluations of people with AI-competence (see Research Article 3 in chapter 5.3). To overcome the weaknesses of self-reported AI-competence, we designed an AI-quiz to create a sample with different levels of AI-competence.

Taken together, the two empirical studies' goals were threefold: first, to understand how non-experts evaluate specific facial analysis AI inferences across two decision contexts that vary with regard to their consequentiality. Second, the analysis of a large corpus of written justifications allowed us to explore the different types of justifications non-experts use and how they change when the decision context changes. This provided a firm ground to further elaborate on the quality of reasoning of non-experts when justifying specific AI inferences. Third and finally, we compared non-experts' evaluations to those of people with AI-competence.

1.7.1 Summary of research contributions

Overall, this final research project conceptualizes a notion of fair AI classification (or fair AI inference-making) that centers around the epistemic and pragmatic justifications of specific AI inferences. In our vignette studies, we combine qualitative and computational methods to understand the argumentative reasoning behind nearly 30.000 written justifications for eight facial analysis inferences. Our research underlines the normative complexity of facial inference-making.

1.8 Understanding the legitimacy of digital socio-technical systems: The nascent field of AI ethics

The following part of the thesis provides a concise outline of the key practices to understand the legitimacy of digital socio-technical classification systems summarized under the umbrella term "AI ethics". This *overview* consists primarily of "technical fixes of bias in AI" and "principlism in AI ethics". In the nascent field of AI ethics, these two approaches have been among the most dominant strategies to identify and solve challenges of legitimacy for novel digital classification systems. I do not, however, suggest that an analysis of the legitimacy of digital classification systems is exclusively tied to the concepts, methodologies, and communities associated with the field of AI ethics.

The concept of "legitimacy" itself is an essentially contested concept: debates about its central meaning or essence are central to the concept itself [47]. Consequently, other discipline-specific approaches offer valuable and indispensable scientific tools, both conceptual and methodological, to advance the study on the legitimacy of digital classification. Aspiring to provide a comprehensive and conclusive understanding on the matter from a single disciplinary perspective cannot account for the diverse and multidimensional nature of digital classification systems, as I described in section 1.3. To reiterate, corresponding to the multidimensional nature of digital classification systems, there are numerous different efforts to understand and verify their legitimacy. A digital classification system may operate legally after a legal compliance check, it may be legitimate with respect to its safety through different tools of security verification, it may guarantee privacy after a privacy assessment, and may be free of bias after a bias audit and so on.

In the following, I engage in a short discussion on "AI ethics" primarily because of my background in the philosophy of technology and computer science – a disciplinary combination that provides specific conceptual and methodological tools to study the legitimacy of digital classification. In the discussion of the thesis (chapter 6), I elaborate on potential weaknesses of current AI ethics approaches (i.e., "technical fixes" and "principlism") based on the findings of our research. Foreshadowing this discussion, I will first argue that the virtue ethics principles of Chinese ethics offers enough interpretative space to support the legitimization of the Chinese SCS within an authoritarian system (Research Contribution 1). Second, technical fixes help mitigate the consequences of biased classification systems, but they do not account for the epistemic and pragmatic normativity of classification procedures (Research Contribution 2) and specific classifications (Research Contribution 3).

Debiasing as a computational method to legitimate AI-based digital classification systems

AI classification systems have been developed and deployed to make decisions in hiring, advertising, or credit lending. In many of such classification scenarios, researchers have noted that AI predictions can result in unfair outcomes along social axes such as race or gender [9, 35]. AI classification systems can lead to unfair treatment when they classify individuals,

for example, who should and should not get a loan, based on datasets that have ingrained in them the structural inequalities present in society [48]. The unfair treatment based on membership of a protected class is unlawful in most, if not all, constitutions in the Western world [49].

Beginning in the early 2010s, a research community formed that started developing computational tools to identify, mitigate, and resolve bias in AI classification system [35]. This line of research has not just led to the successful production of computational tools that effectively make AI fairer but also managed to generate enormous attention of the ethical challenges of AI-based predictions in science and the general public.

A historical excursion into the debiasing of digital systems

One of the first research papers that conceptualized "bias" in relation to "computer systems" was published in 1996 by Friedman and Nissenbaum [50]. While they analyze three types of biases, preexisting (social bias), technical (bias due to technical constraint), and emergent bias (bias that arises from the decision context), they define bias in computer systems as ..."computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others." [50]. For example, the authors warn of "systematic bias" as a result of computer systems replicating social inequalities. They offer several case studies to show how the design of algorithms can inevitably lead to biased decision-making favoring one group over another. A key difference to today's research on bias in AI is that their paper was set in a time predating big data. Hence, there were no equivalent analyses around data analytics or data mining models. Yet, many of the key ideas in this paper laid the path towards the development of two key AI ethics publication communities today, the ACM Conference on Fairness, Accountability, and Transparency (FAccT)[2] and the AAAI/ACM Conference on Artificial, Intelligence, Ethics, and Society (AIES)[3].

In 2008, also considered a pioneering study today, Pedreschi et al. [51] demonstrated a) that AI classification can systematically discriminate across membership of a protected class (ethnicity, gender, race, religion etc.), b) that this discrimination can be direct or indirect, and c) that providing a solution to the challenge of discrimination was a "non-trivial task". Especially considering indirect discrimination (b), the authors find that simply removing or obfuscating the protected attributes from the dataset did not mitigate or resolve discriminatory classification given that other data can serve as proxies for such protected attributes. Since then, there has been an explosion in the number of research studies that use computational means to de-bias or otherwise balance out the discriminatory effect that results from learning classifiers from datasets that contain social inequalities, either direct or indirect [35, 34]. Such studies commonly define "biased classification" as *decision classification*. For example, an individual either is creditworthy or not and therefore receives a loan or does not receive a loan.

[2] https://facctconference.org/

[3] https://www.aies-conference.com/

Other studies found AI classifications to differ significantly in making *accurate* classifications across protected classes, in particular, gender and race [52]. Whereas the harm of decision bias results from a disparate decision *outcome* that follows from a classification, bias in *representational classification* creates harm because the classification *process* is more error-prone in making a classification (or an inference) for underrepresented groups. For example, a 2018 landmark research paper by Buolamwini and Gebru demonstrated that facial analysis AI produced the highest classification error rate predicting gender for darker-skinned women and the lowest classification error rate for lighter-skinned males [53].

In response to these two types of biased classifications – decision bias and representational bias – researchers have applied various programmatic methodologies to balance out disparities in false positive and false negative rates of classifiers across different demographic groups.

Implementing a particular fairness conceptualization is a normative decision

Yet, the various attempts to computationally de-bias datasets to "minimize disparities across different demographic groups" [35] led to a second defining revelation: the concept of "disparity" can be conceptualized in different ways. Thus, picking a particular definition of disparity and subsequently mitigating its effects successfully does not necessarily create an overall fair AI-based classification system. Classification systems can, in most cases, only operationalize *one* fairness conceptualization and *digital* classification systems are no exception. Digital classification systems can only satisfy the conditions of a particular fairness definition and so require the normative acceptance of trade-offs [54, 34].

A now famous example illustrated this conundrum. In 2016, an investigation of the automated recidivism risk tool COMPAS by the independent journalism consortium ProPublica demonstrated how optimizing a classification system for outcome parity can be evaluated as fair by one definition and unfair by another [55]. From Equivant's perspective (then Northpointe), the developers of COMPAS, the system had outcome parity because races were represented proportionally among those with high risk. To Equivant, the system was fair. ProPublica, on the other hand, conceptualized disparity minimization as equal false positive rates among races. In COMPAS, however, among those that were classified to be a high threat but were not a high treat in truth (i.e., false positives), people of color were misclassified in this way twice as often than white people. As a result, more people of color were erroneously kept in custody in comparison to white people. To ProPublica this meant that the system was unfair.

These two conceptualizations of disparity minimization are mutually exclusive in any decision-making process – digital or analog – and cannot be implemented in one classification system. Thus, one can argue that COMPAS optimized its classifications for a particular fairness conception but at the same time perpetuated racial disparities. This example showed that implementing a particular fairness definition is a normative choice as it entails a choice of one fairness definition over the other.

Challenges of current approaches to legitimate AI-based classification systems

A strong focus on "technical fixes"

Given its "historical beginnings" in the 2010s, technical fixes to ethical challenges that result from novel digital classification technologies promise convenience and effectiveness in mitigating disparities and injustices. Technical solutions such as de-biasing datasets are effective in the sense that they fit the technical affordances of single, individual implementations of classification systems. Essentially, ethical challenges are viewed as a distinct type of "optimization problem" [56]. The advantage of such technical fixes is that engineers and practitioners can implement them without necessarily requiring any other ethical competence. There is a sense that ethical challenges that result from the technical realization of classification systems are solved "at the root". However, only focusing on technical fixes bears the risk of pushing other ethical ramifications to the back of the agenda.

A strong focus on the *consequences* of digital classifications

Technical fixes to AI ethics challenges generally align better with a classic utilitarian (or consequentialist) conception of ethics. This approach to ethics evaluates the "goodness" or "badness" of an act solely in the consequences that such an act brings about [57]. Considering AI classification systems, most technical fixes represent statistical manipulations with the aim to generate parity in *outcome*. This satisfies two necessary conditions of utilitarian ethics: first, it presupposes that ethical scenarios consist of discrete, knowable, and commensurable choices and, second, that the overall "goodness" or "badness" of their enactment can be evaluated by the consequences they will bear. Prediction-based AI systems classify individuals into discrete categories (knowable choices) and the consequences that a particular classifier "enacts" can be directly evaluated (e.g., parity in outcome). Thus, applying technical fixes to legitimate digital classification systems is strongly supported by the legitimacy of utilitarian ethics. This combination of technical fixes and utilitarian principles can lead to the negligence of an engagement with the conceptual contestedness of the ethical principles that systems try to optimize for such as fairness as parity in outcome [34] or differential privacy [58].

A strong focus on "principlism" in AI ethics

Several authors have noted the prominence of "principlism" in AI ethics [59, 60]. Around the world, one can observe how companies, governments, non-governmental and other organizations have formulated "principles for ethical AI": Google, Deutsche Telekom, the High-Level Expert Group on Artificial Intelligence appointed by the European Commission as well as the Association of Computing Machinery (ACM) have all developed and published AI ethics principles in their organizational guidelines in the past years. A 2019 analysis of 84 such ethical guidelines revealed "transparency, justice and fairness, non-maleficence, responsibility and privacy" to be the core ethical principles in AI ethics guidelines [61]. Clearly, powerful organizations have become aware that AI-based classification systems create ethical challenges.

The problem with a reliance on AI ethics principles is that they often remain empty concepts. When there is little conceptual engagement with such principles then they may be desirable but lack any specificity regarding their implementation in individual digital classification systems. Abstract ethical principles leave too much room for interpretation so that organizations can pick a conceptualization of fairness, privacy, or responsibility that does not stand in conflict with other organizational targets and aims. This can, in some cases, be a justifiable approach but so far "principlism" often represents not much more than vaguely-defined abstract ethical principles.

Connecting to the criticisms around "principlism", several authors have noted that in many organizations that develop digital classification systems, "ethics work" is often seen as a marketing strategy. Vague ethical principles are particularly prone to forms of "ethics washing" [62]. Organizations can claim they are ethical and engage in ethical oversight because they have formulated a set of ethical guidelines. However, organizations face little accountability in justifying what these principles actually denote, whether they are really enforced, and how they are weighted against the organization's other aims. This has led to AI ethics having been called "toothless" [63], "useless" [64], and a "fig leaf" [60] to reflect organization's lack of facing consequences for non-compliance with vague ethical principles. Often, organizations that have installed internal ethical oversight tend to implement a narrow conception of AI ethics [62]. Such initiatives often look useful and progressive, but may cover up real risks and harms that result from digital classification systems.

Building on the contributions that have been made in the field of AI ethics, the goal of this thesis is to advance our understanding of the legitimacy of different digital socio-technical classification systems. First, taking a system-level perspective, I explore how the Chinese government justifies and implements a nationwide digital social credit system. Second, I investigate procedural normative choices in social media classification and study how social media users perceive normative trade-offs in this context. Finally, using qualitative and computational methods, I analyze how non-experts and people with AI-competence ethically evaluate facial analysis AI.

1.9 Roadmap for next chapters

In the next chapter (chapter 2), I provide a descriptive overview of our research methodologies. This overview will include statistical and computational methodologies, experimental vignette studies, as well as qualitative analyses. Chapter 3 presents our research papers on the Chinese (SCS), chapter 4 our research on the procedural normativity of social media profiling. Chapter 5 presents our papers on facial analysis AI. Finally, I will summarize key takeaways and offer final reflections in chapter 6. Published versions of the research papers can be found in chapter 7.

2 Research Methods

2.1 Standard Statistical Methodologies

2.1.1 Parametric Testing

For our empirical research studies, we applied a range of standard statistical methodologies to explore differences between groups. For example, in our research studies on facial analysis AI (chapter 5), the parametric two-sided Welch two-sample t-test [65] was used to test whether participants' ratings of facial analysis AI would significantly differ between ratings in the low-stake advertising versus the high-stake hiring context.

Welch's t-test compares means between two groups that are independent from each other. As the experiment was a between-subject design, this condition was fulfilled. Such testing for differences across groups presupposes a null hypothesis and an alternative hypothesis. The null hypothesis denotes that there is no significant difference between the two groups. The alternative hypothesis denotes that there is a significant difference between the two groups [66]. The significance of a difference is given by the p-value. A p-value of less than 0.05 is taken to represent a significant difference and, subsequently, the null hypothesis can be rejected [67]. The means of both groups differ significantly.

2.1.2 Analysis of Variance (ANOVA) & Multivariate Analysis of Variance (MANOVA)

Just like a t-test, an ANOVA tests whether the means in one group are significantly different to the means of another group. However, an ANOVA extends t-tests by comparing more than two variables with each other [68]. Conducting several t-tests leads to an increased probability of making a Type I error (rejecting a null hypothesis that is true) [69]. ANOVAS control for Type I errors by conducting all comparisons simultaneously. It produces an F-score that represents the variance between variables divided by the variance within the variables. The F-score can be used to determine whether there is a significant difference between the variables and thus whether the null hypothesis can be rejected.

In our research studies on facial analysis AI (chapter 5), we used ANOVAs to test whether factors (i.e., variables) other than the decision-context had a significant influence on participants' ratings. We also performed an exploratory factor analysis [70, 71] to understand the underlying structure of participants' ratings. This resulted in two constructs that we termed "first-order inferences" and "second-order inferences". Our analysis now had two dependent variables: ratings of first-order inferences and ratings of second-order inferences.

Here, a MANOVA can extend an ANOVA by testing whether multiple independent variables influence multiple dependent variables [72].

2.1.3 Exploratory Factor Analysis (EFA)

EFA is a multivariate statistical instrument to measure the smallest number of constructs that can represent the variance within a set of measured variables [73, 69]. EFA is used for dimensionality reduction. In our research on the ethics of facial analysis AI (chapter 5), we wanted to understand whether participants' ratings for a set of facial AI inferences contained a hidden relationship. That is, rating behavior could be similar for some AI inferences, which can be expressed as "factors" by EFA. EFA is often used as an initial statistical analysis to form an idea about the underlying structural patterns in a collected dataset. In contrast to confirmatory factor analysis (CFA), EFA does not require the pre-specification of the number of factors that one expects to find in the dataset [71]. EFA is commonly used when researchers cannot apply theoretical constructs to their data in order to map potential factors to such theoretical constructs.

There are several methods to determine the number of factors within a set of measured variables for EFA. Among these are parallel analysis, scree plotting or Velicer's minimum average partial (MAP) test [70, 71]. Usually, researchers use several of such tests to justifiably select the number of factors prior to conducting an EFA. If two or more tests converge on a number of factors then this number of factors can be used with confidence for the subsequent EFA analysis.

2.2 Qualitative Content Analyses

Our research on narratives in the Chinese SCS (chapter 3) as well as on the ethics of facial analysis AI (chapter 5) required a methodology to analyze textual data qualitatively. Here, the methodological conceptualization of "qualitative analysis" in relation to textual data denotes the interpretation of text sections by assigning semantic categories (also called "codes") to the text [74, 75]. After qualitative analysis, categories can be further analysed quantitatively (e.g., category frequencies), which is why some researchers refer to qualitative content analysis as a mixed-methods methodology [75].

We applied two different versions of qualitative content analysis in our research. In our research on the ethics of facial analysis AI, we collected nearly thirty thousand individual written answers from participants. Here, the interpretation of participants' written responses was guided primarily by our overall research question. We wanted to understand how participants justify their rating behavior of specific AI inferences. Thus, in creating a code book – the final analytic scheme of categories to be applied to the entire textual corpus – categories were created to reflect participants' underlying reasoning for justifying their ratings. This can be called "conventional content analysis" whereby codes are developed *inductively* [76].

In so doing, one researcher typically starts to label a subset of written responses to formulate a preliminary code book. This preliminary code book is then discussed with another researcher and potential interpretative ambiguities are discussed and resolved. Both researchers then apply the code book to another subset of the written responses and hence meet for another round of discussion and code refinement. If the two researchers cannot resolve an issue, a third researcher is consulted and a decision is made. Once the entire dataset has been labelled, an inter-coder reliability is calculated. This inter-coder reliability provides a quantitative measure of the agreement in labeling between the two coders. It is commonly calculated using Krippendorff's α [77] whereby an α of ≥ 0.7 is taken to represent sufficient reliability. In our first research paper on the ethics of facial analysis AI, we used qualitative content analysis to create a code book. Given the large number of comments, we then used the transformer-based language model roBERTa (see section 2.3.2) to classify the remaining written responses [78].

In our research on the use of narratives in the Chinese SCS (chapter 3), one core research goal was to understand how authors portrayed the moral experiences of protagonists in different moral scenarios. With this purpose in mind, we applied a so-called *directed content analysis* [76]. Directed content analysis uses theoretical constructs from previous research as codes to interpret a given text [79]. In this way, the application of existing theoretical constructs to a novel text corpus reconfirms the validity of these constructs. Thus, researchers need to find textual passages that serve as evidence for a chosen theoretical construct. Directed content analysis follows a *deductive* approach in code development.

2.3 Computational Methodologies

2.3.1 Web crawling and web scraping

Web crawling and scraping are computational methods that automate the systematic collection of public data from accessible web pages. A web crawler is programmed to systematically search the internet by following specified URLs [80, 81]. A web scraper is programmed to extract pre-specified information from web pages. Collecting publicly available data from web pages requires the programming of a web crawler and a web scraper. Web crawling and scraping can be used for different purposes. I will touch upon these briefly but will focus on web crawling and scraping for automated information retrieval whereby crawling results in a link list that is fed into a web scraper that downloads the requested information [82]. This corresponds to how web crawling and scraping were used in the research on the Chinese SCS (chapter 3). Web crawling and scraping allows for the collection of data from web pages when web page owners do not provide any application programming interface (APIs). In the age of big data mining, web crawling and scraping are essential tools to capture the vast amount of data available online for further analysis [83]. In our research on the Chinese SCS, we also point out ethical challenges of web crawling (see research paper in chapter 3.2).

Web crawling

Most prominently, web crawling is the key component of search engines that scan web pages with the purpose to compile an index [84, 85]. Users can make queries against the index and retrieve web pages. Besides their prominent importance in search engines, data archiving systems such as the Wayback Machine[1] apply web crawlers to take screen shots of web pages periodically at different times. To study the key information platforms of the Chinese SCS, the headless browser Selenium[2] that simulates a human browser was used to control a web crawler that creates a link list leading to specific SCS credit records. As we explain in detail in our research paper in chapter 3.2, many of the SCS websites could be structurally represented as trees with nodes thanks to their underlying static Hypertext Markup Language (HTML) with Cascading Style Sheets (CSS) implementation. This structure could be exploited by crawlers. Typically, HTML typically represents the content of the web page while CSS is a design language that presents the content and defines how HTML items are displayed.

The crawler starts accessing a specified landing page (i.e., of a provincial SCS platform, the root of the tree) and adds every HTML reference (also known as "deep link" or "tree leaf") to the link list. In the end, the crawler produces a simple text file with a list of URL links. In this list, every row contains the deep links with the so-called "href" attributes, which the crawler is programmed to fetch. In the case of SCS blacklist and redlist web pages, the deep links represented individual blacklist or redlist *records*. In the end, this link list will be passed on to the web scraper.

Web scraping

Here, the scraping framework Scrapy [86] was used to download pre-specified information from blacklist and redlist credit records. In scrapy, code with all crawling and scraping instructions are defined in a class called "spider"[3]. A spider contains the algorithm to execute a search query, the link aggregation, and the information extraction. For the data extraction of blacklist and redlist records, such information included the reasons for being listed, the gender of the listed entity (if a person), or the Unified Social Credit Code, a unique SCS identifier. Instructions on the information that the spider extracts is defined in a parse function. The Scrapy framework offers a set of relevant features that make scraping significantly more efficient [87]. The framework consists of a scheduler that manages requests and responses, a downloader for web data, and an item pipeline for data storage and transfer to databases, among others [86]. The framework also handles 404 errors, request delays and downloading problems. Given that most SCS web pages used CSS formatting, scrapers could extract information by CSS selectors and transfer the data into a table or a database.

Taken together, web crawling and scraping have become standard tools that, in combination, are primarily used for data extraction on web pages and are used in a variety of application

[1] `https://archive.org/web/`
[2] `https://www.selenium.dev/`
[3] `https://docs.scrapy.org/en/latest/`

fields [88]. In our work on the Chinese SCS, we show an example of a spider we built and applied to crawl redlist and blacklist records (see Auxiliary Material of research paper in chapter 3.2).

2.3.2 Natural language processing (NLP) techniques

A highly active research field called NLP combines linguistic theories with stochastic models and computer science methodologies to automate a range of different language tasks [89]. Machine learning-based approaches to NLP include language understanding, machine translation, question-answering, and text summarization. With the explosion of textual data on the Internet, NLPs are essential to data analytics in order to infer valuable information from raw text data. NLPs played a prominent role in our research article on non-experts' ethical justifications of facial analysis AI (see research article in chapter 5.2).

Term frequency-inverse document frequency (TF-IDF)

One of the most basic NLP techniques is TF-IDF. TF-IDF measures the importance or relevance of a specific word in a set of documents [90, 91]. "TF" stands for term frequency and simply denotes the frequency of a term (or word) in a document (or text). *Relevance* here is defined as the overall informativeness of a term for a document. The relevance of a term in a document is given by the number of times a term occurs in a document divided by the overall number of words in the document. TF-IDF measures the relevance of a word in a set of documents. Simply counting the number of times a word occurs in a set of documents would not necessarily reflect its relevance: stop words ("a", "the", "if"...) occur very often but carry little informativeness [92]. To filter out stop words, TF-IDF weighs terms by dividing the number of times a term occurs by the number of documents in a document corpus that contain the term, which is represented by the "IDF" (inverse document frequency). TF-IDF was used to weigh terms in the narratives on "good" and "bad" Chinese citizens in the Chinese SCS (see research paper in chapter 3.1).

Topic modeling with latent dirichlet allocation (LDA)

Topic models or topic modeling can be described as an automated procedure for coding the content of a corpus of texts (including very large corpora) into a set of substantively meaningful coding categories called "topics" [93, 94, 95]. LDA presupposes that a textual corpus can be represented by a pre-specified number of "latent topics". It further assumes that the meaning of a topic is represented and "embedded" in a cluster of words. As such, topic modeling considers a textual corpus as a "bag of words" that contains hidden topics. Their co-occurrence in a textual corpus is not by chance but corresponds to the existence of a specific topic.

Topic modeling is applied as an unsupervised classification method but there are variations that allow its use in supervised classification [94]. One of the most widely used type of topic modelling is LDA [95, 96]. LDA is a generative, probabilistic model that follows two main

principles. The first principle of LDA is that it considers every document (a discrete text) as a mixture of topics. That is, each text document can consist of a multitude of different topics. The second principle of LDA is that each topic consists of a mixture of different words. It follows that topics do not necessarily differentiate themselves by a set of unique words. The same word can occur in two different topics. It is the mixture of words that make up a discrete topic.

LDA calculates the probability for each word (or term) belonging to a topic [97, 93]. One can then select the 10 terms with the highest probability per topic to represent that topic best, for example. LDA requires pre-processing of the textual data. This pre-processing phase is called tokenization, the corpus is stripped off all of its semantically irrelevant punctuations and stop words and all semantically relevant words are transformed into their canonical form (i.e, their stem). Moreover, LDA, as well as other topic models, does not automatically determine the optimal number of topics in a corpus [97, 93]. It assumes that the number of topics in a text corpus is already known. The right number of topics is relatively small and results in the highest probabilities for words across topics. Researchers have developed different methodologies to optimize the number of topics for LDA [98].

Importantly, researchers need to add what they believe is the most semantically appropriate label to a given topic as LDA does not generate any semantic inferences. Finally, LDA has been used to analyze social media posts [99], newspaper articles [100], or politicians' speeches [96]. In our work on the Chinese SCS, we used LDA to understand the major topics in SCS narratives on "good" and "bad" behavior (see research article in chapter 3.1).

Bidirectional Encoder Representations from Transformers (BERT) & A Robustly Optimized BERT Pretraining Approach (roBERTa)

At their core, language models represent a probability distribution over a word sequence whereby they predict the probability of a word's position in a sentence based on different conditionals [101, 102]. In the "classic" Markov language model the probability of the next word in a sentence is estimated by its preceding word [101]. Language models cannot understand human language, their output is based on learning common semantic associations in large sets of textual data (the entire Wikipedia corpus, for example).

In recent years, natural language processing has seen significant progress on many benchmark tests thanks to development and use of pre-trained language models [103, 78]. Pre-trained language models such as BERT (or its fine-tuned version roBERTa) only require labeling of few data given the model's extensive *pre*-training. Such pre-trained models already contain a significant amount of lexical, syntactic, and semantic knowledge [102]. In BERT, which learns in a bi-directional manner (i.e., left and right), pre-training primaririly consists of learning how to recover words in masking tasks. The bi-directionality of BERT and roBERTa make it ideal for classification tasks, which are language understanding tasks with word sequences in a single document as input and one or multiple labels as output. Today, most pre-trained models incorporate the transformer architecture as proposed by [104].

Critical to the transformer architecture of BERT is the encoder that represents words in a sequence as a numerical vector [103]. The vector representation of each word does not only contain information of the word's meaning but also of its contextuality, which is based on the bi-directional words around it. Pre-trained language models offer "task universality" [101] in that they can be adjusted for different language processing tasks. They already include a significant amount of factual knowledge.

We use the language model roBERTa for the classification of ordinary people's ethical justification of facial analysis AI (see research article in chapter 5.2). This is an ideal task for a transformer-based language model such as roBERTa that has achieved state of the art accuracy for language understanding and classification tasks [105].

2.4 Experimental Vignette Studies

Experimental vignette surveys have become a key methodological tool to study participants' beliefs, judgements, values, choices and so on in carefully curated scenarios [106, 107, 108].

Vignettes consist of two parts: first, the vignette itself and, second, a common survey instrument for measurement [109]. The vignette is a hypothetical scenario that study participants are asked to read. In contrast to standard surveys, vignettes offer a more detailed description of a scenario that participants need to evaluate. In short stories (i.e., vignettes) experimenters can account for different contextual contingencies in the scenarios [106, 107]. Factorial vignettes enable researchers to deliberately vary the information presented in a vignette.

Just like in a real experimental setting, factorial vignettes allow researchers to modify information in the vignettes in such a way that they present different independent variables. Researchers can then measure how these different independent variables influence the dependent variable, which is often represented by participants' ratings on a Likert scale [110]. In a typical factorial design, researchers can vary the "factors" presented in a vignette as well as the "levels" of such factors. For example, in a vignette study on gender income gaps, Steiner et. al. developed vignettes with several factors (education, occupational experience, industry, gender) that can all take on different levels [109]. Using such a factorial design, researcher can "...assess the importance of those vignette factors which causally affect individual responses to the contextualised but hypothetical vignette settings" [106]. Furthermore, researchers can conduct within- and between-study designs, or a mixture of both. Experimental vignette studies add an experimental character to standard survey studies [107]. They offer more realism by approximating real-life scenarios and account for contextual or situational factors that may be of particular importance for a decision-making process, an interaction, a characterization of an action and so on [111].

All of the above make experimental vignette studies a useful instrument to study moral perceptions and judgments *per se* and in the context of artificial intelligence. In a relatively new field called *experimental* philosophy [108], researchers design vignettes to understand how "ordinary" people define the essence of contested philosophical concepts such as intentionality,

knowledge, belief and so on. One prominent discovery that resulted from experimental philosophy has been termed the *Knobe Effect* after its discoverer Joshua Knobe [112]. Knobe tried to understand people's perceptions of an intentional action by an agent, that is, the conditions under which one can justifiably say that an agent has performed an action intentionally. The *Knobe Effect* focuses on the intentionality of side effects that emerge from the primary action of an agent. It describes the phenomenon that people claim that an agent acted intentionality when the side effects of a primary action result in negative consequences. They do not ascribe intentionality when the side effects result in positive consequences.

In our research, we designed vignettes to study what ordinary people (non-experts) perceive as a fair inference in the context of facial analysis AI. Here, we could vary the decision context to understand whether ordinary people perceive the same AI inference more fair in one decision context in comparison to another decision context (see research paper in chapter 5.2). Moreover, we repeated the experiment with an expert group (high AI knowledge) to understand whether the level of AI knowledge correlates with different perceptions on fair AI inferences in facial analysis (see research paper in chapter 5.3). We also used a single-treatment vignette study for our research on social media users' perceptions of social media user profiling (see research paper in chapter 4.1).

3 Published Articles Part 1: The Chinese Social Credit System

Research Article 1: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior (2019)

Research Article 2: Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness (2021)

Research Article 3: Ordinary people as moral heroes and foes: Digital role model narratives propagate social norms in China's Social Credit System (2022)

Please note that the published articles are slightly modified mainly to allow for unification of format and reference style. References for each research paper appear in the overall bibliography at the end of the doctoral dissertation. Published versions of the research articles are appended to end of the doctoral dissertation in chapter 7.

3.1 Research Article 1: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior

Authors

Severin Engelmann, Mo Chen, Felix Fischer, Ching-Yu Kao, Jens Grossklags

Publication Outlet

FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency; January 2019; Pages 69–78; https://doi.org/10.1145/3287560.3287585

Abstract

China's Social Credit System (SCS, 社会信用体系 or shehui xinyong tixi) is expected to become the first digitally-implemented nationwide scoring system with the purpose to rate the behavior of citizens, companies, and other entities. Thereby, in the SCS, "good" behavior can result in material rewards and reputational gain while "bad" behavior can lead to exclusion from material resources and reputational loss. Crucially, for the implementation of the SCS, society must be able to distinguish between behaviors that result in reward and those that lead to sanction. In this paper, we conduct the first transparency analysis of two central administrative information platforms of the SCS to understand how the SCS currently defines "good" and "bad" behavior. We analyze 194,829 behavioral records and 942 reports on citizens' behaviors published on the official Beijing SCS website and the national SCS platform "Credit China", respectively. By applying a mixed-method approach, we demonstrate that there is a considerable asymmetry between information provided by the so-called Redlist (information on "good" behavior) and the Blacklist (information on "bad" behavior). At the current stage of the SCS implementation, the majority of explanations on blacklisted behaviors includes a detailed description of the causal relation between inadequate behavior and its sanction. On the other hand, explanations on redlisted behavior, which comprise positive norms fostering value internalization and integration, are less transparent. Finally, this first SCS transparency analysis suggests that socio-technical systems applying a scoring mechanism might use different degrees of transparency to achieve particular behavioral engineering goals.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

3.1.1 Introduction

Moral thinking and action necessarily depend on informational resources. When an individual asks: "What is the right thing to do?", he or she essentially relies on information that renders a conclusion morally justifiable. In philosophy and anthropology, descriptive morality refers to how groups or societies negotiate codes of conduct (or norms) that are morally acceptable or unacceptable [113, 114]. As a consequence, an individual's moral accountability tends to be proportional to his or her knowledge of good and bad moral behavior underlining the epistemic character of morality [115]. In 2014, the Chinese government issued a plan for a nationwide digital scoring system known as the Chinese Social Credit System (SCS) classifying behavior into morally "praise-" and "blameworthy" [116]. Thereby, all legal entities including companies and public institutions (among others) receive an 18-digit ID called the Unified Social Credit Code,[1] which corresponds to the 18-digit ID card number for Chinese citizens. Presumably, based on these IDs, the SCS will collect and evaluate behavioral data and may assign scores that result in material benefits and reputational praise or material exclusion and reputational loss. Or, in the words of the Chinese government, the goal of the SCS is to "allow the trustworthy to roam everywhere under heaven while making it hard for the discredited to take a single step" [117, 116].

But how can citizens, companies, and social institutions know what behaviors are "good" and "bad" in the SCS? Put differently, how transparent is the current SCS in providing information on "good" and "bad" behaviors? Answering this question requires a conceptualization of transparency. Here, we rely on the definition proposed by Turilli and Floridi, which conceptualizes transparency as "the choice of which information is to be made accessible to some agents by an information provider" [118]. First, this definition distinguishes between an information provider, which makes information accessible, in this context the Chinese government, and agents or entities that depend on this information for their decision-making. Secondly, this definition recognizes that information transparency is an "ethically impairing or enabling factor when the information disclosed has an impact on ethical principles" [118]. Both of these components are highly relevant for the SCS since participants are dependent on the information provided to make decisions that can lead to reward or punishment.

Recently, the Chinese government has started issuing behavioral information on several platforms (see Section 5.1.2 for more information). In this empirical study, we review a subset of this behavioral information released on two central SCS platforms: the official SCS national website "Credit China" and its equivalent municipal outlet "Credit China (Beijing)". On the former site, we collect and analyze 156 news reports about "good" behaviors (we refer to as "positive" cases), and 789 equivalent reports about "bad" behaviors ("negative" cases). In these "negative" portraits, individuals are commonly stereotyped as so-called "Laolai (老赖)" – the epitome of a financially dishonest individual in China. Since all stories we collected are news reports about real-life events portraying a morally "good" or "bad" individual, they all include descriptive norms highlighting "desirable" and "undesirable" characteristics of individuals in Chinese society today.

[1] `http://www.gov.cn/zhengce/content/2015-06/17/content_9858.html`, last accessed on November 19, 2018.

Next, on "Credit China (Beijing)", we retrieve a large number of records of "good" and "bad" behavior from the so-called Redlist and Blacklist. Thus, our approach is as follows: first, we collect and statistically analyze close to 200,000 Blacklist and Redlist records from "beijing.gov.cn/creditbj", the SCS's information platform for China's capital, Beijing. Hence, based on machine learning topic modeling and manual text coding, we identify the common semantic patterns of close to 1000 reports on "good" and "bad" behavior published on the national SCS platform "www.creditchina.gov.cn".

We show several informational asymmetries that characterize the current degree of transparency of the governmental SCS's information platforms. Finally, we discuss how degrees of transparency could correspond to different incentive strategies of socio-technial systems that rate legal entities in society. Our paper has the following structure. In Section 5.1.2, we discuss the development of China's SCS and review related work. In Section 3.1.3, we present our data acquisition and data analysis approach. We conduct our analysis in Sections 4.1.8 and 3.1.5. We discuss our results and offer concluding remarks in Section 5.2.6.

3.1.2 Background

The implementation of the SCS rests on at least three main factors: First, lack of honesty and trust[2] in Chinese society has become a serious issue manifested in the numerous news reports about food poisonings, chemical spills, financial and telecommunications fraud, and academic dishonesty over the past two decades [119, 120]. It is estimated that Chinese enterprises suffer from a loss of 600 billion RMB (around 92 billion USD) per year due to dishonest activities[3]. According to a survey conducted by Ipsos Public Affairs [121], "moral decline" was regarded as the most serious issue in China in 2017. 47% of Chinese respondents ranked it as one of the top 3 greatest concerns, while the same issue was only mentioned by 15% of total respondents worldwide.

Secondly, China's SCS is expected to boost the domestic economy. The Chinese government hopes that the SCS will give millions of Chinese citizens without a financial history access to credit and investment opportunities in the domestic market. China has the largest unbanked population in the world (in absolute numbers), with more than 225 million citizens having no bank account [122]. So far, only 320 million Chinese citizens have a credit record[4]. However, the sustainability of China's economic growth partially depends on an increase in domestic spending. Through the SCS, citizens could apply for loans based on trustworthiness scores without having to prove their financial creditworthiness.

[2]The characters "诚信 (chengxin)" literally mean both honesty and trust in Chinese.

[3]This information is included in the "Report on China's Honesty Building Situation (Zhongguo Chengxin Jianshe Zhuangkuang Baogao)". The full report is not publicly available, but parts of the report (in Chinese) are accessible through: `http://society.people.com.cn/n1/2016/0523/c1008-28370202.html`, last accessed on November 19, 2018.

[4]See "Inspiration of the US Non-traditional Credit Information Mechanism" available on the platform of "Credit China" at `http://www.creditchina.gov.cn/zhengcefagui/tashanzhishi1/201712/t20171207_98701.html`, last accessed on November 19, 2018.

Finally, in Chinese society, the concept of personal identity is largely determined by Confucian principles [123, 124]. Accordingly, personhood is supposed to extend from the private to the public sphere thereby somewhat losing its private and public boundaries. In other words, normative expectations on individuals hardly account for the distinction between a private and a public sphere. The division between a private and a public persona is often conceived as trying to be secretive as privacy is commonly conceived as hiding something shameful [125]. In fact, until recently, privacy was primarily protected under the right of reputation in Chinese civil law [126]. At the same time, the public interest ranks highly in Chinese civil law [127]: "private information protected from disclosure refers to information that is irrelevant to the public interest or to the interests of other persons." However, while the Chinese concept of privacy is evolving, it is expected to remain distinct from other societies [128]. Overall, the introduction of the SCS is hardly perceived as a privacy-violating system in Chinese society, which is perhaps surprising from a Western perspective [129].

Current state of the SCS

At the current stage, the SCS remains fragmented, being developed at national, provincial, municipal, and ministerial levels with no clear unified structure. In the past years, provinces and cities have developed various prototype models for the SCS [130, 131]. Importantly, the SCS also takes companies, government departments and judicial organizations as its targets [116]. This means that some companies have a special role in the SCS. Since 2015, eight companies were granted permission to run individual credit services with the purpose to implement pilot SCS programs [132]. Individually, none of the eight companies received a licence to continue individual credit services after the two-year trial period ended in 2017. Instead, together with the China Internet Finance Association (run by the People's Bank of China), they recently have become common shareholders of a company called Baihang Credit, which received the first credit scoring licence in February 2018.

Related Work

We are unaware of *any* research project that conducts a data-driven analysis of the currently observable data practices of key sites of China's SCS. However, we have identified two empirical research studies that help understand how the SCS is being communicated and discussed by Chinese media [132], and how it is being perceived by Chinese citizens [129].

Ohlberg et al. collected official Chinese news articles and public communications, as well as social media postings on Chinese blogs, forums, and bulletin board services about the SCS for a six-month period in 2017 [132]. The large majority of news articles has a positive focus and highlight the SCS as a "cure-all for social and economic problems". Criticism is mostly aimed at the slow implementation progress or directed at commercial initiatives in the SCS. Citizens' social media postings rarely address privacy issues and rather focus on how to game the system to achieve a higher social credit score within commercial SCS applications. Of relevance to the latter point, the implications of gamifying social credit are also being discussed from a non-empirical perspective by other scholars [133, 134].

Kostka [129] conducted an online survey with about 2,200 Chinese citizens that was distributed via different channels including websites and apps. Due to the widespread internet surveillance in China, the validity of such online surveys remains questionable at least to some extent. According to her findings, about 80% of the respondents have a positive perception of governmental and commercial SCS initiatives. Interestingly, older and more educated respondents have a higher approval rating. In contrast, these demographic factors are typically associated with higher privacy concerns in Western societies (see, for example, [135]). Several policy papers address the relationship between the SCS and the danger of mass surveillance (e.g., [136]).

Finally, there is rigorous work on comparing financial credit reporting systems [137], which, however, predates the emergence of the SCS in China and focuses on the financial aspects of credit reporting. Likewise, privacy considerations concerning private entities facilitating credit and background reporting have, for example, been explored by Hoofnagle [138].

Ethical Issues

Our analysis is built on publicly available data from key sites of China's SCS, which is posted with the intent of public scrutiny. Our paper includes screenshots from the *currently available* implementations. We have blurred any personally identifiable data.

3.1.3 Methods

We used computer-assisted content analysis methods to explore the level of transparency of current behavioral information published on the two previously mentioned SCS websites. First, the column-and-row structured records of both the Blacklist and the Redlist on the SCS's Beijing platform[5] were crawled and statistically evaluated. Hence, to understand the semantic and structural patterns of both "positive" and "negative" case studies, we crawled news reports on "bad" behavior labeled as "Typical Cases (典型案例)"[6] and on "good" behavior labeled as "Stories of Integrity (诚信人物/故事)" under the section of "Integrity Culture (诚信文化)"[7]on the national SCS information platform "Credit China"[8]. We then applied statistical topic modeling based on Latent Dirichlet Allocation (LDA) to all available 156 news reports on "good" behavior ("positive" cases) and 789 news reports on "bad" behavior ("negative" cases) on August 12, 2018.

We preprocessed the downloaded documents by applying *jieba*[9] for segmentation and stopword filtering of Chinese text. We used the stopword corpus compiled by the Chinese search engine Baidu[10]. After tokenization of the given text, we applied *tf-idf* to re-weigh term counts.

[5] `http://www.creditbj.gov.cn/xyData/front/creditService/initial.shtml%20?typeId=4`.
[6] `https://www.creditchina.gov.cn/home/dianxinganli1/?navPage=6`.
[7] `https://www.creditchina.gov.cn/chengxinwenhua/chengxingushi/`.
[8] `https://www.creditchina.gov.cn/`.
[9] https://github.com/fxsjy/jieba.
[10] http://www.baiduguide.com/baidu-stopwords.

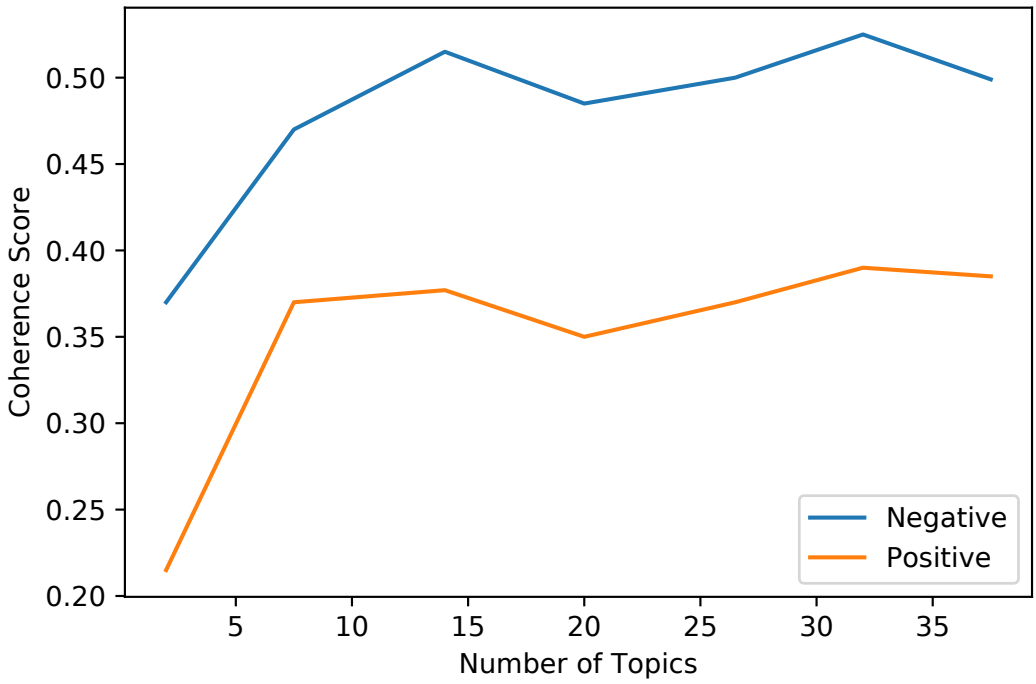


Figure 3.1: Coherence score C_v for topic models of negative and positive case studies using different topic counts.

As we had no reasonable expectation for the number of topics k to be detected within the given document corpus, we performed optimal topic number search. Thereby, we created several LDA models for "positive" and "negative" case studies and calculated the topic coherence measures C_v as proposed in [139]. We started with $k = 2$ and increased the number of topics until an upper bound of $k = 40$. As shown in Figure 3.1, coherence values of models for both document sets increased until $k = 15$ before flattening out. Therefore, we investigated the top-30 most salient terms for each of the fifteen topics produced by these models [140]. Thereby, we set $\delta = 0.6$ within the applied relevance metric [141]. Moreover, we also reviewed the results for $k = 10$, $k = 20$, and $k = 30$ in order to further manually verify the optimal topic number. We found the optimal model with $k = 10$ for both "positive" and "negative" cases. Finally, we further selected 5 main topics for the "positive" cases and 7 topics for the "negative" cases (see Table 3 in the Supplementary Materials for topics selected for the "positive" cases, and Table 4 in the Supplementary Materials for "negative" case topics).

Based on our topic modeling results, we selected the 4 most related cases (highest predicted probability of belonging to the topic) for each of the topics.[11] We then manually analyzed 20 "positive" cases and 26 "negative" cases[12] in detail. One author first reviewed 5 "positive" and 5 "negative" cases, respectively, and drafted a coding guide, which was then reviewed iteratively by another author, refined, and retested to generate consistent definitions. As a result, we developed two coding schemes for "positive" and "negative" cases (see Table 3.1 for the coding scheme applied to "positive" cases and see Table 3.2 for the coding scheme used to analyze "negative" cases). After reliability was established, we examined all 46 cases for structural and thematic commonalities. Each coding sheet contained the information from one "positive" or "negative" case. Once the coding sheets were completed, we grouped and analyzed the information contained in them.

3.1.4 Results

Blacklists

On the platform of "Credit China (Beijing)", we found three publicly accessible databases providing information on "bad" behavior, all of which could be queried by search term. Translated from Chinese (see Figure 3.2), they were termed the following: 1) Blacklist (1,137,546 entries), 2) Special Attention List (9,229,179 entries), and 3) Administrative Punishment (14,885,789 entries).

The Blacklist further contained 16 subcategories for "bad" behavior. For the Blacklist, we crawled two of these subcategories, one containing records of individuals that have been banned from participating in the securities market (Securities Market Entry Prohibition, 422 entries) and one listing companies with debts (Blacklist of Company Debtors, 1,116,707 entries = 98.2% of all Blacklist entries). For the Blacklist of individuals, all 422 entries included extensive explanations for the punishment (e.g., length of ban) referencing financial law (see

[11] For "negative" cases, there are only three cases for Topic 6 (measures taken against crime) and Topic 7 (public transport regulation violation), respectively.

[12] There were only 3 cases for 2 out of the 7 topics.

Pattern	Definition	Example
Bio-info	full name	今年70岁的刘某某，为了一句诺言，一辈子踏踏实实做一名“小村大医生”。
	age	古亭村77岁的老人蓝某某为了归还欠银行的一笔500元死账。
	living place	蓝某某出生在遂昌县云峰街道古亭村。
	profession	这位一天两次捡到钱包的"好运人"就是蒙阴一中的的英语教师耿某某。
Social class	low	一个清贫的普通农家，父亲、儿子、孙女毫无怨言地赡养一位无任何血缘关系的"外人"。
	middle	陈某的妻子说，他们家里也就是普通家庭，上有老下有小。
	high	这句话时常在内蒙古明泽集团董事长王某某的心里翻腾着。
Sacrifice for the common interest	material sacrifice	他隔天检查药柜，受潮的药直接销毁，损失的药费自己承担。
	non-material sacrifice	每天为他做三餐，每天打针吃药，就连端屎端尿的活也揽下来。
Rewards	reputational rewards	他被评为全国农村青年创业致富带头人、北京市优秀农村实用人才。
	material reward refusal	钱包的主人一个劲地要给她塞钱。肖某某坚决地拒绝了。
Virtue cascade	trustworthy and honest	为了不让养殖户遭受损失，彭某某把风险留给自己，仍按照回收合同原价收回了养殖户的肉鸭。
	hardworking	虽然有时一天连饭都顾不上吃，还帮助菜农一起装菜卸菜，忙到了深夜还要了解市场信息、掌握蔬菜的价格趋向。
	self-discipline	虽然银行减免并注销了这笔贷款，但放在我私人账户的钱一定要还上。
	helpful	积极参加协会组织的慰问残疾人、资助贫困大学生活动。
	care-taking	他们一家三代几十年如一日地照顾着丁某某老人。
	sense of responsibility	她以当好水资源质量的守望者为己任。

Table 3.1: Coding scheme for "positive" cases. All "positive" cases included biographical information of the individual and indicated his or her social class. Other codes described the individual's sacrifice for the common interest, the rewards obtained, and the further attribution of other virtues (virtue cascade).

Pattern	Definition	Example
Bio-info	anonymous (for individuals, surname only)	当宁陵县法院执行干警在被执行人郭某家的楼顶将其抓获时，郭某无奈地低下了头。
	anonymous (company name not provided)	原告北京某装饰工程有限公司为被告北京某文化有限公司所有的房屋进行建设、装修。
Implementing Agency	the court	海淀法院3月6日出动执行法官、法警等共计50余人，对15起案件进行集中强制执行。
	Public Security Bureau	华龙区法院的执行法官远赴拉萨，与当地公安机关通力合作。
	telecommunication company	由商南法院向中国移动、联通、电信三大通信运营公司出具协助执行通知书，对失信被执行人实行彩铃和短信曝光。
Causes for punishment	refusing to repay individuals	当地法院判决吕某赔偿梦某医疗费、残疾赔偿金等损失46万元。吕某拒不履行赔偿义务，甚至远走他乡。
	refusing to repay banks	岫岩法院判决某食用菌公司偿还银行贷款本金380万元及相应利息。判决生效后，食用菌公司一直没有履行。
	refusing to repay companies	原告北京某装饰工程有限公司为被告北京某文化有限公司所有的房屋进行建设、装修，施工结束后，被告拖欠原告工程款400余万元。
Reasons to fulfill obligations	actions taken by the court	在中牟法院执行干警的全力配合下，成功将被执行人吕某拘留。
	threatened to be placed on Blacklist	法院将肖某纳入了“老赖”名单里，将他的大头照向社会公布。

Table 3.2: Coding scheme for "negative" cases. All cases provided anonymized biographical information, an entity implementing the punishment, justification of the punishment, and descriptions on why the obligations were fulfilled in the end.

Figure 3.2: Three lists publishing records of "negative" behavior: from left to right, the first arrow points to Blacklist, the second arrow to Special Attention List, and the third arrow to Administrative Punishment.

Figure 3.3). Apart from the censored ID card number, the full names of all individuals were published.

Due to the large amount of company records we found on the Blacklist, the Special Attention List, and Administrative Punishment, we crawled the first 1000 pages for these lists. For the Blacklist of companies with financial debt, this resulted in a total of 131,485 entries all of which featured information on why an entity had been blacklisted (see Figure 3.4). Out of these 131,485 entries, 128,006 entries specified that the financial obligation had not been fulfilled at the time of crawling (corresponding field not shown). Entries included a reference to legal regulation and specified the full name of the company (see Figure 3.5). Note that some companies listed had multiple entries corresponding to multiple breaches. Together with these explanations, we crawled the date of publication on the Blacklist for each entry. We found that on one day in June 2018, 95.6% of all entries (125,747) had been published on the Blacklist for companies (see Figure 3.6). This probably indicates that these records had already been collected and processed by another entity before being transferred to and published on the Blacklist.

For the Special Attention List, we collected 30,625 entries containing information on companies that had violated business operation regulations. For all records collected, companies had been blacklisted for providing various types of false information to the authorities (see Figure 3.7).

Finally, our crawler returned 32,719 entries for the Administrative Punishment register that contained information on both individuals and companies (see Figure 3.8). As Figure 3.9 shows, the majority of records of the Administrative Punishment register reported traffic rule violations. Correspondingly, fines were the most widely used measure (see Figure 3.10). We also found that only company entries of the Administrative Punishment register and the

数据来源：	证监会
数据类别：	市场禁入
主体名称：	吴[redacted]
加密证件号码：	110104**********4X
证件类型：	身份证
个人代码：	
处罚处理名称：	证监法律字[2006]12号 市场禁入决定书（梁[redacted]、吴[redacted]、孙[redacted]）
处罚处理日期：	2006/11/27
处罚处理种类：	市场禁入(5年)
处罚对象类型（1组织机构，2个人）：	
真实证件号码：	
信息类型：	
处罚机关：	中国证监会
处罚决定书id：	
处罚处理内容：	证监法律字[2006]12号市场禁入决定书当事人：梁[redacted] 女，1948年出生，北京中兴信托投资有限公司（以下简称中兴信托）法定代表人、健富投资有限公司（以下简称健富投资）法定代表人，住址北京市东城区前门东大街1号2单元207号。吴[redacted] 女，1949年出生，中兴信托北京亚运村营业部经理、健富投资经理，住址北京市朝阳区北四环东路106号3号楼1804号。孙[redacted] 男，1942年出生，健富投资顾问，住址北京市东城区前门东大街1号2单元207号。依据原《中华人民共和国证券法》（以下简称《证券法》）的有关规定，本会对中兴信托北京亚运村营业部等机构违反证券法律法规行为进行了立案调查、审理，并依法向当事人告知了实施市场禁入措施的事实、理由及依据及当事人依法享有的权利，现已调查、审理终结。经查明，自2000年4月20日起，健富投资在中兴信托北京营业部开设并控制"健富公司"、"华捷经贸"、"华捷发展"、"兴发机械"、"兴发广告"、"国利通达"、"张[redacted]"、"孙[redacted]"、"郭[redacted]"、"孙[redacted]"、"孙[redacted]"、"张[redacted]"等资金账户，下挂个人股东账户842个，买卖证券。上述事实，有相关账户开户资料、客户对账单、资金划转凭据、情况说明等证据在案证实，证据确实、充分，足以认定。健富投资的上述行为违反了原《证券法》第七十四条"在证券交易中，严禁法人以个人名义开立账户，买卖证券"的规定，构成了原《证券法》第一百九十条所述"违反本法规定，法人以个人名义设立账户买卖证券"的行为。根据当事人违法行为的事实、性质、情节与社会危害程度，依据《证券市场禁入暂行规定》第二条第七项等相关规定，本会决定认定当事人梁[redacted]、吴[redacted]、孙[redacted]为市场禁入者，自本会宣布决定之日起5年内不得担任上市公司高级管理人员和从事证券业务。二○○六年十一月二十七日

Figure 3.3: An entry from the Blacklist of "Securities Market Entry Prohibition". The first column, from top to down: the first arrow points to "name of punishment" and the second points to "content of punishment". The table on the right side of the second arrow shows the detailed explanation of the punishment.

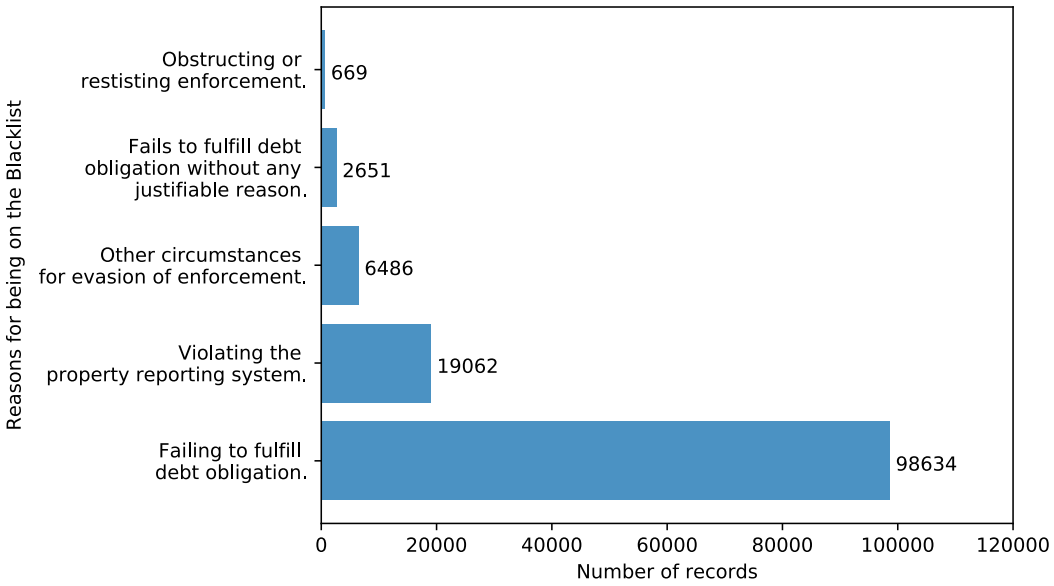


Figure 3.4: The top 5 reasons for being on the Blacklist of company debtors.

数据来源：	高法
数据类别：	失信被执行人名单
案号：	（2018）鲁1092执47号
主体名称：	威海久誉建材有限公司
企业法人姓名：	
组织机构代码：	558932122
执行法院：	威海经济技术开发区人民法院
地域名称：	山东
执行依据文号：	（2016）鲁1092民初1085号
作出执行依据单位：	威海经济技术开发区人民法院
法律生效文书确定的义务：	被执行人给付申请人呢共计3023657.49元
被执行人的履行情况：	全部未履行
失信被执行人具体情形：	有履行能力而拒不履行生效法律文书确定义务
发布时间：	20180525
立案时间：	20180118
已履行部分：	暂无
未履行部分：	暂无
最新更新日期：	2018-06-11

Figure 3.5: Screenshot of a company's Blacklist entry. Left column, the first arrow points to a field explaining the specific context of the case, the second arrow points to the date of publication of this entry on the Blacklist.

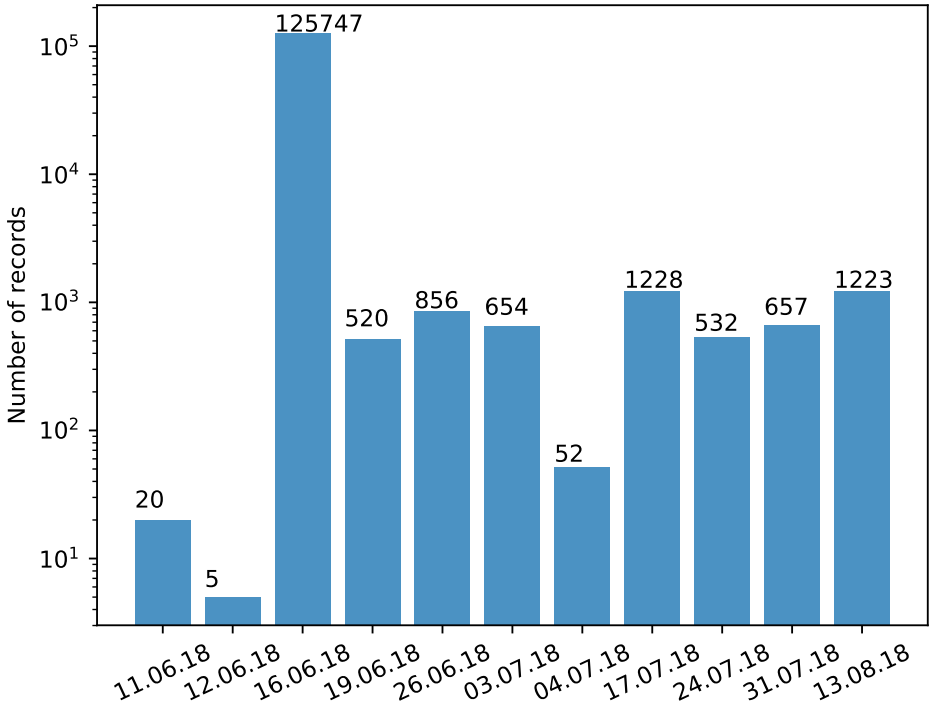


Figure 3.6: Publication dates of Blacklist entries for company debtors.

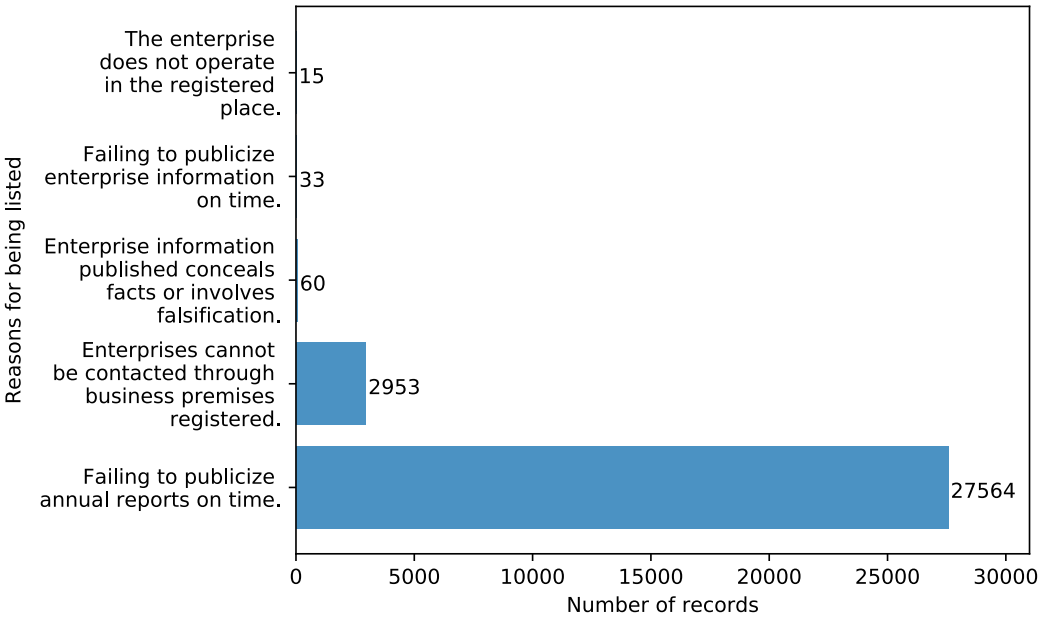


Figure 3.7: The top 5 reasons for companies to be on the Special Attention List.

Blacklist consistently featured the Unified Social Credit Code.

On the national SCS information platform "China Credit", we found another Blacklist issued by the Civil Aviation Administration of China (中国民用航空局)[13]. This list, which is updated every month, publishes information on individuals that are excluded from aircraft travel for a period of one year due to misbehavior on airplanes or airports (data collected on August 10, 2018; see Figure 3.11). According to the list published in August, 2018, 946 individuals were banned from air travel for one year. Among others, the list provided full name, censored ID number, and explanations why individuals had been punished (see three arrows in the first row of Figure 3.11). Being banned from air travel resulted from taking illegal objects on airplanes, smoking on airplanes, or boarding airplanes with a fake passport. The figure also indicates that the list contained names and ID numbers of non-Chinese citizens providing some evidence that foreigners were not excluded from the SCS.

Redlist

We found one type of list documenting information on "good" behavior - the Redlist. It contained a total of 1,206,944 entries distributed across 24 categories (3 categories for redlisted individuals, 21 categories for redlisted companies). The categories for individuals, translated from Chinese, are: 1) Taxi Star (1557 entries), 2) Top Ten Tour Guides (14 entries), and 3) Five-Star Volunteer (603 entries). For all entries, the full name of the person and his or her partially censored ID number were given. The Five-Star Volunteer category displayed the

[13] `https://hmd.creditchina.gov.cn/`, last accessed on November 5, 2018.

主体名称：	赵某
统一社会信用代码：	
处罚类别1：	罚款
处罚事由：	过度疲劳仍继续驾驶的
处罚依据：	暂未入库
处罚名称：	过度疲劳仍继续驾驶的
处罚类别2：	
组织机构代码：	
工商登记码：	
税务登记号：	
法定代表人名称：	
处罚结果：	交通警察总队武清支队对赵文选进行罚款的处罚
处罚期限：	
处罚机构：	交通警察总队武清支队
处罚部门：	
处罚决定日期：	2017/01/03
当前状态：	正常
严重程度：	
地方编码：	120000
创建时间：	2018-07-11 00:00:00

Figure 3.8: A record of the Administrative Punishment register. The first column, from top to down: the first arrow points to the field "type of punishment" and the second points to the field "reasons for punishment".

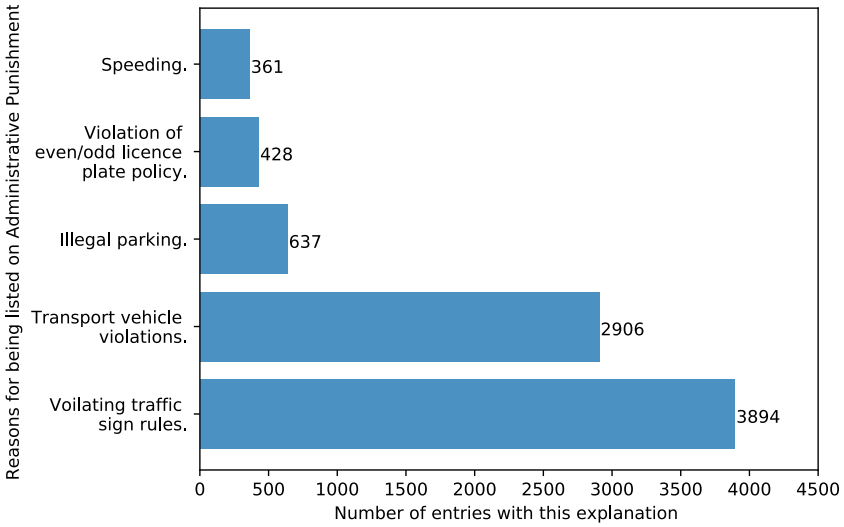


Figure 3.9: The top 5 reasons why individuals or companies are placed on the Administrative Punishment register.

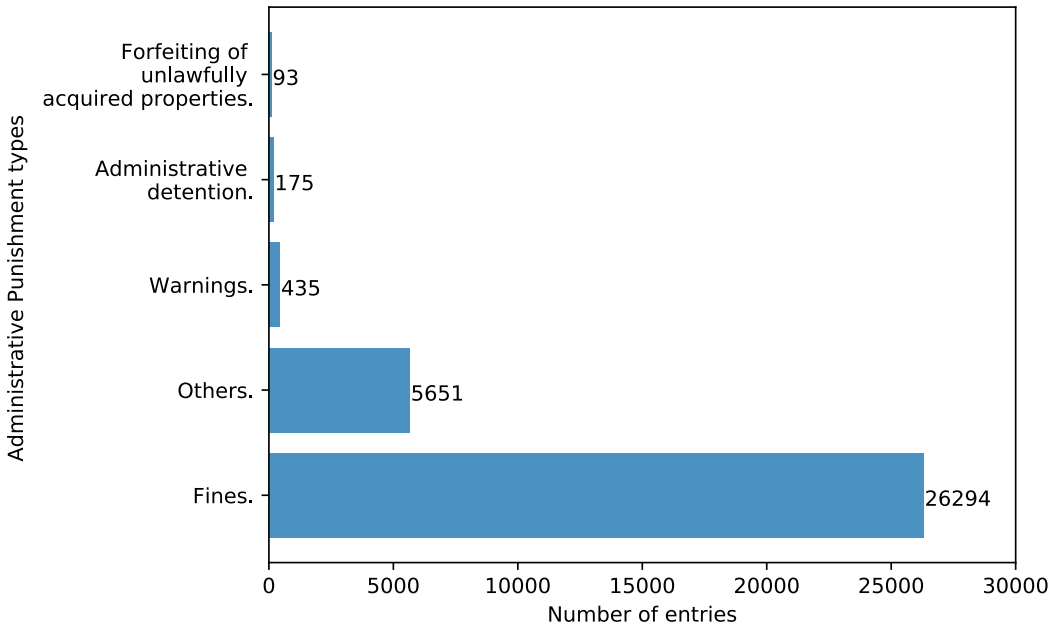


Figure 3.10: The 5 types of Administrative Punishments.

256	刘[redacted]	420281****0919691X	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	湖北省公安厅机场公安局直属分局航站区派出所 行政处罚决定书	鄂机公直航【行罚决】字（2018）第49号
257	姜[redacted]	370103****12182531	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	北京首都国际机场公安分局公安行政处罚决定书	京机公分（法）决字[2018]第296号
258	李[redacted]	340121****0918915X	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	安徽省公安厅机场公安局合肥机场派出所行政处罚决定书	直公（机）行罚决字[2018]64号
259	赵[redacted]	532331****06090635	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	云南省公安厅民用机场公安局直属公安分局 当场处罚决定书	0548462018050702
260	SAURAM[redacted]	AA7242****	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	深圳市公安局机场分局 行政处罚决定书	深公（机）行罚决字【2018】00280号
261	SHOTAYRVNUR[redacted]	N11668****	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	民航新疆机场公安局 行政处罚决定书	民航新机公（侯派）行罚决字[2018]009号

Figure 3.11: A screenshot of the Blacklist for individuals that are banned from flying on commercial airplanes. In the first row, from left to right, the first arrow points a field containing the full name of the individual; the second to censored ID number; and the third to explanations why individuals have been punished. Two arrows at the bottom left indicate entries of two foreign passengers.

gender of the person as well as the amount of volunteering hours carried out per person. The lowest amount of volunteering hours documented was 1500 (which was probably the necessary threshold to be listed) and the highest was 25,400. None of the entries we collected from the Redlist provided an explanation justifying why such a honorary title had been awarded to that person (see Figure 3.13). Thus, we cannot report any observations about justifications on "good" behavior from our Beijing Redlist analysis.

Company categories referred either to tax awards (e.g., A Class Taxpayer) or to other honorable statuses such as Harmonious Labor Relations or Excellent Contributor to Developing Chinese Socialism. Just like the Redlist entries of individuals, there were no justifications explaining why a honorable title had been awarded to a company. No Redlist entry contained the Unified Social Credit Code. Generally, Figure 3.13 shows a single record of an entity that can display several "positive" and "negative" entries. Thus, there is reason to believe that the interface shown in Figure 3.13 functions as the governmental SCS information template: recording and making transparent information on rewards and/or sanctions to the public.

Importantly, every Blacklist and Redlist record we collected featured a "Disagreement/-Correction (异议/纠错)" function (see Figure 3.12). This function allowed citizens to object to a Blacklist or Redlist decision by providing a statement of up to 2000 Chinese characters (submission required 18-digit ID number).

Coding results for "positive" cases on "good" behavior

News reports on "good" behavior were introduced as "Stories of Integrity (诚信人物/故事)" posted under the section of "Integrity Culture (诚信文化)" on the national SCS information platform "Credit China". All of the 20 "positive" cases selected described how a protagonist sacrificed his or her self-interest (both material and non-material) for the common good. Moreover, all cases centered on "trustworthiness" and "honesty" as key SCS virtues. The stories all followed the same narrative structure: they first provided detailed biographical information of a person (full name, social class, profession, family status), followed by a dilemma: the protagonist could either engage in "dishonest" behavior winning him or her an immediate small reward or get a large future reward by being "honest". Once the person had enacted the "honest" behavior, which happened in all the "positive" reports we analyzed, the narratives ended with a virtue cascade.

Take, for example, cases in which individuals found and returned lost property to an owner. Here, all four cases assigned to the topic"return lost property to owner" ended by further attributing "self-discipline", "helpfulness", "care-taking for others", and a "sense of responsibility" to the protagonist as part of a virtue cascade. Another commonality across the selected cases was that all protagonists were morally "praised" by their social environment. Also, the protagonist was recognized for his or her "good" behavior by official agencies or the media in the form of "honors", "decorations", or a "cute nickname". On the other hand, when a material reward was offered for the "good" behavior, as in all cases with topics "family and community relationship and repayment", "return lost property to owner", and "social entrepreneurship to help people out of poverty", the protagonist refused the material reward

Figure 3.12: Example of a company's Blacklist entry. The black circle on the upright corner indicates the "Disagreement/Correction (异议/纠错)" function.

Figure 3.13: Example of a Redlist entry for an individual with the honorary title Five-Star Volunteer. The record does not justify why the honorary title was awarded.

at all times.

Coding results for "negative" cases on "bad" behavior

Reports about "bad" behavior were labeled as "Typical Cases (典型案例)" on the homepage of "Credit China" with the sources being both local newspapers and the platform itself. The 26 selected "negative" cases relating to 7 topics all revolved around one common theme, the "Laolai (老赖)": a term specifically referring to individuals and companies refusing to repay debts. These cases were presented in two ways. The 4 cases with the topic "public shaming" were about the courts' actions in solving repayment problems. The remainder of the stories were about specific individuals or companies. All individuals and companies were anonymous in the selected cases. Local courts collaborated with local telecommunication companies in all 4 cases with the topic "public shaming", and the Public Security Bureau played an important enforcement role in all cases with topic "public transport regulation violation". In these reports, both the compulsory actions taken by the court and the threat of being placed on the Blacklist forced the "Laolai" to fulfill the stated obligation. Generally, both "positive" and "negative" case studies we analyzed were homogeneous in structure, framing, and content. This could indicate that they had been deliberately formulated to propagate the SCS's conceptualization of "good" and "bad" behavior.

3.1.5 Analysis

The results of our content analysis demonstrate that there are currently multiple informational asymmetries in both datasets.

Listed companies versus listed individuals

Currently, companies make up the majority of entries on both the Blacklist and Redlist of Beijing's SCS platform. We found that companies which are involved in the construction of the SCS were also included in the list. For instance, Alibaba (with Zhima Credit) and Tencent (with Tencent Credit) were both granted permission to start individual pilot credit service programs in 2015 and have provided digital data collected from online shopping and social media to the SCS. Both Alibaba and Tencent were listed as A-level Taxpayers on the Redlist. Since we only crawled the Beijing SCS platform, we cannot make any claims about the transparency of other SCS Blacklist and Redlist websites.

Our analysis of "positive" and "negative" cases demonstrates the opposite: here, the majority of reports on either "good" or "bad" behavior focuses on individuals' behaviors. For our manually coded sample, only 15.4% of "negative" reports and 30.0% of "positive" reports featured companies. In both "negative" and "positive" cases that featured companies, however, reports centered on the person in charge of the company typically highlighting the CEO's virtues and vices. In other words, it is not the company as such that is "blamed" or "praised," but rather the person responsible for the company. Such portraits, therefore, signal that

individuals are not shielded by large institutions but can be made responsible for their "good" or "bad" decision-making.

Justifying punishments versus justifying rewards

All entries of the Blacklist explain why a person or company is currently registered on the Blacklist. Moreover, Blacklist explanations include legal terms and refer to laws and regulations. In other words, Blacklist explanations make transparent the mechanism of punishment by specifying a causal link between behavior and consequence. This is perhaps best illustrated by the Blacklist on individuals excluded from air travel (see Figure 3.11). The legal threat contained in the entries of the Blacklist could furthermore signal that a specific "dishonest" behavior can be detected and sanctioned.

On the other hand, not a single entry of the Redlist includes a formulated explanation on why a person or company has been awarded a honorary title. We found that fulfilling legal obligations (Class A Taxpayer), performing professional (Taxi Star) or volunteering (Five-Star Volunteer) duties can result in reputational gain in the current SCS. However, the mechanisms or criteria determining when an individual or a company secures a place on the Redlist are not further explained. Taken together, the current SCS makes behaviors leading to punishments more transparent than behaviors resulting in rewards. More generally, our study could not identify publicly available information associating specific behaviors to a scoring or rating mechanism.

Types of punishments versus types of incentives

The most common reason for a company to be placed on any of the "negative" lists is failure to pay back debt (the second most common reason is informational misconduct). Failure to pay back debt is also the most prominent reason given for why protagonists of the "negative" cases are registered on the Blacklist. The Chinese term for "Laolai" appeared 481 times in the 789 "negative" reports we collected. All "negative" stories we manually coded report on the activities of a "Laolai" person (either as an individual or as the legal representative of a company). In terms of punishment, individuals and companies face both the material loss specified in the corresponding legal regulation as well as the consequences of being publicly shamed on the Blacklist. In more than 40% of the narratives on "negative" behavior, an individual is threatened to be placed on the Blacklist leading to the immediate compliance of the individual.

On the other hand, individuals and companies on the Redlist receive moral "approval" and reputational gain. Similarly, "positive" cases report on individuals that gain reputational rewards, while at the same time rejecting material incentives when offered as a consequence of their "role-model" behavior. Still, being listed on the Redlist is not mentioned or even indicated by any individuals as a motivational factor for their behaviors. All stories we analyzed emphasize that a morally "praiseworthy" activity is "praiseworthy" when it is "genuinely" moral rather than instrumental in obtaining a material reward. Furthermore,

all "positive" stories feature a virtue cascade: once an individual is described as "genuinely honest" or "trustworthy", he or she is attributed other "positive" virtues as a consequence.

3.1.6 Discussion & Concluding Remarks

In this first study of key websites of the Chinese SCS, our goal was to understand how transparent the SCS currently is in providing information on "good" and "bad" behavior. To this end, we collected and analyzed 194,829 Blacklist and Redlist entries from the Beijing SCS website "beijing.gov.cn/creditbj" and applied a machine learning topic modeling algorithm to almost 1000 reports on "positive" and "negative" behavior crawled from the national SCS information platform "www.creditchina.gov.cn". Finally, we manually coded a sample of these texts to understand what kind of specific behavioral information they contain.

The main question arising from our findings, we believe, is whether the degree of the current SCS's transparency is intentionally engineered or whether it is simply a manifestation of work in progress. Is there a purpose in explicitly describing and publishing the causal link between behavior and sanction while leaving information on getting rewards deliberately vague? First, the asymmetries in information provided between the Redlist and the Blacklist could be motivated economically: while an infinite amount of people can be excluded from valuable material resources, only a finite amount can be given valuable resources (e.g., a first-class train ticket). Detailed instructions on how to win rewards could therefore lead to distribution problems since many individuals could implement them. On the other hand, another explanation for the current informational asymmetries of the SCS might be that already existing records of legal offenses were used to start filling Blacklists. Consequently, these records entail more justifications since they refer to specific legal articles or regulations.

The degree of transparency of the SCS observed in this work could also be motivated by behavioral engineering goals. Let's imagine for the moment the system were completely inscrutable (i.e., the system did not justify a score increase or decrease and eventually a given punishment or reward, respectively). In this case, individuals would have little possibility to understand when the SCS rewarded and when it sanctioned specific types of behaviors. Moreover, besides being oblivious to the moral code of conduct, individuals would not have the ability to contest the system's decision-making process (again, to negotiate a norm one must have the necessary epistemic resources to do so). Note that this issue is also debated in the context of the "Right to Explanation" of the European Union's General Data Protection Regulation [142, 143]. A fully transparent scoring system, on the other hand, would precisely map behaviors to rewards or sanctions. Indeed, in the context of a nationwide digitally-implemented scoring system, full transparency must account for the mechanism that leads to the distribution of rewards or sanctions. This degree of transparency would offer individuals the possibility to understand the system's decision-making procedures at least to a certain extent. In our analysis of SCS Blacklist and Redlist records, we did not identify an explicit SCS scoring mechanism. We have shown, however, that the SCS already enables citizens to dispute single Blacklist and Redlist records. On the other hand, a fully transparent SCS would possibly create other problems: if the SCS became fully transparent in regard to its

scoring mechanisms, complying to a norm would likely become a market transaction. In fact, research on intrinsic and extrinsic motivation suggests that introducing an external reward to a norm-guided behavior turns this behavior into a commodity that can be bought [144, 145]. This phenomenon, termed "crowding-out effect", results in fewer people engaging in this behavior since the consequences of failing to act can simply be compensated by financial means [146, 147, 148]. For example, if one reliably receives monetary compensation for being honest, being honest will no longer be evaluated as a moral behavior for both the actor and the recipient. As this line of research suggests, individuals will likely stop attributing a genuine moral character to individuals with a high score in a fully transparent SCS.

Our analysis provides evidence that the currently implemented SCS possibly attempts to counter such a transformation of moral behavior into market transactions. All of the "positive" case studies unambiguously emphasize that norm conformity is "good" because it is "morally valuable" – for both average citizens as well as CEOs. None of the Redlist entries describe a connection between moral behavior and external material reward. Rather, they contain virtue signals and reputational gains by awarding symbolic honorary titles (e.g., Five-Star Volunteer). On another sub-page of the national SCS website, we found the publication of 32 ancient Chinese fables (not shown) also promoting self-concepts comprising virtues of being a morally "good" Chinese citizen. In contrast, our analysis on the corpus of "negative" case studies demonstrates the propagation of a "negative" self-concept ("Laolai") attributable to a specific offense (i.e., intentionally not paying back debt). Taken together, our analysis suggests that degrees of transparency can serve different behavioral engineering goals in the context of a digital scoring system.

3.2 Research Article 2: Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness

Authors

Severin Engelmann, Mo Chen, Lorenz Dang, Jens Grossklags

Publication Outlet

AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; July 2021; Pages 78–88; https://doi.org/10.1145/3461702.3462535

Abstract

The Chinese Social Credit System (SCS) is a novel digital socio-technical credit system. The SCS aims to regulate societal behavior by reputational and material devices. Scholarship on the SCS has offered a variety of legal and theoretical perspectives. However, little is known about its actual implementation. Here, we provide the first comprehensive empirical study of digital blacklists (listing "bad" behavior) and redlists (listing "good" behavior) in the Chinese SCS. Based on a unique data set of reputational blacklists and redlists in 30 Chinese provincial-level administrative divisions (ADs), we show the diversity, flexibility, and comprehensiveness of the SCS listing infrastructure. First, our results demonstrate that the Chinese SCS unfolds in a highly diversified manner: we find differences in accessibility, interface design and credit information across provincial-level SCS blacklists and redlists. Second, SCS listings are flexible. During the COVID-19 outbreak, we observe a swift addition of blacklists and redlists that helps strengthen the compliance with coronavirus-related norms and regulations. Third, the SCS listing infrastructure is comprehensive. Overall, we identify 273 blacklists and 154 redlists across provincial-level ADs. Our blacklist and redlist taxonomy highlights that the SCS listing infrastructure prioritizes law enforcement and industry regulations. We also identify redlists that reward political and moral behavior. Our study substantiates the enormous scale and diversity of the Chinese SCS and puts the debate on its reach and societal impact on firmer ground. Finally, we initiate a discussion on the ethical dimensions of data-driven research on the SCS.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

3.2.1 Introduction

In 2014, the Chinese government published the *Planning Outline for the Construction of a Social Credit System (2014-2020)* as part of its 12th five-year plan [116]. Following its release, media and research have offered various perspectives on the Chinese Social Credit System (SCS, 社会信用体系). Some Western media have characterized the SCS as a mass surveillance apparatus, with the purpose of calculating a digital "sincerity score" for each Chinese citizen based on a wide range of personal data [117, 149, 150]. Below a certain point level, citizens would face multiple restrictions, such as exclusion from air travel and high-speed trains. A positive score, on the other hand, would lead to discounts and preferential treatment for a variety of products and services. This "dystopian perspective" sees the unification of an authoritarian regime's policies and artificial intelligence (AI) to enforce social order by means of a sincerity score. Some media outlets have since revised their original viewpoints regarding such comprehensive sincerity scoring [151, 152].

Academic scholarship on the SCS has largely been theory-driven, which has led to the independent development and discussion of different conceptualizations. The SCS has been defined as a novel administrative policy program with the main goal of strengthening compliance of citizen and organizations with laws and regulations [153, 154]. The novelty consists in the public (at least temporary) disclosure of already existing citizen and organizational records on so-called digital blacklists and redlists. Blacklists publicly showcase non-complying individuals and organizations, while redlists, as their normative counterpart, show complying entities. In this perspective, the SCS deploys reputational tools with some similarity to company rankings or background checks on individuals in Western economies.

Other authors have called the SCS a big data empowered system that collects, processes, and evaluates vast amounts of personal data [127]. These data are ultimately aggregated and published as public credit information (PCI) on digital platforms. This line of research argues that PCI creates transparent citizens, not least due to the lack of a sufficient legal framework that protects personal data in China [155]. Some scholars have noted an all-encompassing application of credit to society's political, economic, and social activities. Thereby, the SCS marks the emergence of a so-called reputation state [156, 157]. As a governance tool, the SCS seeks to harness reputational information for purposes that go beyond neoliberal notions of regulating market failure. Still other perspectives frame the SCS as a social management program [158]. Drawing on concepts from systems engineering, a social management program considers society to be a complex system that can be optimized using digital technologies.

While these accounts disagree in many important regards, three points of agreement can be identified: first, multiple independent initiatives have been labelled as "SCS" [159]. One SCS is driven by the apps and services of big data companies (e.g., Sesame Credit) that distribute scores to consumers in voluntary promotion programs [129, 160]. Here, "voluntary" denotes consenting to the terms and conditions of the service. Second, *local* governments have tested SCSs that integrate different scoring systems in "prototype cities" (社会信用体系建设示范城市), such as Rongcheng and Suzhou. Participation in these local "credit scoring experiments" is mandatory for residents in these areas. Such policy experiments [161] can

serve as models for other local SCS implementations but they are not necessarily a model for national implementation. Third, government-led SCS measures have been realized nationally. There are various types of blacklists (黑名单) and redlists (红名单) run by government agencies at different levels of administrative divisions (ADs) including municipalities and provinces, but also government departments at the national level. These platforms publicly display information to "shame"[14] or "praise" natural and legal persons (e.g., companies) for non-compliance or compliance with a variety of legal and social norms [162, 163, 132, 155, 164]. No entity can opt out from being listed. Depending on the type of list, entities are subjected to different types of reward or punishment over a wide range of areas, a process that has been termed "joint reward and punishment mechanism" (JRP) by the Chinese government [116]. Both natural and legal persons on specific blacklists or redlists will be punished or rewarded under the rules defined in Memoranda of Understandings (MoUs). Different government agencies have jointly signed and started enforcing these MoUs [165].

To summarize, the government-run SCS operates blacklists and redlists throughout the entire country. It enforces regulations with reputational and material means and requires mandatory participation. *This* SCS has regulatory "teeth". However, no research has conducted an empirical analysis of this nationwide SCS blacklist and redlist infrastructure.

This lack of knowledge is troubling, as the SCS will likely shape the behavior of about 1.4 billion Chinese citizens and all companies doing business in China. Further, important international long-term technology policy challenges are dependent on the success of systems such as the SCS, as highlighted by Antony Blinken in his confirmation hearings, when he argued that "*whether techno democracies or techno autocracies are the ones who get to define how tech is used (. . .) will go a long way toward shaping the next decades*" (2021 U.S. Secretary of State confirmation hearings [166]).

This study investigates the design and technical implementations as well as the number and types of blacklists and redlists across 30 Chinese provincial-level ADs. Our exploratory study shows the diversity of SCS lists in granular detail and outlines the informational consistency between social credit records of the same type of list on different SCS platforms. We find that SCS listings focus on economic activities but also capture reputational rewards for moral and political behavior. Moreover, we show that the SCS listing infrastructure is flexible, as observed in a second round of data collection during the COVID-19 outbreak: when necessary, new types of lists can regulate novel forms of transgression and thereby help accomplish new policy goals.

Figure 3.14: Screenshot of an overview of the SCS information platforms of the different ADs listed on the national SCS platform "creditchina.gov.cn". Taiwan, Hong Kong and Macao were previously listed together with other ADs on the landing page of the "Credit China" website, but without a valid link. The listings were then removed in July 2019. Data collection was conducted via the SCS platform of each AD. Color-coding: orange represents municipality under the direct administration of central government; blue represents provinces; purple represents autonomous administrative regions; green represents the Xinjiang production and construction corps (Bingtuan), an economic and paramilitary organization in the Xinjiang Uyghur Autonomous Region, which is not included in our analysis due to an insignificant amount of credit data. Translations of AD names added by the authors.

3.2.2 Study procedure

Policy-making in China: Provinces implement blacklists and redlists

SCS implementation is largely left to regional rather than central government, a common trait of China's policy-making process that tends to follow a principle of "centralized planning, decentralized implementation" [167, 168]. As a planning polity, central policy-makers outline policy goals in top-level policy documents valid for a specific policy-making cycle. Commonly, a first policy document (called *jianyi*/建议) includes general guidelines for a new cycle of policy-making. A second, more refined, but still broad, policy outline (called *gangyao*/纲要) sets more specific policy goals [161].[15] Importantly, the *implementation* of the policy goals outlined in top-level policy documents is left to provincial, county, and city governments. This also applies to the SCS: provincial-level administrative authorities (i.e., those in charge of provinces, municipalities under the direct administration of central government, and autonomous regions) are, to some extent, free to determine *how* they implement nationwide policy goals for their AD [169, 170].

The SCS's *gangyao* includes vague instructions regarding social credit record applications for broadly defined commercial and social sectors (e.g., [127, 165, 155]). SCS implementation rests on the commitment of provincial-level ADs[16] to realize general instructions laid out in

[14]The authors use quotation marks to communicate a neutral standpoint towards SCS-specific normative concepts (e.g., "positive", "negative", "reward", "sanction/punishment"). For the remainder of the article, quotation marks will be omitted for the sake of reader-friendliness.

[15]Generally, policy-making in China is accompanied by a multitude of other policy documents. Engaging in a comprehensive description of Chinese policy-making would go beyond the scope of this study.

[16]In China, provincial-level ADs comprise provinces (e.g., Sichuan), municipalities under the direct administration

top-level policy documents. As such, understanding the nationwide SCS listing infrastructure requires an empirical assessment of all SCS platforms at the provincial level. As each province is responsible for the implementation of its own SCS blacklist and redlist, we expected to find differences in the technological setup, interface design, and list types (i.e., differences in types of rewards and sanctions) between the provincial-level SCS platforms.

We conducted two rounds of data collection. First, between June 2019 and December 2019, we collected data on blacklists and redlists from 30 Chinese provincial-level ADs comprised of 22 provinces, 5 autonomous regions and 4 municipalities under the direct administration of central government. Second, in February 2020, we started collecting data on blacklists and redlists related to the coronavirus outbreak.

As we describe in more detail in the methodology section, our study approach is fundamentally *exploratory*. Data collection and analyses were intended to understand SCS implementation with regard to three high-level research questions, as follows.

- `RQ1`: Are there technological and design differences in credit lists and records between the provincial SCS platforms?
- `RQ2`: How do provincial SCS platforms differ in the number and types of blacklists and redlists?
- `RQ3`: How do SCS blacklist and redlist *records* of the same type of list differ in terms of the information displayed across provincial SCS platforms?

Methodological approach

Data

Our analysis pertains to blacklists and redlists implemented at the AD level from June 2019 to December 2019. Data collection was aimed at provincial-level blacklists and redlists from 31 ADs (22 provinces, 5 autonomous regions, 4 municipalities under the direct administration of central government) listed on China's national SCS platform "creditchina.gov.cn" (Figure 3.14).[17] For the follow-up study of coronavirus-related lists, we inspected the same SCS platforms again between February 2020 and April 2020.

Data collection primarily refers to a) the types of lists implemented in each AD (RQ2) and b) retrieving individual credit records from the most commonly implemented blacklist and redlist across all 31 ADs (RQ3). Collecting list types and credit records enabled an analysis of the technical realization and interface designs of SCS platforms and credit records (RQ1).

of central government (e.g., Beijing, Shanghai) and autonomous regions (e.g., Inner Mongolia, Tibet).

[17] This list also included the Xinjiang production and construction corps (Bingtuan). However, we did not include these data in our analysis for two reasons: first, Bingtuan is a unique state-owned economic and paramilitary organization in Xinjiang and, second, at the time of data collection, Bingtuan's SCS platform had published only a very small amount of credit information (9 blacklist and 7 redlists entries).

Our data collection was organized to produce a *descriptive* study of SCS implementation. Our core analyses focused on the diversity of list types across ADs and the structural differences between list records, in particular, their interface designs and the information provided in individual credit records. For several reasons, we did not conduct a quantitative analyses on published records. First, during data collection, we observed that the number of published SCS records changed on a day-to-day basis for all SCS platforms. We refrained from drawing general inferences on SCS credit records based on a onetime quantitative analysis. Second, when we began to scrutinize different SCS platforms, we observed large differences in the amount of credit records uploaded. Some SCS platforms had not published any credit records, while some displayed multiple millions (note that only a few SCS platforms indicated the total number of credit records). Third, given the early stage of SCS development, a comprehensive quantitative analysis of the economic and societal impacts of credit records was not possible at the time of data collection. This impact may need several years to materialize as SCS measures begin to influence the economy, government administration, and social processes at large. Fourth, as we discuss in the next subsection, we encountered challenges in accessing and retrieving public credit information from SCS platforms.

Data collection obstacles

The first obstacle was obtaining access to the 31 AD SCS platforms. Access from our location was severely impeded, so we tested the accessibility of different SCS websites from various locations. To accomplish this, we sent web requests from 44 servers spread around the world to each AD's SCS website.[18] SCS server accessibility from outside China was generally possible but unstable.[19] To investigate SCS platforms, we used a virtual private network of servers located in China. Requests from China provided more stable access to SCS servers than from other locations. All SCS servers, apart from the SCS server of the municipality of Chongqing, responded to requests from a Chinese server. For the server of the municipality Chongqing, no data could be retrieved at any time, as the server did not respond to requests for the entire data collection period from any location. Thus, our final data collection represented 30 ADs. Overall, it took 6 months to access all SCS platforms and to document the different types of blacklists and redlists, verify them through revisits, and collect credit records for each AD.

While documenting the different types of lists for each province, we observed that each AD operated a different web server with different implementations of front-end, back-end and database design. Moreover, we did not find a public API on any of the AD SCS platforms. Taken together, this made data collection for *credit records* complicated, as each AD SCS platform required the programming of a unique web crawler and scraper.

The systematic sampling of public credit records from each blacklist and redlist on all SCS platforms was not possible for several reasons. First, the number and therefore types of lists implemented varied between the ADs. Some ADs had more than 10 types of lists,

[18]The analysis was conducted with the Uptrends online monitoring service (`www.uptrends.com`). Data available from the authors.

[19]The most frequent return values were: HTTP connection failure, HTTP protocol error, HTTP timeout, and TCP connection failure.

while others only displayed a single list (see Results). We saw that some ADs with only a single implemented blacklist or redlist used this list to present different types of sanctions or rewards. Second, some ADs had only one list but no records to show at all. Third, SCS platforms differed in how credit records were displayed. For example, some SCS platforms displayed a number of credit records on a single page and offered page tabs that opened the next page, displaying the next set of credit records. This interface style allowed page visitors to go through all available credit records. Other SCS platforms only showed a selection of credit records and instead of page tabs provided a search bar for specific queries. Here, visitors could not see all available credit records. Finally, some AD SCS platforms deployed captchas and bot blockers that sometimes led to time-out denials such as temporary or even permanent IP address suspension.

Given these restrictions on the collection of credit records, systematic and unbiased sampling of credit records across all SCS platforms was not possible. However, the goal of our study was not to measure effects between credit record samples to generalize to the SCS as a single system. Instead, for the credit record analysis, our research goal was to explore informational differences in credit records across the SCS platforms. For this purpose, homogeneous convenience sampling was sufficient to compare the information provided on credit records on the same list between SCS platforms. Homogeneous convenience sampling differs from conventional convenience sampling by constraining sampling by one factor (see e.g., [171]). We did not sample any credit record on any type of list (i.e., we did not conduct conventional convenience sampling). We directed the analysis of credit records toward the most frequently implemented type of blacklist and redlist across all SCS platforms. Consequently, different crawling and data extraction (scraping) robots were programmed to extract pre-specified information on credit records from the most common type of blacklist and redlist.[20] The two main frameworks and tools used for the crawling and scraping process were ThoughtWorks Limited open source headless browser Selenium and Scrapinghub Limited open source framework called Scrapy. The extracted data were eventually pushed into a noSQL database (MongoDB) as a horizontally scaling non-relational database was the better solution given the different SCS platform implementations.

Finally, the obstacles described above naturally led to credit record samples of varying size. On some SCS platforms, we managed to retrieve thousands of public credit records. On other platforms we obtained less than a hundred; some platforms did not have *any* credit records at all during the entire data collection period (for an overview of sampling results, see Table 2 in the Auxiliary Material). The differences in sample size were not due to any systematic sampling error committed by us but reflected the arbitrariness of the credit record display across the SCS platforms during the data collection period.

[20]We provide a code example of a crawler and a spider in the Auxiliary Material.

3.2.3 Results

Technical implementation and design of blacklists and redlists

Each SCS platform operated a different web server with its own front-end, back-end and database design. We observed that the designs of the blacklists and redlists differed between ADs but was, overall, simple and plain.

All SCS platforms implemented either a Hypertext Markup Language (HTML) document with classic Cascading Style Sheet (CSS) structure or advanced dynamic scripting technology (JavaScript) for lists and individual records.

The majority of ADs (21) displayed only a selection of records but enabled targeted queries via a search bar. The remaining ADs showed all available social credit records with the help of a page tab. For example, on Guangxi's SCS platform, blacklist records could be accessed via 6852 tabs, each displaying 10 records. By contrast, Shanghai's blacklists showed ten blacklist records with no option to access more entries other than with a targeted query (Figure 3.15).

The design differences extended to individual credit records. Blacklist and redlist records were either structured as two column tables (Figure 3.16), multiple column tables (Figure 3.17) or continuous text documents.

Inner Mongolia and Shandong enabled sharing of blacklist and redlist records through Chinese social media platforms (e.g., Wechat, Sina Weibo, and Baidu Tieba). We found that eight SCS platforms offered citizens and organizations the possibility to contest published social credit records via a standardized interface option (e.g., Figure 3.16 top right corner).

Our data indicate that there are technological and design differences in credit lists and records between provincial SCS platforms (RQ1). The current design and implementation of SCS platforms prioritize the display of social credit records rather than any aspect of their reputational effects. All SCS platforms had a binary rating system for good and bad behaviors – redlists and blacklists. Other than this binary classification, however, ADs did not apply other rating measures, such as numerical or continuous scoring. Indeed, we did not observe any social credit score at all communicated on any provincial-level SCS platform across China. Different types of lists were not put into relation with each other by means of a sorting or ranking. For example, no system of reputational ordering was found between individual records that highlighted severe transgressions more prominently than less severe cases. Five ADs showed numerical aggregation when a citizen or company had multiple social credit records. Entities with additional record entries were not displayed more prominently than entities that had a single credit record entry. Currently, the design of the SCS lists serves as a digitally accessible repository for citizen and company records and does not use any advanced features characteristic of other digital reputation systems [172].

Diversity and comprehensiveness: Number and types of blacklists and redlists

In response to RQ2, our data provide evidence for substantial differences in the number and types of lists between ADs (compare Figures 3.18 & 3.19). This confirms that regional governments determine the number and types of blacklists and redlists for their administrative

Figure 3.15: Shanghai's "Dishonest legal persons subjected to enforcement" (Lao Lai) blacklist of companies only displayed 10 record entries, requiring visitors to make a targeted search query. Translations by the authors.

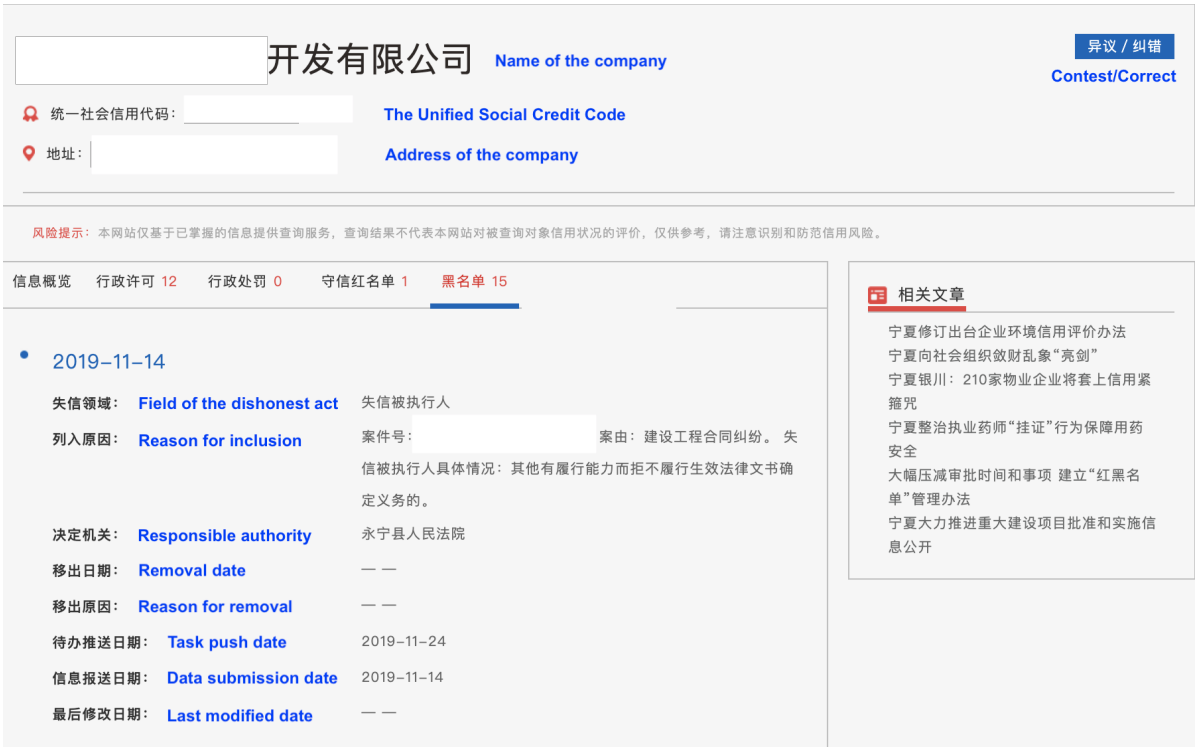


Figure 3.16: A two-column example credit record of the "Lao Lai" blacklist published on Ningxia's SCS platform. Translations by the authors.

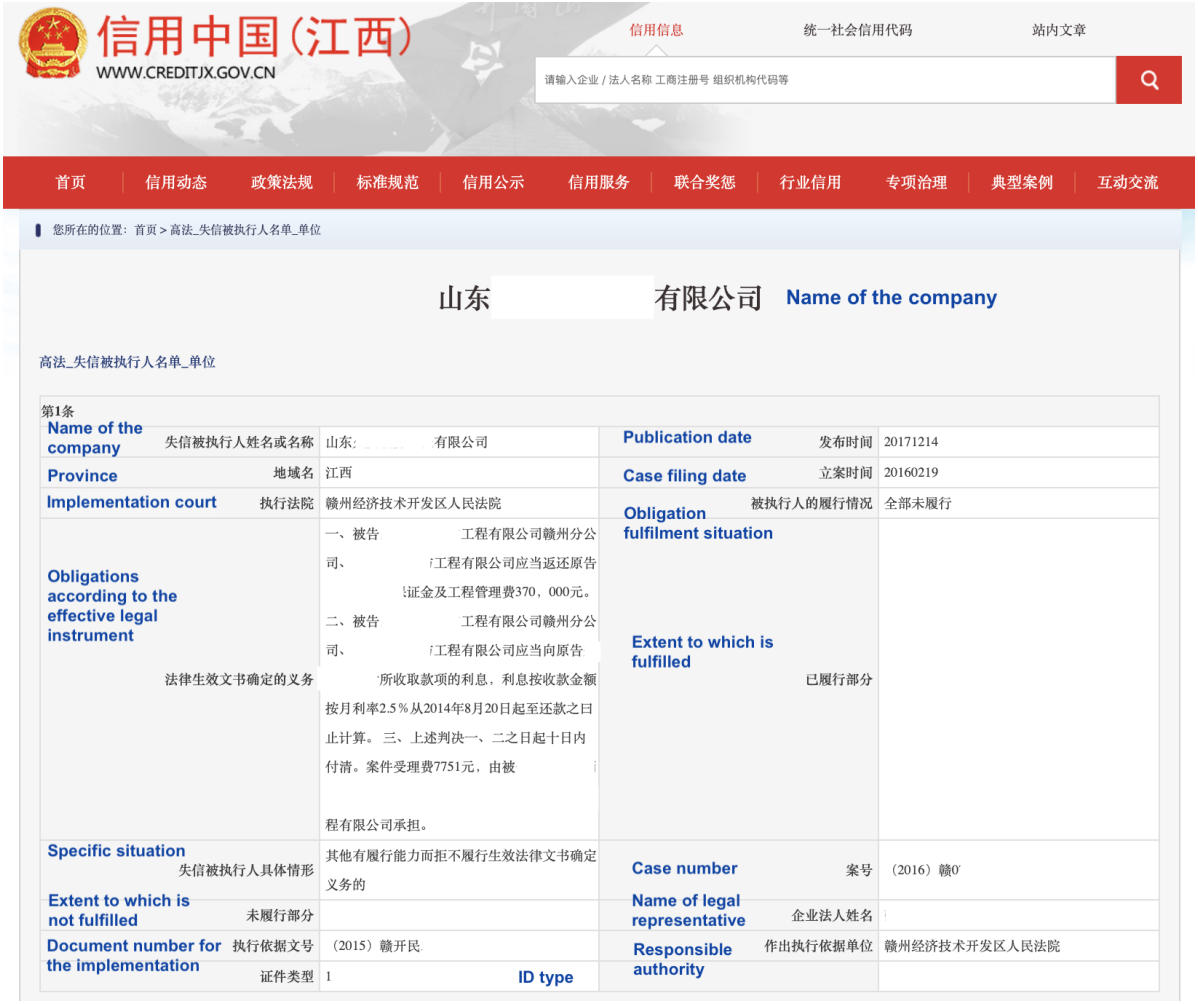


Figure 3.17: A multi-column example record of Jiangxi's "Lao Lai" blacklist (失信被执行人名单). Translations by the authors.

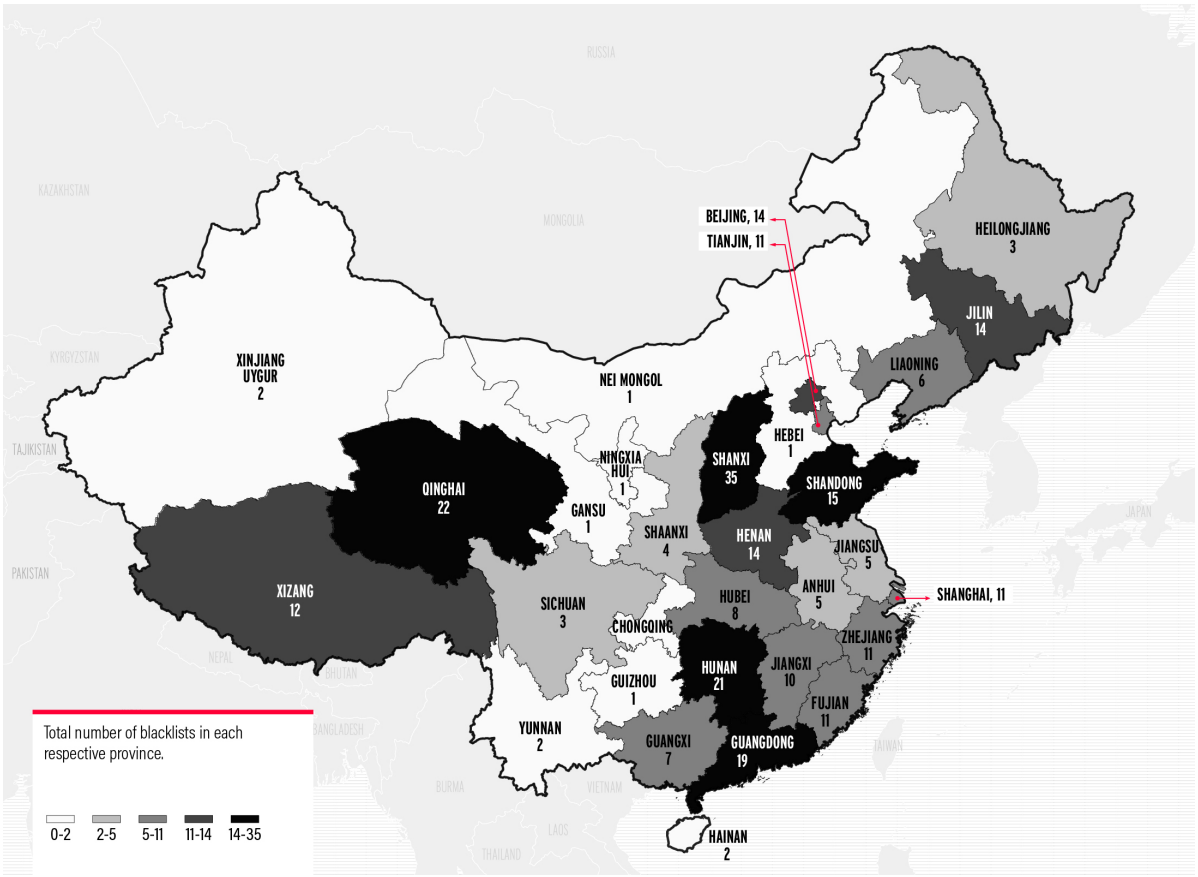


Figure 3.18: The number of blacklists implemented across 30 ADs. Shanxi had implemented most blacklists (35), followed by Qinghai (22), Hunan (21), Guangdong (19) and Shandong (15).

region. For example, Beijing, Tianjin, Tibet, Guangdong, Hunan, Shanxi and Qinghai each operated more than ten different types of blacklists and redlists. In contrast, Inner Mongolia, Ningxia, Gansu, Guizhou, and Hebei each had implemented only one blacklist *and* one redlist. At present, it is impossible to say why some ADs run multiple lists and some only a single list. The number of lists did not correlate with economic, demographic, or geographic factors (data not shown).

In total, more blacklists (273) were published than redlists (154). We first grouped the 273 blacklists into 41 categories and the 154 redlists into 45 categories. We then created a taxonomy consisting of eight types of blacklists and eight types of redlists that currently make up the entire SCS AD listing infrastructure (Table 3.3). Note that different types of lists emphasize compliance with the legal and social norms that an AD wants to improve on. Thereby, the SCS influences behavior through two common reputation strategies [173]. With a minimum threshold strategy, blacklisting stresses the need for conformism. This technique tries to bring all entities to the same level of compliance. Redlisting, on the other hand, highlights praiseworthy performers that are intended to serve as behavioral role models.

The majority of blacklists displayed companies and citizens that have not fulfilled a court order, have committed commercial or transactional fraud, or have not complied with specific industry regulations. *All* ADs had implemented a "List of Dishonest Persons subject to Enforcement" also called the "Lao Lai" blacklist. This blacklist published information on citizens and companies that have failed to fulfill a court order. The "Lao Lai" blacklist aims to tackle China's court order enforcement problem [156, 165]. It forms a critical part of the JRP by which listed citizens face multiple restrictions, such as being banned from taking flights and high speed trains. Restrictions for "Lao Lai" companies include denial of licenses, reduced possibility to win bids for public contracts, or being subject to additional requirements for mandatory government approval for investments in sectors where market access is usually not regulated. Beyond the "Lao Lai" blacklist, we did not find any other type of blacklist implemented on all SCS platforms. The other types of blacklist most commonly found targeted non-compliance in tax payment (12 out of 30 ADs), untrustworthy behavior in financial activities (9/30), illegal import or export of products (8/30), delay or failure to compensate migrant[21] workers (8/30, companies only), or failure to protect the environment (7/30, companies only). We found blacklists that sanctioned fraud in marriage registrations or charity donations (social fraud), companies that had failed to comply with product quality standards (especially in food and drug production), or companies that had bad employment relationships.

The most frequently implemented redlists displayed entities that complied with tax law (18 out of 30 ADs) and import and export regulations (10/30). Usually, redlists serve to reward particularly "praiseworthy" behaviors. We made the surprising observation that many types of redlists highlighted regular compliance with laws and regulations. Some redlists, however, showcased individuals and companies that distinguished themselves politically or morally. For example, Beijing's SCS platform published a list called "4th Beijing Excellent Builders of

[21] "Migrant" here refers to rural citizens moving into urban centers for employment.

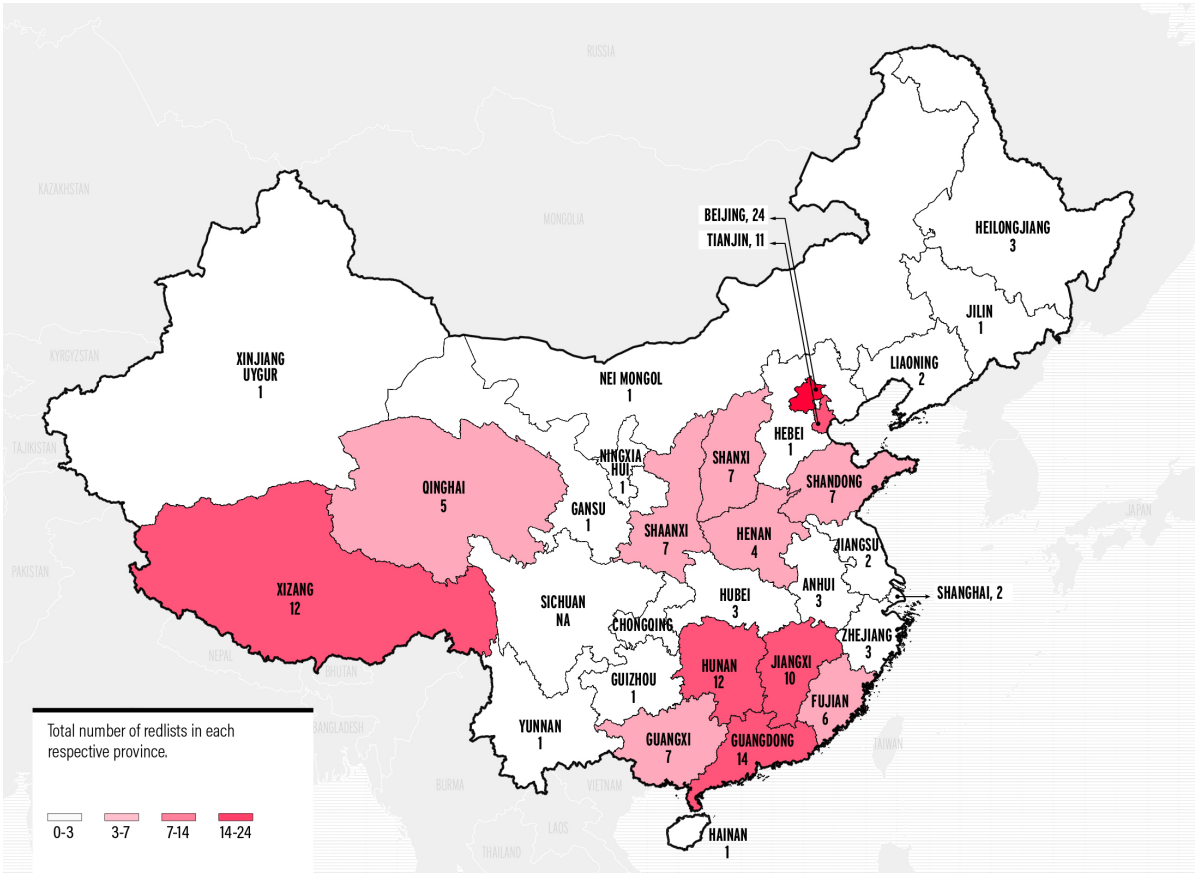


Figure 3.19: The number of redlists implemented across 30 ADs. Beijing had implemented the most redlists (24), followed by Guangdong (14), Xinjiang (12), Hunan (12), Tianjing (11), and Jiangxi (10).

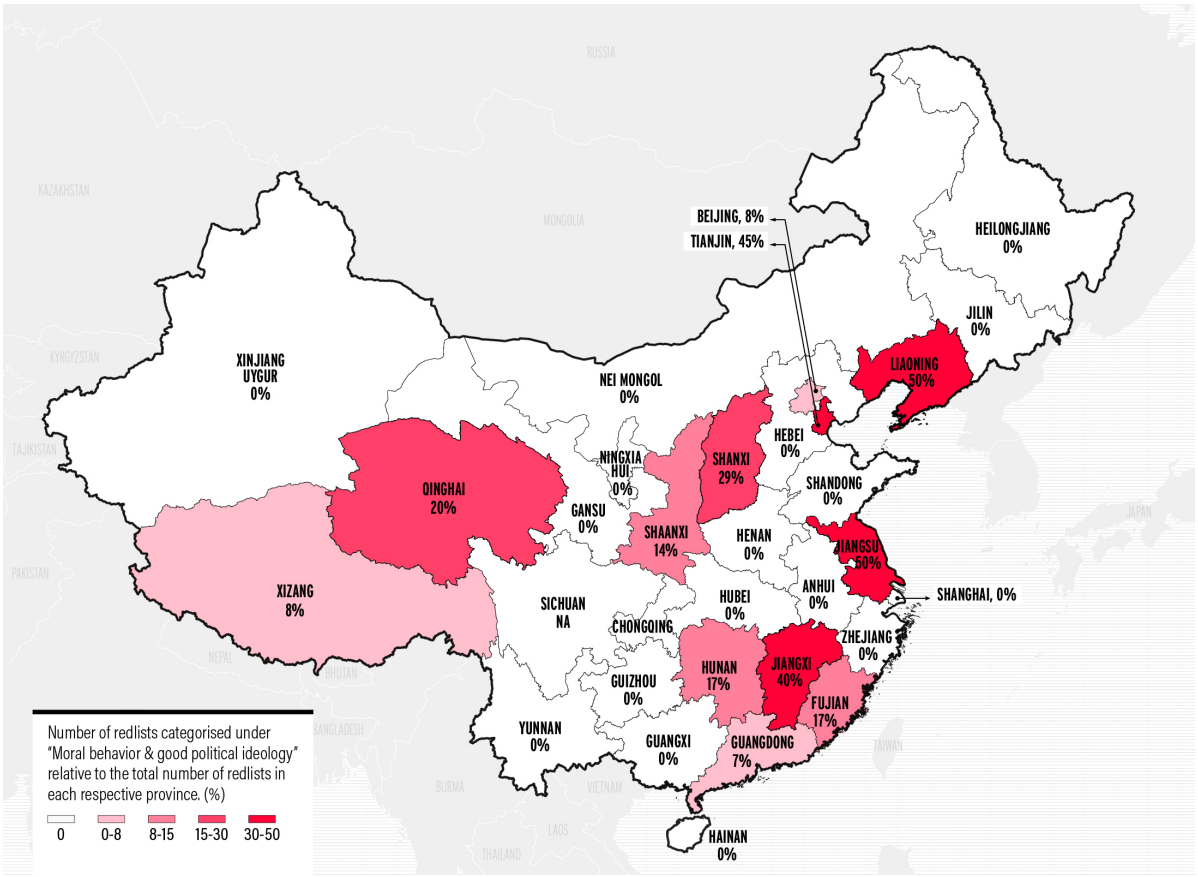


Figure 3.20: Ratios of redlists for moral behavior and good political ideology to total redlists across the 30 listed Chinese ADs.

Socialism with Chinese Characteristics", and Jiangxi and Tianjin listed citizens that had been rewarded the "May Fourth Medal". Tianjin had implemented two lists titled "Tianjin Good Man" and "Tianjin Ideological and Moral Model". Tibet had a similar redlist called "Moral Models & Good Political Ideology" (Figure 3.20). Other redlists were dedicated to citizens that had volunteered, given to charity or won awards in education, science or technology. Overall, the redlist infrastructure was less elaborate than its blacklist counterpart: not a single type of redlist existed in all ADs. Three ADs had published a single redlist with no data (Xinjiang, Gansu, and Jilin).

Informational consistency on credit records of the most common blacklist and redlist

To address RQ3, we explored the informational differences among the credit records of the most frequently implemented types of lists: the "Lao Lai" list (blacklist) and the "Class A Taxpayer" list (redlist). With the exception of Jilin and Tibet, the remaining 28 ADs had published credit records in their "Lao Lai" lists. We compared ADs based on the provision of five types of information in "Lao Lai" credit records: 1) the unified social credit code

Types of blacklists	Blacklists of untrustworthy entities (general)	Blacklists of commercial and transactional activities	Other blacklists	Employment relationship blacklists	Product quality blacklists	Blacklists of financial fraud	Industry blacklists	Blacklists of social fraud	Coronavirus blacklists
Beijing	2	4	3	0	1	4	0	0	NA
Shanghai	4	2	2	0	0	2	1	0	1
Tianjin	6	1	0	1	0	3	0	0	1
Hebei	1	0	0	0	0	0	0	0	1
Shanxi	5	5	5	1	3	2	12	2	1
Liaoning	2	1	3	0	0	0	0	0	0
Jilin	1	6	3	1	2	1	0	0	NA
Heilongjiang	3	0	0	0	0	0	0	0	1
Shandong	2	6	2	1	1	2	1	0	1
Jiangsu	2	0	2	0	0	1	0	0	1
Zhejiang	3	4	2	0	1	1	0	0	1
Anhui	4	0	1	0	0	0	0	0	1
Fujian	3	2	2	2	0	1	1	0	NA
Jiangxi	2	2	1	1	0	2	2	0	1
Henan	5	4	1	0	1	1	2	0	1
Hubei	2	5	0	0	1	0	0	0	0
Hunan	5	5	3	1	4	0	3	0	0
Sichuan	2	1	0	0	0	0	0	0	0
Guangdong	1	5	5	4	2	1	1	0	0
Gansu	1	0	0	0	0	0	0	0	0
Hainan	2	0	0	0	0	0	0	0	NA
Qinghai	3	4	2	1	3	2	7	0	NA
Guizhou	1	0	0	0	0	0	0	0	1
Yunnan	2	0	0	0	0	0	0	0	1
Shaanxi	1	1	1	1	0	0	0	0	1
Tibet	1	4	0	6	0	0	1	0	0
Inner Mongolia	1	0	0	0	0	0	0	0	1
Guangxi	3	1	1	2	0	0	0	0	0
Ningxia	1	0	0	0	0	0	0	0	0
Xinjiang	2	0	0	0	0	0	0	0	0

Types of redlists	Redlists of trustworthy entities (general)	Redlists of commercial and transactional activities	Redlists of moral behavior & good political ideology	Employment & customer relationship redlists	Product quality redlists	Industry redlists & redlists of professions	Redlists of intellectual property & awards	Other redlists	Coronavirus redlists
Beijing	1	5	2	2	2	4	8	0	NA
Shanghai	1	1	0	0	0	0	0	0	0
Tianjin	0	3	5	0	0	0	3	0	0
Hebei	1	0	0	0	0	0	0	0	1
Shanxi	0	3	2	0	0	1	0	1	0
Liaoning	0	1	1	0	0	0	0	0	0
Jilin	1	0	0	0	0	0	0	0	NA
Heilongjiang	1	2	0	0	0	0	0	0	1
Shandong	1	3	0	0	1	1	0	1	1
Jiangsu	1	0	1	0	0	0	0	0	0
Zhejiang	2	1	0	0	0	0	0	0	1
Anhui	0	3	0	0	0	0	0	0	1
Fujian	1	3	1	1	0	0	0	0	NA
Jiangxi	0	2	4	0	0	3	1	0	0
Henan	0	3	0	0	0	0	1	0	1
Hubei	1	2	0	0	0	0	0	0	0
Hunan	1	2	2	0	0	3	4	0	0
Sichuan	0	0	0	0	0	0	0	0	0
Guangdong	0	7	1	0	0	1	5	0	0
Gansu	1	0	0	0	0	0	0	0	1
Hainan	1	0	0	0	0	0	0	0	NA
Qinghai	0	2	1	0	0	1	1	0	NA
Guizhou	1	0	0	0	0	0	0	0	1
Yunnan	1	0	0	0	0	0	0	0	1
Shaanxi	1	3	1	0	0	2	0	0	1
Tibet	0	4	1	0	2	2	1	2	0
Inner Mongolia	1	0	0	0	0	0	0	0	0
Guangxi	0	1	0	0	0	0	6	0	0
Ningxia	1	0	0	0	0	0	0	0	0
Xinjiang	1	0	0	0	0	0	0	0	0

Table 3.3: The different types of blacklists and redlists implemented by ADs in China. Shading indicates the number of blacklists or redlists for a given type. N/A denotes no access to the SCS platform.

(companies) or identification number (natural persons), 2) specification of a data source or responsible authority, 3) reasons for listing (i.e., a justification), 4) information on the fulfillment of the requirements, and 5) information on a future removal date of the record (see Figure 3.21).

Information on "Lao Lai" blacklist credit records

Based on the samples of credit records obtained, out of the 28 different ADs, only 14 ADs had provided either the unified social credit code (8/28) or the natural person's identification number (6/28). The remaining ADs either listed an organization code (3/28) for companies or simply the name of the natural person listed (3/28). 23 ADs specified the data source of the record (i.e., where the data had been generated), the name of the executive court (12/28) or a responsible agency.

In all, 24 ADs provided at least some explanation for why an entity had been listed. In the majority of cases, the credit records referred to a specific law that was to be enforced. Finally, 12 ADs indicated whether the requirement had already been fulfilled or not, and only 6 ADs displayed the removal date of the record.

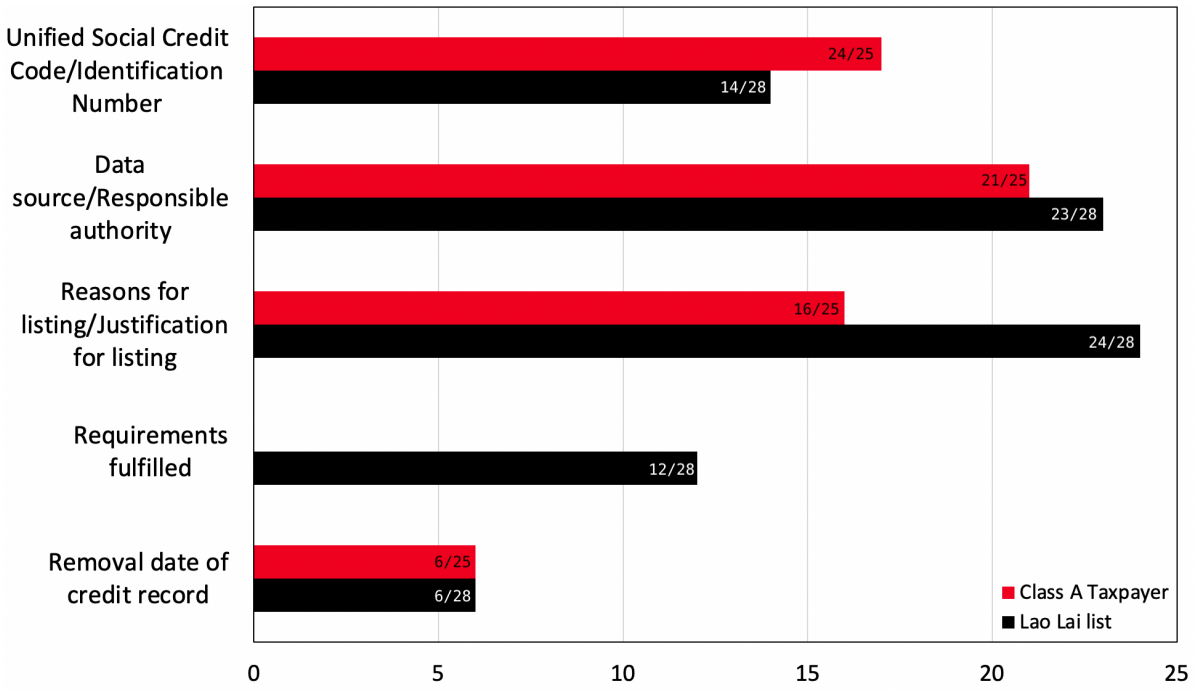


Figure 3.21: A comparison of the information provided on credit records collected from the most frequently implemented type of blacklist and redlist across all ADs.

Information on "Class A Taxpayer" redlist credit records (including unspecified redlists)

For ADs without a "Class A Taxpayer" list, we inspected records from the only list available. 25 ADs had provided redlist records on their SCS platforms. 17 ADs had explicitly used the term "unified social credit code" in their records, and 7 listed a "taxpayer identification number". The remaining ADs simply presented the name of the listed entity. All ADs that published redlist records provided some form of identifying information. Of these, 21 ADs indicated the responsible authority for the case in question, and 16 ADs included a justification for being listed (commonly termed "reason for inclusion" or "honor content"). 6 ADs indicated the record's expiration date. An example record of a Class A Taxpayer List is shown in Figure 3.22.

Flexibility: Blacklists and redlists regulate behavior during the COVID-19 epidemic

Finally, we found that novel types of norm transgression can be quickly subjected to blacklisting and redlisting. Between February 27 and March 30, 2020, we collected data from the same SCS platforms to understand whether blacklisting and redlisting were used to regulate social behavior in an exceptional state of emergency. During this second round of data collection, we had access to 25 of the 31 ADs.[22] We identified coronavirus-related blacklists in 15 ADs and redlists in 10 ADs. Pursuant to our first analyses, blacklist and

[22]We did not have access to the SCS platforms of Jilin, Beijing, Fujian, Qinghai, Chongqing, and Hainan.

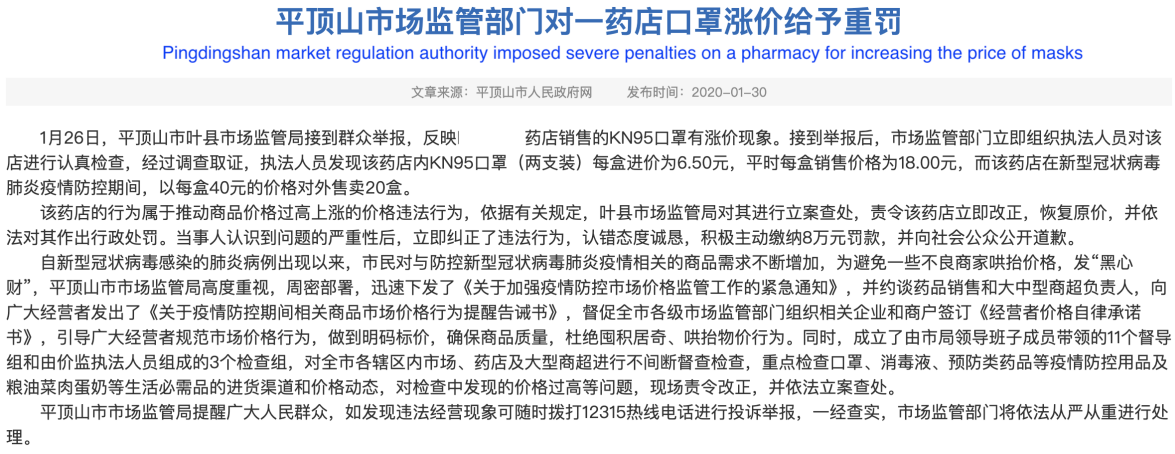
Figure 3.22: A screenshot of a redlist record from the "Class A Taxpayer List" published on the Fujian SCS platform. Translations by the authors.

redlist records targeted natural persons and companies. We found that coronavirus blacklists included entities for selling fake preventive health products, violating quarantine regulations, organizing or participating in gatherings during lockdown, or illegally operating transport vehicles as ambulances. Blacklists were presented in different formats across the 15 ADs: they were either given in a row-and-column format (5) or in narrative-like news reports (10) (see Figure 3.23). Coronavirus redlists reported on devoted professionals such as doctors, nurses, volunteers, and border control officials, as well as on companies and individuals that had donated health products. All coronavirus redlist records were presented as narrative news reports.

3.2.4 Summary and Concluding Analysis

We conducted an empirical investigation on the diversity, flexibility, and comprehensiveness of provincial-level SCS blacklists and redlists in China. Overall, we highlighted that SCS listing designs facilitate public access to social credit records. The majority of SCS platforms display a selection of credit records and enable targeted queries. SCS platforms serve as digital reputation systems because redlists and blacklists digitally showcase entities' good and bad behaviors. However, with the exception of a few ADs that aggregated credit records for a single entity or allowed sharing of credit records to social media platforms, we did not observe any automated classification, ranking or scoring on any of the current SCS listings.

The SCS comprises hundreds of blacklists and redlists across provincial-level ADs. Currently, the majority of these types of lists target compliance with a wide range of laws and



平顶山市场监管部门对一药店口罩涨价给予重罚

Pingdingshan market regulation authority imposed severe penalties on a pharmacy for increasing the price of masks

文章来源：平顶山市人民政府网　　发布时间：2020-01-30

1月26日，平顶山市叶县市场监管局接到群众举报，反映[redacted]药店销售的KN95口罩有涨价现象。接到举报后，市场监管部门立即组织执法人员对该店进行认真检查，经过调查取证，执法人员发现该药店内KN95口罩（两支装）每盒进价为6.50元，平时每盒销售价格为18.00元，而该药店在新型冠状病毒肺炎疫情防控期间，以每盒40元的价格对外售卖20盒。

该药店的行为属于推动商品价格过高上涨的价格违法行为，依据有关规定，叶县市场监管局对其进行立案查处，责令该药店立即改正，恢复原价，并依法对其作出行政处罚。当事人认识到问题的严重性后，立即纠正了违法行为，认错态度诚恳，积极主动缴纳8万元罚款，并向社会公众公开道歉。

自新型冠状病毒感染的肺炎病例出现以来，市民对与防控新型冠状病毒肺炎疫情相关的商品需求不断增加，为避免一些不良商家哄抬价格，发"黑心财"，平顶山市市场监管局高度重视，周密部署，迅速下发了《关于加强疫情防控市场价格监管工作的紧急通知》，并约谈药品销售和大中型商超负责人，向广大经营者发出了《关于疫情防控期间相关商品市场价格行为提醒告诫书》，督促全市各级市场监管部门组织相关企业和商户签订《经营者价格自律承诺书》，引导广大经营者规范市场价格行为，做到明码标价，确保商品质量，杜绝囤积居奇、哄抬物价行为。同时，成立了由市局领导班子成员带领的11个督导组和由价监执法人员组成的3个检查组，对全市各辖区内市场、药店及大型商超进行不间断督查检查，重点检查口罩、消毒液、预防类药品等疫情防控用品及粮油菜肉蛋奶等生活必需品的进货渠道和价格动态，对检查中发现的价格过高等问题，现场责令改正，并依法立案查处。

平顶山市市场监管局提醒广大人民群众，如发现违法经营现象可随时拨打12315热线电话进行投诉举报，一经查实，市场监管部门将依法从严从重进行处理。

Figure 3.23: Screenshot of the coronavirus blacklist from the SCS platform for Henan province. Translation: *On January 26, the Market Supervisory Authority of Ye County Pingdingshan City received reports from the public reporting that ** Pharmacy increased the price of KN95 masks. After receiving the report, the authority immediately sent out law enforcement officers to conduct a serious inspection of the store and found that the purchase price of the KN95 masks (2 pieces in one package) was 6.5 RMB for the store and the sale price was usually 18 RMB. However, the pharmacy sold 20 packages of the masks at the price of 40 RMB during the epidemic period. The pharmacy was thus in violation of the price regulation. Following relevant regulations, the Market Supervisory Authority filed a case for the investigation and ordered the pharmacy to restore the price to its original level. The authority also imposed administrative penalties on the pharmacy according to law. The pharmacy realized the seriousness of the problem and immediately halted the illegal behavior, admitted its misconduct, proactively paid a fine of 80,000 RMB, and apologized to the public.* Translations by the authors.

regulations. Thereby, SCS blacklists focus on "Lao Lai" entities, which are citizens and companies that have not fulfilled a court order. The SCS first displays "Lao Lai" on its digital listings and hence excludes them from future cooperative opportunities through its JRP. Based on these two mechanisms, the SCS seeks to turn "Lao Lai" into cooperators by attaching an exceptionally high cost to defection. We also observed redlists that highlight praiseworthy political and moral behaviors. Further development of lists that go well beyond legal or regulatory norms could substantially increase the social control characteristics of the SCS.

We have exemplified the flexibility of SCS listings by a case study on the COVID-19 outbreak. Digital blacklists and redlists might be a particularly powerful regulatory measure because they can be adapted to help accomplish novel policy goals quickly and at relatively low costs.

There are several outstanding questions for future research. For example, will SCS platform design incorporate more reputational affordances? Will the governmental and commercial branches (i.e., big data apps) of the SCS cooperate to share and analyze different data streams? Will SCS mechanisms really produce their intended regulatory effects? We believe that asking such questions is crucial and we hope to have laid a useful foundation for future empirical and conceptual studies on the SCS.

3.2.5 Ethical dimensions of the study

We now turn to initial ethical considerations of data-driven research on SCS implementation. First, our analysis was based on publicly available data found on key platforms of China's SCS. These data are posted to enable public scrutiny. Our paper includes screenshots from the currently available implementations (see Figures 3.14, 3.15, 3.16, 3.17, 3.22, 3.23). Our data collection and analyses are privacy-preserving: we blurred any personally identifiable data to protect the privacy of listed companies and citizens. Our methodological approach does not result in any unfavorable consequences or costs for any of the data subjects. We are transparent in our methodology and provide a representative code example of a web crawler and spider we used in this study (see Auxiliary Material).

Second, our account adheres to the principles of ethical web crawling and scraping [174, 175, 176, 177]. For each SCS platform, we checked for a specified *robots.txt* file. At no point during our data collection did we find a robots.txt file that specified rules for web crawlers. Accordingly, when platforms make data publicly available, do not specify a robots.txt file, and do not provide a data collection interface (e.g., API), then robots are free to gather data (see, e.g., [174, 177]).

Third, the purpose of our study is ethically justifiable on its own. In the absence of systematic empirical accounts, uncertainty will inevitably help foster misconceptions about the SCS (whether overly positive or negative). Given China's geopolitical prominence, governments of other countries may be inspired to copy China's SCS [157]. This is particularly likely for neighboring countries [178]. Data-driven research on SCS implementation can help prevent hasty SCS adaptations by other governments based on false assumptions. Empirical and conceptual analyses on the SCS allow for a more informed public debate about the

development of digital socio-technical systems. As our data indicate, *currently*, there is little evidence that blacklists and redlists operate as AI-driven reputation systems. Apart from two SCS platforms that enable sharing of credit records to social media platforms, at the moment, there is no evidence that credit records are subjected to other means of digital reputation mechanisms such as classification, ranking, or profiling based on AI. It is possible that future developments might implement AI-based reputation mechanisms. As we have argued, additional empirical work on the SCS is necessary given that Chinese policy-making rests on often vaguely formulated policy goals. We show a considerable diversity of SCS blacklist and redlist implementation that cannot be concluded from policy analysis alone. Our study raises important questions that also matter for non-Chinese citizens and organizations. For example, is stable access to blacklists and redlists from outside China justifiable when non-Chinese citizens and companies are listed [162, 179]? Should China distribute licenses or special APIs to allow non-Chinese entities to ascertain whether they are listed? Or will Chinese authorities directly notify non-Chinese entities when they are listed?

The Chinese SCS is already one of the most comprehensive reputation systems in the world. Given that the government generates the reputation signals, we believe that SCS blacklisting and redlisting could have a strong influence on societal behavior at large.

Finally, this research extends growing calls for more open data in computational social science [180] with a case for more data availability *in China*. As this body of research has shown, open government data can significantly improve our understanding of societies' most important challenges in the context of equality, health, or employment. Even if data collection obstacles are likely to persist, we hope that our study underlines the importance of future data-driven research on the Chinese SCS.

Acknowledgements

We thank Rogier Creemers, Bilge Kobas, and Marianne von Blomberg for their helpful input. We also thank the anonymous reviewers for their constructive comments. We gratefully acknowledge funding support from the Bavarian Research Institute for Digital Transformation (bidt). Mo Chen further thanks for the support through a postdoc research stipend of the Fritz Thyssen Foundation. Responsibility for the content of this publication rests with the authors.

3.2.6 Auxiliary Material

Documentation: Example crawler and spider for Guangdong province

The following code sections are an excerpt of the crawling and scraping methodology to systematically collect data from public blacklists and redlists of the Chinese Social Credit System. The crawler for collecting relevant data and the spider for extracting specific information from the data are demonstrated for the example of the Guangdong province below. Please note that the collection methodology may have to be adjusted, if the collection site is undergoing changes. You also may want to revisit the discussion on the ethics of data crawling in our paper (see Section 3.2.5).

Crawler example Guangdong province:
This section shows how the link lists are created, in particular, the methodology to collect the deep links that lead to the entry records of blacklists and redlists. A headless browser (like Selenium) is used, which is basically a normal web browser remotely controlled by a programmed robot.

In the following, an example of a web crawler is given:

```
class GuangdongSelenium():
    def crawl_red(self):
        link = 'https://credit.gd.gov.cn/opencreditAction!getOpencreditList_new
            ↪ .[...]&tbType=1'
        print_start("Guangdong␣Redlist")
        linkliste = []
        file = open("linklist_guangdong_red.txt", "a")

        driver.get(link)
        driver.find_element_by_css_selector('#newtype␣>␣option:nth-child(8)').
            ↪ click()
        driver.find_element_by_css_selector('label.search_button').click()

        while '下一页' in driver.page_source:
            try:
                categorylist = driver.find_elements_by_css_selector('tbody␣>␣tr:
                    ↪ nth-child(1)␣>␣td␣>␣div␣>␣a')
                for i in categorylist:
                    print(i.get_attribute('href'))
                    s = i.get_attribute('href')
                    linkliste.append(s)
                driver.find_element_by_css_selector('a.next').click()
                time.sleep(10)
```

```
            except():
                print ("Error, no next page available!")
                break

        print("Length of final linklist: ", len(linkliste))
        linkliste = list(dict.fromkeys(linkliste))
        print("This is the lenght of the list after removing all duplicates: ",
            ↪  len(linkliste))
        for e in linkliste:
            file.write(e + "\n")

        print("Crawled links are written into the final file.")
        print("File created")
        file.close()
        driver.close()
        sys.exit()

    def crawl_black(self):
        link = 'https://credit.gd.gov.cn/opencreditAction!getOpencreditList_new
            ↪ .[...]&tbType=2'
        print_start("Guangdong Blacklist")
        linkliste = []
        file = open("linklist_guangdong_black.txt", "a")
        driver.get(link)
        driver.find_element_by_css_selector('#newtype > option:nth-child(2)').
            ↪ click()
        driver.find_element_by_css_selector('label.search_button').click()
        try:
            while '下一页' in driver.page_source:
                wait = WebDriverWait(driver, 10)
                wait.until(ec.visibility_of_element_located((By.CSS_SELECTOR, 'a.
                    ↪ next')))
                time.sleep(10)
                categorylist = driver.find_elements_by_css_selector('tbody > tr:
                    ↪ nth-child(1) > td > div > a')
                for i in categorylist:
                    print(i.get_attribute('href'))
                    s = i.get_attribute('href')
                    file.write(s + "\n")
                    linkliste.append(s)
                driver.find_element_by_css_selector('a.next').click()
                time.sleep(5)
```

```
        except:
            pass
            print("Error, no next page available!")
        print("File created")
        file.close()
        driver.close()
        sys.exit()
```

The desired output should be a collection of links stored in corresponding files 'linklist_guangdong_black.txt' or 'linklist_guangdong_red.txt'.

```
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=
    ↪ FF89EED12BC14E21BF36360E9044FC45
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=
    ↪ FF89EED12BC14E21BF36360E9044FC45
[...]
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=
    ↪ FF89EED12BC14E21BF36360E9044FC45
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=
    ↪ FF89EED12BC14E21BF36360E9044FC45
```

Spider example Guangdong province:
This section shows a web scraping spider, a methodology that follows the web crawling process. A web scraper's task is to sequentially work through the web crawler's link list and extract specific data.

In the following, an example of a web scraper is given:

```
import scrapy, re

class GuangdongSpider(scrapy.Spider):
    name = "guangdong"
    file = open("linklist_guangdong_black.txt", "r")
    start_urls = [i.replace("\n", "") for i in file]

    def parse(self, response):
        table = response.css('table > tr > td')
        yield{
            'case_number' : table[1].css('::text').extract_first(),
            'lost_trustee_name' : table[3].css('::text').extract_first(),
            'gender' : table[5].css('::text').extract_first(),
            'age' : table[7].css('::text').extract_first(),
```

```
        'ID_number_desensitization__organization_code' : table[9].css('::
            ↪ text').extract_first(),
        'corporate_legal_person_name' : table[11].css('::text').
            ↪ extract_first(),
        'executive_court' : table[13].css('::text').extract_first(),
        'execution_basis_number' : table[15].css('::text').extract_first(),
        'basis␣for␣execution' : table[17].css('::text').extract_first(),
        'obligation_established_by_the_law' : table[19].css('::text').
            ↪ extract_first(),
        'implementation_of_the_person_being_executed' : table[21].css('::
            ↪ text').extract_first(),
        'untrustworthy_enforcer' : table[23].css('::text').extract_first(),
        'release_time' : table[25].css('::text').extract_first(),
        'filing_time' : table[27].css('::text').extract_first(),
        'fulfilled_part' : table[29].css('::text').extract_first(),
        'unfulfilled_part' : table[31].css('::text').extract_first(),
        'hyperlink' : response.url
    }
```

Table: Summary of credit record collection for blacklists and redlists

AD	No. of blacklist records	Avg. size blacklist record	No. of variables	No. of redlist records	Avg. size redlist record	No. of variables
Municipalities						
Beijing	100	1700 B	35	50	776.9 B	27
Shanghai	10	156.5 B	3	10	157.8 B	3
Tianjin	1501	1100 B	5	2000	306.6 B	5
AR						
Guangxi	30281	265.7 B	8	27692	547.5 B	15
Inner Mongolia	10	795.9 B	15	10	319.5 B	5
Ningxia	20	853.3 B	12	19	714.5 B	12
Xinjiang	3	1100 B	12	no data	-	-
Tibet	no data	-	-	no data	-	-
Provinces						
Anhui	190	926.5 B	15	190	315.8 B	6
Fujian	99	477.6 B	9	78	380.5 B	7
Gansu	20	1200 B	21	no data	-	-
Guangdong	160	1900 B	17	90	476.1 B	6
Guizhou	38	1600 B	6	39	2900 B	6
Hainan	40	817.3 B	17	40	654.6 B	13
Hebei	311	663.9 B	11	652	515.2 B	11
Heilongjiang	24	804.2 B	6	7	939.7 B	14
Henan	180	218.0 B	2	180	218.0 B	2
Hubei	50	588.4 B	11	50	465.5 B	8
Hunan	20	174.1 B	4	79	129.9 B	3
Jiangsu	50	1700 B	26	50	440 B	8
Jiangxi	2413	1600 B	16	482	1300 B	13
Jilin	no data	-	-	no data	-	-
Liaoning	4	1100 B	14	8	356.1 B	8
Qinghai	19	1000 B	15	18	928.6 B	15
Shaanxi	49	1100 B	15	47	748.6 B	15
Shandong	100	672.3 B	14	100	361.5 B	7
Shanxi	53	2100 B	21	73	1100 B	21
Sichuan	320	226.4 B	10	10	650.9 B	10
Yunnan	50	752.0 B	9	42	516.8 B	9
Zhejiang	1950	163.0 B	4	5580	217.0B	5
$\sum$	38065			37596		

Table 3.4: The "No. of blacklist records" and "No. of redlist records" indicate the number of credit records retrieved from each AD SCS platform for the most commonly implemented type of blacklist and redlist, respectively. Numbers show varying sample sizes due to several data collection obstacles (see Section 3.2.2). "Avg. size blacklist record" denotes the average byte size of a blacklist record for each sample. "No. of variables" indicates the number of informational variables on each credit record in the sample.

3.3 Research Article 3: Ordinary people as moral heroes and foes: Digital role model narratives propagate social norms in China's Social Credit System

Authors

Mo Chen, Severin Engelmann, Jens Grossklags

Publication Outlet

AIES'22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society; July 2022; Pages 181–191; https://doi.org/10.1145/3514094.3534180

Abstract

The Chinese Social Credit System (SCS) is a digital sociotechnical credit system that rewards and sanctions economic and social behaviors of individuals and companies. As a complex and transformative digital credit system, the SCS uses digital communication channels to inform the Chinese public about behaviors that lead to reward or sanction. Since 2017, the Chinese government has been publishing "blameworthy" and "praiseworthy" role model narratives of ordinary Chinese citizens on its central SCS information platform creditchina.gov.cn. Across many cultures, role model narratives are a known instrument to convey "appropriate" and "inappropriate" social norms. Using a directed content analysis methodology, we study the SCS-specific social norms embedded in 100 "praiseworthy" and 100 "blameworthy" role model narratives published on creditchina.gov.cn. "Blameworthy" role model narratives stress social norms associated with an "immoral" SCS identity label termed "Lao Lai" — a "moral foe" that fails to repay debt. SCS role model narratives familiarize Chinese society with SCS-specific measures such as digital surveillance, public shaming, and disproportionate punishment. Our study makes progress towards understanding how a state-run sociotechnical credit system combines digital tools with culturally familiar customs to propagate "blameworthy" and "praiseworthy" identities.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

3.3.1 Introduction

In the past eight years, the Chinese government has made efforts to reshape its domestic power structure. The government removed the term limits for the Chinese presidency, created an anti-corruption ministry, and launched a "propaganda" app called "Xuexi Qiangguo" (学习强国, literally translated as "study and make the country strong").[23] Further, after four decades of rapid economic growth, *domestic* demand-driven models aim to consolidate economic sustainability [182, 183].

In 2014, the government published a *Planning Outline for the Construction of a Social Credit System (2014 - 2020)*; a high-level policy document that mandates a nationwide digital social credit system referred to as the Chinese Social Credit System (社会信用体系, SCS). The SCS's purpose is to evaluate, reward and punish the behavior of individuals, as well as commercial and societal organizations [116]. The outline describes two key SCS-specific regulatory measures: first, a digital "shaming"[24] and "praising" reputation system and a "joint punishment and reward mechanism" that distributes disproportionate "punishments" and "rewards", respectively [184, 185, 162, 186, 187, 136]. The Chinese SCS is a novel regulatory instrument enforcing reputational *and* material incentives and sanctions with the help of a large-scale digital infrastructure. The regulatory idea of the SCS rests on a broad conceptualization of "credit" that covers economic and social behaviors. SCS policy documents specify 14 different economic (e.g., production safety, finance, construction, e-commerce, etc.) and 10 different social sectors (e.g., health care, social security, and labor and employment) for credit application [116]. This "credit everywhere" directive subjects Chinese society to an all-encompassing concept of metrics with the aim to build a "socialist harmonious society" without "social contradictions" [116].

The establishment of a large-scale digital SCS to enforce social norms[25] corroborates the government's efforts to govern society through mechanisms that go beyond common legal and regulatory practices. In order for citizens to comply with SCS-specific social norms, the government must create awareness and understanding of these norms. This research focuses on the central SCS platform "Credit China" (`creditchina.gov.cn`). Run by the National Center for Public Credit Information, the platform functions as the main SCS platform on all SCS-related developments. "Credit China" provides public access to official policy documents of the SCS, presents different types of reputational blacklists, and publishes SCS role model narratives and SCS news reports; as such, the platform also propagates SCS-specific social norms to the Chinese public.

[23] Civil servants, and employees of state-owned enterprises, particularly party members, are "encouraged" to use the app [181].

[24] Throughout this paper, the authors use quotation marks to communicate a neutral standpoint towards SCS-specific normative concepts (e.g., "praiseworthy", "blameworthy", "shaming", "praising").

[25] This paper uses the term *social norm* in a purely functionalist manner (see, e.g., [188]). A functionalist account defines social norms as deliberate measures by one party or group to establish social order over another. While other accounts of social norms study their natural emergence in individual or group interaction (see, e.g., [189]), a functionalist account puts emphasis on the exogenous dimensions of social norms attributable to the Chinese SCS.

While previous research on the SCS has largely focused on policy document analysis, here, we contribute to a more precise understanding of how the Chinese government makes SCS social norms intelligible to society at large. SCS policy documents describe vague instructions on SCS development, a common trait of policy documents issued by the central government [167]. Moreover, the broad public does not tend to engage with policy documents. Second, the SCS digitally publishes credit records on citizens, companies, and other organizations on so-called SCS blacklists (displaying "blameworthy" behavior) and redlists (displaying "praiseworthy" behavior). While blacklist and redlist records provide some information on why an entity was listed (i.e., punished or rewarded) [186], such justifications are written in legal and technical jargon. They do not offer causal or contextual clarifications for the sanctioned or rewarded behaviors [162]. We observe that, since 2017, the national SCS platform "Credit China" (`creditchina.gov.cn`) has been regularly publishing *SCS role model narratives* on "praiseworthy" and "blameworthy" behaviors. SCS role model narratives explicitly convey SCS-specific social norms to a broad audience. They vividly illustrate how ordinary Chinese citizens comply with or transgress SCS-specific norms and what consequences they experience. Narratives, stories, or folklore are as old as civilization. In the Chinese SCS, narratives on ordinary citizens are integrated into a digital infrastructure. They are published online and readers can share narratives to Chinese social media platforms amplifying the messages they seek to convey.

China has a long cultural tradition of propagating social norms through narratives, stories, and portraits of *model individuals* (e.g., [190]). First, Chinese ethical scholarship formulates principles through narratives, rather than through abstract principles. Second, besides a plethora of ancient moral narratives that still profoundly influence moral education in China today,[26] the Chinese government today uses narratives to showcase moral exemplars through reader-friendly stories and portraits (e.g., famous and popular narratives on moral heroes such as Huang Jiguang and Lei Feng). In the context of the SCS, we find that the government employs a similar strategy. Consequently, their analysis enables a more substantive understanding of the specific social norms the Chinese government wants the public to comply with and internalize with regard to SCS implementation.

We apply a directed content analysis methodology to systematically study the SCS-specific social norms embedded in 200 "blameworthy" and "praiseworthy" role model narratives on `creditchina.gov.cn`. Our study exemplifies how socio-cultural traditions influence and resurface in the implementation of a large digital sociotechnical system. Role model narratives on `creditchina.gov.cn` represent a prime example of how "...state actors appropriate technologies to support broader ideological shifts in their discourse" [191]. In addition, digital narratives present the biographical information and moral judgments of ordinary Chinese citizens that, as we show, can be distributed to large social media networks. SCS narratives demonstrate the problematic coupling of traditional values and socio-political policy plans by large digital infrastructures.

[26] For example, the *Twenty-four Stories about Filial Piety* written by Guo in the Yuan Dynasty.

3.3.2 Background

In Western media, the SCS has been linked to a national metric system assigning social credit scores to individuals (e.g., [192, 193]). While this perspective needs clarification, the SCS allows for more government supervision of individuals, companies, and institutions through digital information technologies. First, big information technology companies contribute to the construction of the SCS and distribute trustworthiness scores to individuals in promotion programs (e.g., Zhima Credit) [129, 184]. Second, *local* governments have tested different rating systems in "prototype cities" such as Rongcheng and Suzhou. Here, social credit ratings grant or deny citizens access to various public services and products [117]. Participation in these local "credit scoring experiments" is mandatory. However, these local "policy experiments" do not necessarily serve as a model for national policy implementation.

A number of SCS-specific measures operate at the national level. Early research accounts noted the existence of different types of SCS blacklists and redlists. With these lists, the SCS uses digital platforms to publicly "shame" or "praise" natural and legal persons for non-compliance or compliance, respectively, with a variety of legal and social norms (e.g., [194, 132, 195, 196, 162, 186]). Another national SCS-specific measure is the SCS joint punishment and reward mechanism. Thereby, "praiseworthy" or "blameworthy" behavior in one specific area leads to "reward" or "punishment" in different areas of life. To give just one example, blacklisted individuals have been barred from booking 26.8 million flights and nearly 6 million high-speed train trips since June 2019 (according to the National Development and Reform Commission).[27] Scholars note that public "shaming" and "praising" platforms as well as joint punishment and reward mechanisms differentiate the Chinese SCS from other social credit systems [187, 162].

SCS implementation as a digital transformation of culturally and politically familiar customs

Social science and legal scholarship has mainly focused on the privacy implications that result from the surveillance measures of the Chinese SCS (e.g., [197, 195]). A key observation is that the Chinese SCS is able to collect, process, and analyze personal data for a broad range of different purposes [127, 198]. As a "surveillance system", the SCS is a critical stepping stone for the government not only to monitor, but also to regulate and shape people's behaviors [136]. However, prior research seems to indicate that Chinese citizens do not primarily associate the SCS with the dangers of surveillance [129]. Compared to the astonishment and criticism from some Western media (e.g., [196, 199, 200]), Chinese citizens appear to perceive the SCS favorably rather than critically [129]. The high approval levels can partially be explained by the effort of the government to base SCS mechanisms on culturally familiar customs and practices. For example, blacklists and redlists are common modes of shaming and praising schemes in Chinese society. In kindergarten, it is not uncommon for children to receive "praise" and "blame" via so-called "Honor Rolls" and "Critique Rolls", respectively. Beyond kindergarten, "praise" and "blame" mechanisms include public presentation of photos of

[27] Refer to `http://www.sohu.com/a/327229387_120054409`, last accessed on May 21, 2022.

individuals on banners at the entrance of buildings such as hospitals, schools, and companies. The distribution of reputational "reward" and "punishment" by institutions represents a culturally accepted regulatory instrument.

Second, according to survey research, Chinese citizens voice little doubt regarding the political legitimacy of the government to ensure social order through surveillance and monitoring systems [129]. Characteristics of what has been referred to as the "surveillance tradition" of the government date back to the "personal file system" *dang'an* [187, 195]—a national archive system that was set up in 1949 to systematically collect, record, and store information on citizens' and organizations' attitudes and behaviors [201]. Similar to the *dang'an*, SCS measures apply to individual citizens, companies, and social organizations. Given the longstanding surveillance practices represented by the *dang'an* system, Chinese society is unlikely to perceive the implementation of data-rich digital reputation lists by the government as an illegitimate political measure. This is not to say that Chinese citizens attach a low value to their privacy in principle. When it comes to using corporate digital services such as WeChat, for example, Chinese citizens do raise concerns about their privacy but are less likely to take corresponding privacy actions [202]—this "privacy paradox" is prevalent among users in Western societies, too [203].

Narratives as instruments for propagating ethical norms and political propaganda

Across cultures, stories, poems, and plays are an indispensable and prevalent source of ethical principles [204, 205, 206, 207]. Narratives naturally raise ethical questions and present possible model behaviors, good and bad. The narrative format is particularly suitable to illustrate complex ethical scenarios in a comprehensible manner. In William Shakespeare's *King Henry V* soldiers face the moral trade-off whether to fulfill the king's demands for war when they believe that the king's motivation for war is irrational and unjust. Or take Mark Twain's *The Adventures of Huckleberry Finn*. The story illustrates the moral tensions of Huckleberry Finn who decides to protect his escaped enslaved friend Jim rather than returning Miss Watson's "lost property". Narratives are powerful media for ethical deliberations, they place moral choices in specific, real-world contexts. The narrative format may not be suitable for generalizing abstract principles, but it vividly reveals the conditional trajectories that cause protagonists to face moral trade-offs or dilemmas [205].

Deontological and utilitarian ethics are typically concerned with the conceptual development of ethical principles. These ethical traditions justify a moral imperative conceptually and take them to be universally valid across contextual conditions. In contrast, Chinese ethics has a practical focus and demands practical solutions to specific ethical conflicts [208], and is *"skeptical that highly abstract theories will provide a response that is true to the complexities of that problem"* [209]. As such, Chinese moral philosophy takes a predominantly virtue ethics approach. Its emphasis lies on the development and presentation of a particular moral character in the face of a particular problem [209]. Here, the narrative format plays an indispensable role in conveying ethical deliberation and decision-making in Chinese ethics. Examples of Chinese role model narratives abound. The *Biographies of Exemplary Women*,

compiled two millennia ago, is the earliest extant book of Confucian ethics solely devoted to the education of women. It includes 125 biographical accounts of exemplary women in ancient China. Well-known to the Chinese today is the famous *Twenty-four Stories about Filial Piety*. Written about 700 years ago, this collection of stories aims to educate the public on the virtue of Confucian filial piety. In Confucian ethics the virtue of filial piety represents a constitutive element of "communitarianism". Narrated scenarios illustrate virtuous acts that cover moral conflicts. For example, the passage *7A35* in the book *Mencius*, places the protagonist in the following situation: would one hand over one's own father to the state if he has committed a murder? Another "virtuous exemplar" of filial piety—perhaps better known to the Western world—is the young girl called Mulan. An entire collection of poems called the *Ballad of Mulan* documents her courage and sense of duty in China 1500 years ago.[28]

In the 20th century, the Chinese government has used role model narratives to underline "praiseworthy" moral dispositions. For instance, Huang Jiguang is highly decorated as a revolutionary martyr for "sacrificing" himself during the Korean War in the 1950s. Another example is the story of Lei Feng—a socialist hero during the 1960s and a famous hero in contemporary Chinese society [210]. He is glorified for his "unconditional loyalty" to the Chinese Communist Party (CCP). More recently, stories praising and blaming citizens regularly appear on Chinese television. In 2016, the state's television station China Central Television produced a special program called "Role Model/榜样". In each season, the program presents the "stories" of ten CCP members, praising their dedication and steadfastness in their faith as CCP members. The Chinese public is familiar with the use of narrative portraits of role models that propagate political and ideological ideals. Narratives published on `creditchina.gov.cn` follow this tradition and instill a representation of everyday moral life in citizens' minds [211]. This work presents evidence that the Chinese SCS uses narratives of ordinary Chinese citizens to familiarize society with digital surveillance practices and digital reputation listings to enforce SCS-specific norms.

3.3.3 Data and methods

Data

In September 2017, the national SCS platform `creditchina.gov.cn` started the regular publication of "blameworthy" role model narratives about "dishonest"/"untrustworthy" natural and legal persons. These "blameworthy" role model narratives can be accessed on the landing page of `creditchina.gov.cn` (titled "representative cases/典型案例)"[29]. In November 2017, the platform also started publishing "praiseworthy" role mode narratives of "honest" and "trustworthy" individuals and representatives of companies. These "praiseworthy" role model narratives can be accessed on the sub-page "credit culture (诚信文化)" under the headline "integrity characters/stories (诚信人物／故事)". Both "praiseworthy" and "blameworthy"

[28] For a comprehensive overview of narratives in Chinese ethics, see [209].

[29] This section only included "blameworthy" narratives when we crawled the data in August 2018. Now, this section includes both "blameworthy" and "praiseworthy" narratives.

Title: *Quzhou Court: For the first time, capturing a "Lao Lai" using the measure of "temporary control"!*

The charged person, [redacted] (last name), is a "contractor". In 2017, he hired three people ([redacted], [redacted] and another person) to work for a steel company in Fengnan District, Tangshan City. He did not pay the workers their salary which amounted to 14,300 RMB and was subsequently sued by the court. The court of Quzhou County ordered [redacted] to pay 14,300 RMB for labour remuneration to [redacted], [redacted], and others. After the verdict came into effect, [redacted] refused to fulfill his obligation, and the case entered the enforcement process.

The court of Quzhou County dealt this case as one involving people's livelihood and tried to educate and persuade [redacted] directly or through his family members. But [redacted] still refused to fulfil his obligation. He went out to work and played the game of "hide and seek" with court executives. [redacted] was then put on the "List of Dishonest Persons Subject to Enforcement" according to law, and became a "Lao Lai". Due to [redacted]'s long-term concealment and evasion of execution of assets, in July this year, the court of Quzhou County applied "temporary control" in accordance with the law with the help of the public security bureau in July. Only three days later, the Yonghongqiao Police Station, Lunan Branch of Tangshan Public Security Bureau came with the good news: [redacted] was successfully captured. The court of Quzhou County dispatched executives who drove more than 1,200 kilometres overnight to take [redacted] back. Frightened of the strong enforcement, [redacted] contacted his family on his way back to the court. Finally, his family then sent the money to the court.

Figure 3.24: Translation of a "blameworthy" role model narrative from `creditchina.gov.cn`. This is an excerpt of the complete role model narrative. The narrative also provided the following information: publication date (July 30, 2018), original source of the role model narrative (Jiaotong Wang), and the category of the role model narrative (Representative Cases); as well as a sharing function with links to the platforms of Wechat, Weibo, Baidu Tieba, and Renren.

narratives are either created and published by `creditchina.gov.cn` itself or selected and taken from city, provincial, and other national government-associated news outlets.

We crawled and scraped publicly available "blameworthy" and "praiseworthy" narratives on `creditchina.gov.cn`. This resulted in a corpus of 798 "blameworthy" and 156 "praiseworthy" role model narratives. To generate comparable datasets, we used the random number method (e.g., [212]) to select 100 "praiseworthy" and 100 "blameworthy" role model narratives. We found that protagonists in all "praiseworthy" narratives were individuals and their full names were provided. In contrast, 11 out of 100 "blameworthy" narratives (11%) portrayed companies. Only in 2 "blameworthy" cases (2%), a full name of the protagonist was included, while in the remaining 98 cases the protagonist's name was partly anonymized (only the family name was provided). In the process of coding, we obscured the protagonist's name, living address and related companies' names to reduce the risk of re-identification. Translations of a "blameworthy" and a "praiseworthy" narrative can be found in Figures 3.24 and 3.25, respectively.

Title: *Twenty years of upholding "honesty and trustworthiness" and giving back to the home village*

, born in November 1963, is a member of the Communist Party of China, secretary of the party branch of Village, Shuitun Town, Yicheng District, Zhumadian City, Henan Province, general manager of Human Resources Co., Ltd. and Technology Co., Ltd. has always adhered to the life tenet of "honesty and trustworthiness is gold and virtuous". He has set up his own "Poverty Alleviation Convoy" with the idea of "facilitating labour with passenger transport, promoting poverty alleviation with labour". For 20 years, he has behaved according to the virtues of honesty and trustworthiness, exempting transport fares for migrant workers from his home village for over 50 million yuan, sending more than 1.6 million people to the south for employment, and helping more than 3,000 families to get rid of poverty. He tried every means to persuade five companies to settle in Village, fulfilling the dream of poor households seeking employment and poverty alleviation at the doorstep of his home. He is enthusiastic about public welfare and donated more than 3 million yuan to roads and bridges construction, education, earthquake relief, and supporting students in need. In recent years, has been awarded more than 30 titles including "National Outstanding Migrant Workers", "Outstanding Migrant Workers from Henan Province" and Zhumadian City "May 1st Labour Medal".

Figure 3.25: Translation of a "praiseworthy" role model narrative from `creditchina.gov.cn`. This is an excerpt of the complete role model narrative. The web-page also provided the following information: publication date (April 2, 2018), original source of the role model narrative (Credit China), and the category of the role model narrative (Trustworthy Figures); as well as a sharing function. It also featured an image of the protagonist and an audio recording of the narrative.

3.3.4 Research ethics

Our analysis is built on publicly available data from key sites of the Chinese SCS, which is posted with the intent of public scrutiny. The two main frameworks and tools used for the crawling and scraping process were ThoughtWorks Limited open source headless browser Selenium and Scrapinghub Limited open source framework called Scrapy. Our methodological approach conformed to the legal and ethical principles of web scraping [213]. Moreover, our research adheres to ethical guidelines on crawling publicly available SCS data raised in [186]. These include protecting the privacy of data subjects at all times and checking for robots.txt files before crawling.

Method

We applied a directed content analysis to map out social norms propagated through role model narratives published on `creditchina.gov.cn`. Directed content analysis draws on existing research when identifying appropriate codes for textual analysis (see, in particular, [214]). We developed four codes based on Tappan and Brown's work on the analysis of narratives about individuals that experience a moral conflict [215]. A first code termed "moral conflict" (Code 1) documented the moral conflict of an individual in a given role model narrative. Next, we developed codes that helped us explore the nature of the moral experience of the protagonist when confronted with the moral conflict. Tappan and Brown suggest that the moral experience of an individual in the context of moral conflict requires analysis of the *cognitive*, *affective*, and *conative* dimensions of the protagonist's experience [215]. These codes

allowed us to pose the following questions: given the moral conflict, *what does the protagonist think?* (Code 2); *what does the protagonist feel?* (Code 3); and *what does the protagonist do?* (Code 4). Codes 2, 3, and 4 made the reflective, emotional, and behavioral dimensions of the moral experience intelligible.

We also wanted to understand whether the assignment of a single virtue or vice led to the attribution of other virtues or vices, respectively. We termed this code "virtue/vice cascade" (Code 5). First, being attributed multiple virtues for carrying out a specific virtuous act indicates a special importance of this virtue. Second, this code allowed us to define the broadness and specificity of the SCS conceptualization of its key virtues "honesty and trustworthiness" (as outlined in the official SCS documents, see [116]).

Table 3.5: Coding scheme for "blameworthy" role model narratives.

Categories	Codes	Examples
Narrative context	(1) Decision scenario	Owing debts of 30 million RMB
	(2) The protagonist's thoughts	"It is only 2000 RMB. I do not have to repay."
	(3) The protagonist's feelings	"I feel deeply regretful".
	(4) The protagonist's actions	Refusing to repay debt with various excuses.
Virtue/Vice	(5) Vice cascade	He fails to repay debt, ..., he lied.
Social norm expression	(6) Injunctive norm	"Neighbors will not come into contact with the Lao Lai."
Identity	(7) Identity labeling	"Lao Lai (老赖)" Owing debts of 30 million RMB... still lives a luxury life.

Furthermore, we took into account social norm messages that have proven to be effective in nudging individuals into a desired behavior [216, 217]. Two types of social norm messages are typically distinguished: *injunctive* and *descriptive* social norm messages. Injunctive norms refer to behavior other individuals approve of (e.g., 80% of individuals think activity x is morally good), while descriptive norms directly refer to the desirable behavior of others (e.g., 80% of individuals engage in desirable activity x) [217, 188, 218]. To avoid redundancy in our analysis (see Code 1 "moral conflict" and Code 4 "the protagonist's actions"), we only used injunctive norms for our analysis (Code 6).

Finally, we applied a code to understand how the author of a role model narrative interpreted the overall moral identity of the protagonist. In role model narratives, authors

construct moral identities [219, 220]. A particular interpretation of the individuals' moral experiences (see Codes 2, 3, 4) by the authors signals the virtues and vices a model citizen, company, or organization is supposed to conform to. As is common in Chinese ethics, virtues and vices tend to be connected to a particular identity ("the moral exemplar"). In order to capture such a moral identity in the role model narratives, we created a code termed "identity labeling" (Code 7). Our final coding scheme included three categories with seven codes in total (for the coding schemes for "praiseworthy" and "blameworthy" narratives, respectively) (see, e.g., Table 3.5).

3.3.5 Results

Text lengths and SCS keywords: The average length of "praiseworthy" narratives was 1,423.27 Chinese characters, more than two times longer than that of "blameworthy" narratives (544.77 Chinese characters). "Praiseworthy" but not "blameworthy" narratives featured either a real photo of the protagonist (46 narratives) or an audio recording of the narrative (50 narratives).

A word frequency analysis revealed the terms "honest/诚实", "trustworthiness/守信" and "honest and trustworthy/诚信" were mentioned altogether 348 times in "praiseworthy" narratives. In "blameworthy" narratives, the contrary concept "untrustworthy/失信" was mentioned only 145 times. However, we found that the term "Lao Lai/老赖" appeared 198 times across "blameworthy" narratives and at least once in every "blameworthy" narrative in our sample. "Lao Lai" refers to individuals or companies that do not repay debt and is commonly known as a substitute of "dishonest person subject to enforcement (失信被执行人)".

Finally, we wanted to understand the occurrence of different SCS-specific and non-specific sanction and detection measures in "blameworthy" role model narratives (see Figure 3.26). 36 "blameworthy" narratives included the term "blacklist". "Public shaming" was explicitly mentioned in 16 of the "blameworthy" narratives. Here, the protagonist's personal information (e.g., passport photo) was posted either online (e.g., social media) or offline at bus stops in the protagonist's living area. 23 "blameworthy" narratives used the term "joint punishment". In these narratives, the protagonist failed to repay debt and was subsequently banned from taking high-speed trains, boarding flights, participating in village elections, departing from and entering China, applying for loans from the bank, gaining job promotions as a public servant, and/or indulging in luxury consumption. In five narratives, the "joint punishment" mechanism sanctioned the protagonist's family members. For example, the protagonist's child could not go to a private school (with high tuition fees) due to the father's transgressions (a measure that is also formulated in the relevant SCS policy document).

Other narratives described how the government was capable of effectively capturing "Lao Lai". "Temporary control" (临控)" is an online or offline surveillance measure operated by the public security organs to monitor an individual's activities. Online accounts and information taken from social media were collected to track the protagonist in four narratives. In three narratives, other surveillance strategies were applied such as video surveillance. "Blameworthy" narratives also highlighted data sharing practices between public security

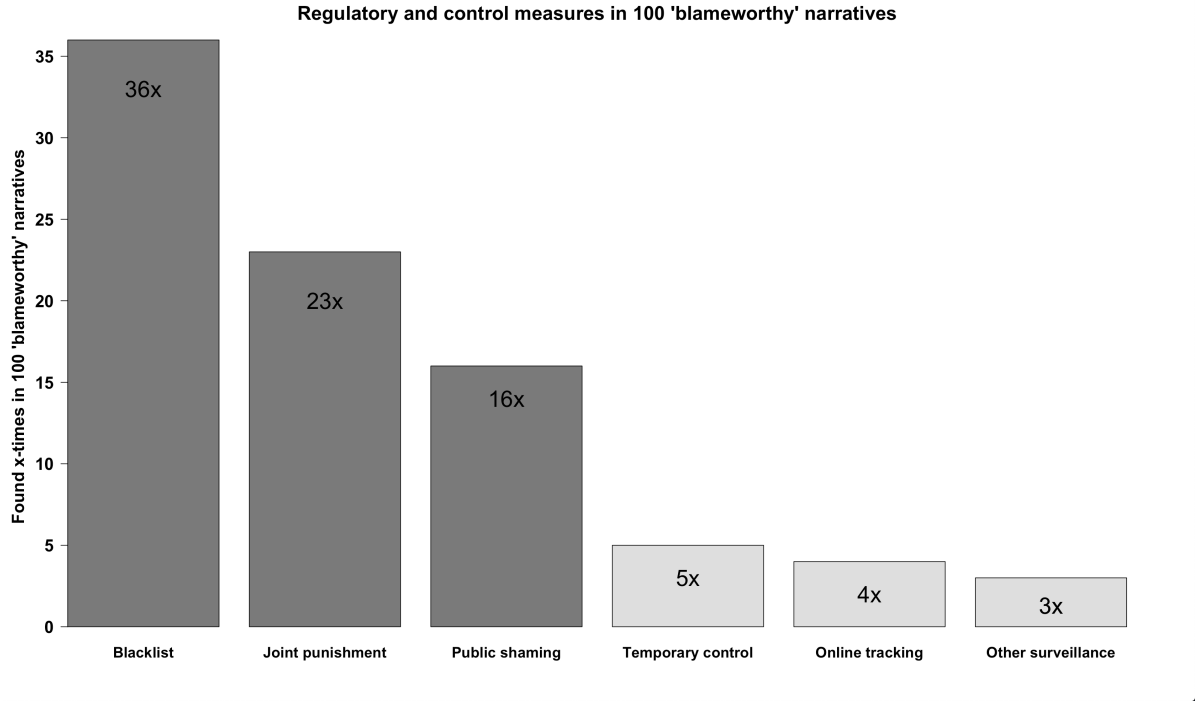


Figure 3.26: Number of different regulatory and control mechanisms in "blameworthy" narratives. Dark gray: SCS-specific mechanisms. Light gray: three other types of regulatory and control mechanisms including online tracking (e.g., social media tracking).

services, hotel registries, and train ticket booking sites for surveillance purposes.

Biographical information of protagonists: Protagonists in "praiseworthy" narratives were individuals. 11 "blameworthy" narratives portrayed companies; eight described a legal representative of the company.

In our sample, 99 "praiseworthy" narratives communicated the gender of the protagonist (75 males, 24 females), 73 "praiseworthy" narratives indicated the age. For "blameworthy" narratives, 49% of the sample indicated the gender of the protagonist (39 males, 5 females). The protagonist's living location was given in 94 "blameworthy" narratives.

Qualitative content analysis

The narrative's storyline

"Praiseworthy" narratives covered a variety of different moral conflicts. These dealt with ostensibly incommensurable trade-offs between protagonists' interests and the interests of the collective (see Figure 3.27). Protagonists were confronted with a moral conflict that tempted them to further their own self-interests at the expense of civic honesty. Protagonists in the "praiseworthy" narratives always chose to be honest towards other members of society. In "praiseworthy" narratives, we identified 141 decisions in total (narratives could include multiple conflicts). 31 of these decisions were about paying back debt or salary. The protagonist

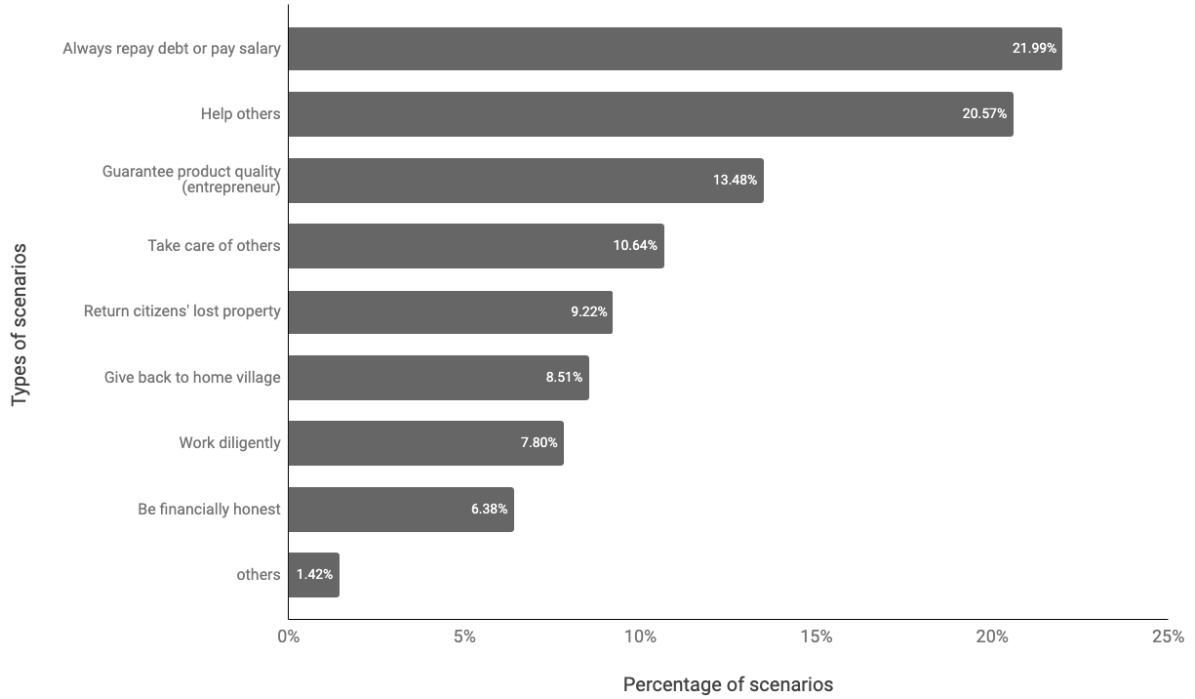


Figure 3.27: Scenario analysis for "praiseworthy" narratives. "Other" mostly referred to various economic virtues: pay employees on time, take care of consumers' rights, and obey the CCP under any circumstances. Numeric values represent the percentages of texts that feature a given scenario.

typically repaid his or her debt faithfully; often despite modest financial possibilities. 29 scenarios showed protagonists helping others financially or non-financially. In another 19 narratives, businessmen guaranteed product quality at the cost of their own economic interest. Other scenarios included taking care of both family and non-family members in various contexts (15), returning lost property of others under various circumstances (13), giving back to one's home village financially and non-financially (12), and working diligently for the public good (11).

All "blameworthy" narratives portrayed an individual who deliberately failed to fulfill a financial obligation, i.e., a repayment of debt—ranging from 300 USD to about 16 million USD. A typical "blameworthy" narrative explained how a Chinese court used various surveillance technologies to identify and sanction "Lao Lai". Across the "blameworthy" narratives, the list of sanctions included exclusion from high-speed trains and any form of political participation, public shaming, detention, and imprisonment.

The protagonists' moral experiences

What the protagonist thinks (cognitive): 95 "praiseworthy" narratives described the cognitive experience of the protagonist when facing the moral conflict (see Table 3.6). Protagonists

Table 3.6: Coding results.

	"Praiseworthy" role model narratives, frequencies (%)	**"Blameworthy" role model narratives, frequencies (%)**
Narrative context of the moral story		
Moral conflict	100% about voluntary sacrifice for public good	100% about debt obligation and the court's action
The protagonist's thoughts	95%	27%
The protagonist's feelings	63%	38%
The protagonist's actions	100% about sacrifice of self-interest	100% about the escape from debt obligation
Virtue & vice cascade		
Virtue cascade	88%	/
Vice cascade	/	16% about vice cascade
Social norm		
Injunctive norm	79%	9%
Identity	100% about honest and trustworthy; 100% justified	100% about "Lao Lai"; 41% justified

either reflected on the importance of being trustworthy in the role they had in society (e.g., as a citizen, lawyer, or doctor) or on the general well-being of others (e.g., "the owner of the lost wallet must be worried").

In contrast, only 27 "blameworthy" narratives described the protagonist's thinking. "Blameworthy" narratives showcased the protagonist's *misrepresentation* of the moral scenario. For example, a "Lao Lai" falsely believed that he was not responsible for the debt and therefore not obligated to repay. In another narrative, a "Lao Lai" with debt falsely thought that the court could not take effective measures against him because of his low economic status. After being threatened with detention he paid back the debt. In another example, an individual owed a relatively small amount of money to another citizen (2000 RMB, around 300 USD) and thought the court would not enforce any sanctions, which turned out to be false.

What the protagonist feels (affective): 63 "praiseworthy" narratives described the emotional state of the protagonist. The most common emotive attitude displayed by protagonists was a "rewarding sense of responsibility" and "satisfaction" as a result of being "honest" toward other citizens.

38 "blameworthy" narratives described how protagonists felt about their behavior. "Lao

Lai" either felt "apologetic" or "regretful" for their actions or feared the consequences of being punished: for example, being detained by the police or being publicly shamed on blacklists. The emotions of "Lao Lai" were described only after their misbehavior had been revealed.

How the protagonist acts (conative): In all "praiseworthy" narratives, individuals acted according to what they believed was expected of them by society: A "good" citizen returns the lost property of another citizen, a "good" doctor treats everybody regardless of their financial background, and a "good" entrepreneur pays employees on time.

In "blameworthy" narratives, protagonists escaped debt obligations by moving to another province, hiding in another family's home, or secretly transferring assets to another person. After the court had taken a certain enforcement action, "Lao Lai" fulfilled the debt obligation. For example, one protagonist lived a luxury life based on debt and frequently showed his wealth on social media. When the individual was identified and punished by public shaming he was reported to have paid back the debt immediately.

Virtue & vice cascade

In our sample, 88 "praiseworthy" narratives featured a "virtue cascade": when protagonists were reported to be "honest" or "trustworthy", protagonists were attributed multiple other virtues. These included diligence, kindheartedness or benevolence, filial piety, and a sense of responsibility to the society.

In contrast, only 16 "blameworthy" narratives featured a corresponding "vice cascade". 11 of them highlighted that a "Lao Lai" was also a "liar". Two "blameworthy" narratives told the story of a "Lao Lai" that was "dishonest" to his friends that had previously helped him.

Injunctive norm expression

79 "praiseworthy" narratives incorporated multiple different injunctive norms such as positive comments from co-workers and villagers, friendly nicknames given by members of the social circle (e.g., "the secretary for children"), and official honorary awards (e.g., "Good People in Anhui Province").

Only 9 "blameworthy" narratives used an injunctive norm. In one "blameworthy" role model narrative, the injunctive norm was expressed by the protagonist: "My neighbours would not come into contact with me once they knew that I am a Lao Lai". In five "blameworthy" narratives, injunctive norms were propagated through the activities and words of relatives who fulfilled debt obligations for the "Lao Lai".

Identity

"Praiseworthy" role model narratives did not include a specific label that served to emphasize a morally ideal identity. In contrast, "blameworthy" narratives fostered a strong link between a specific "immoral" behavior (i.e., deliberately avoiding to repay debt) and a specific "blameworthy" identity, the "Lao Lai". In only one narrative, the individual himself expressed explicitly that he was a "Lao Lai". In all other "blameworthy" narratives (99), the identity

"Lao Lai" was attributed to the protagonist by the authors of the role model narratives. 41 narratives provided a justification for assigning the identity label "Lao Lai" to the protagonist. For example, a "Lao Lai" went on luxurious trips and lived in a high-end hotel while refusing to pay back debt. In the remaining 59 "blameworthy" narratives, however, the authors of the narratives did not justify the attribution of the "Lao Lai" label.

3.3.6 Analysis

Role model narratives underline the SCS's priority for "sincerity" in economic activities

The SCS national platform propagates social norms through narratives focusing on transgressions in the context of economic activities. Across the narratives, businessmen and businesswomen were the most represented profession. Business activities ranged from selling breakfast on the street to producing an annual output worth over 100 million RMB (15 million US dollars). As such, different from traditional Chinese ethical narratives that cover a wide range of virtues, the SCS narratives have a specific focus—moral behaviors in an economic context. In addition, all "blameworthy" narratives reported on an individual or a company that failed to repay debt. This indicates the importance of economic development as a goal of the SCS: China's corporate defaults hit a record high of 62.59 billion RMB (9.67 billion USD) in the first half of 2021.[30] The ratio of household debt to GDP hit an all-time high of 62.4% in September 2021.[31] Investigating individual households, one can observe that the thriftiness culture and the tradition of savings are fading in China [221, 222]. Preventing debt defaults is a pressing economic issue in China and the SCS purports to be part of its solution. The strong focus on the detection and subsequent punishment of "Lao Lai" provides evidence that the SCS makes financial dishonesty very costly.

In addition, the SCS represents a new measure to evaluate the creditworthiness of individuals and companies. The broad conceptualization of "credit" enables evaluation of businesses based on *trustworthiness* rather than on *financial* creditworthiness. Here, SCS redlists and blacklists further aim to decrease informational asymmetry between cooperating entities [223, 224].

SCS role model narratives use ordinary people as moral heroes and familiarize the public with SCS-specific surveillance

A result of reading "blameworthy" narratives is that the readership inevitably becomes familiar with the different forms of technological and administrative surveillance measures. Here, the narrative format allows authors to introduce the state's range of surveillance tools: online tracking, digital blacklisting, temporary control. Narratives clarify the purpose for which they can be used and showcase the near unconditional success of surveillance technologies in finding those that have not complied with laws. Narratives on `creditchina.gov.cn` are able

[30] Data source: *Reuters* at `https://reut.rs/3B6a6H9/`, accessed on May 26, 2022.

[31] Data source: *CEIC* at `https://www.ceicdata.com/en/indicator/china/household-debt--of-nominal-gdp`, accessed on May 26, 2022.

to accomplish what neither the SCS policy documents nor the SCS blacklists or redlists achieve: they combine empirical with fictional elements to portray the power of the state's surveillance apparatus in sanctioning defectors and transgressors. They can be swiftly accessed on the platforms and are easy to read.

Role model narratives use *ordinary* people rather than heroes as moral exemplars. The one-sided emphasis on ordinary people echoes what Turner has referred to as "demotic turn" [225]. It denotes an increasing visibility of ordinary people in mass media. The media not only celebrates ordinary people through reality TV, journalism, radio, and user-generated content but actively creates culturally intelligible identities around them. Scholars of narratives have argued that life stories of ordinary citizens are a *"marker for a society that is losing faith in the more established sacred narratives of religion, preferring more prosaic accounts for advice and guidance"* [226]. In China, there has been an increasing use of ordinary public idols such as socialist heroes and other non-elite figures since the 1950s [227]. Popular Chinese television programs such as *Touching China* (感动中国) and *Civilian Heroes* (平民英雄) illustrate this transformation.[32] However, *currently*, we cannot find a TV program focusing on the SCS specifically. SCS narratives are potentially powerful instruments for propagating SCS-specific social norms to a broad audience. Their sharing to all relevant Chinese social media platforms effectively increases their visibility.

The emergence of the "Lao Lai" as an "immoral" SCS identity

The strict categorization into "praiseworthy" and "blameworthy" role models corresponds to the two ideal moral role models in Confucianism, one of the most prominent traditions of Chinese ethics. In Confucianism, the *Junzi* represents the gentleman (literal translation), while the *xiaoren* literally refers to a "small man" [228]. In the *Analects, Book 4.16*, for instance, Confucius stated that *"The gentleman comprehends righteousness; the small man comprehends profit"*. In traditional Chinese narratives, a particular virtue is exemplified across different social scenarios by the *junzi*, or in contrast, by the *xiaoren*. Such an exemplary person displays virtuous or immoral acts for the public to imitate or to refrain from, respectively. It is for this reason that Chinese ethics is often referred to as "exemplarism" [229], whereby ethical judgment is fundamentally based on "analogical reasoning" [230, 208]. The communication of such "exemplarism" unfolds best in the narrative format: stories inspire an audience to strive for the moral character of the *junzi* or to refrain from being labeled as the *xiaoren*.

Authors of role-model narratives deliberately use stylistic features to strengthen the distinction between "praiseworthy" and "blameworthy" moral characters. "Praiseworthy" narratives attempt to create sympathy and empathy with protagonists when they illustrate the reflective and emotional dimensions of virtuous intentions and convictions. The presentation of a photograph and the detail of biographical information further emphasize that protagonists are worthy of moral emulation in "praiseworthy" narratives. In contrast, the lack of a visual depiction and the informational reduction to a stereotypical label "Lao Lai" of protagonists in "blameworthy" narratives aim to produce a dissuasive effect. The attribution of the label "Lao

[32]Both TV programs focus on the moral lives of ordinary Chinese citizens.

Lai" lacks justification. In "blameworthy" narratives, protagonists' intentions and beliefs are revealed retrospectively, concealing the reasons that led to the borrowing of money and the subsequent failure to repay.

Generally, "blameworthy" narratives do not specify why the protagonist is in a debt situation in the first place. While there are many—perfectly justifiable—reasons why a person can end up in a debt situation (e.g., sickness, loss of employment), authors of "blameworthy" narratives only attended to the reflective and emotional experience of protagonists after they have been captured and sanctioned. An insufficiently justified identity label likely creates stereotypying and possibly discrimination against members of this group [231, 232]. Labels function as external identity markers, constituting an influence on an individual's identity beyond the individual's control [233]. Being assigned such a label may carry a number of negative connotations, treating an individual as if they were generally rather than specifically in the wrong. Subsequently, such individuals could be gradually cut off from participation in more conventional (group) activities, denied ordinary means of carrying out the routines of everyday life, and may eventually find themselves in social isolation. As is illustrated by the "blameworthy" narratives, reports on "Lao Lai" regularly appear on TV news programs, in newspapers, on websites, on social media, or in public areas such as train stations and bus stops.

In a recent study on the relationship between folklore and economic prosperity in 958 societies, Michalopoulos & Xue find that the depiction of "tricksters" or "cheaters" is among the most common archetypes in narrative traditions around the world [234]. Importantly, cultures with more narratives on tricksters that are unsuccessful and that get punished for their antisocial behavior are more trusting and prosperous today than cultures with narratives in which tricksters often get away. The authors argue that such *"folklore-based measures of historical attitudes are robust predictors of contemporary values and economic choices"* [234]. Observing that "Lao Lai" are always identified, captured, and sanctioned in the role model narratives we studied, leads us to believe that SCS narratives could work as powerful portraits of antisocial behavior in Chinese society nowadays.

3.3.7 Concluding remarks

We analyze 100 "blameworthy" and 100 "praiseworthy" role model narratives on `creditchina.gov.cn`. We find that these narratives help to instill a sense of "folk morality", showcasing, partly empirically and partly fictionally, how individuals comply with social norms, how they transgress them, and what consequences they experience. By authorial choice, narratives are rich in biographical detail, which helps readers believe in their presented realities. They are short stories and, as such, everything they contain is there for a reason. Indeed, SCS role model narratives are not "just-so stories" that are first and foremost entertaining in nature. They effectively model "blameworthy" and "praiseworthy" social norms in an epistemically viable manner: they explain a particular causal trajectory in the past, reconstructing specific episodes of moral decision-making coherently and vividly. They reflect the author's perceptions on the moral ills of social life in China.

Over time, social norms change, in particular, when societies face enormous challenges. We found that, in May 2020, `creditchina.gov.cn` started publishing narratives on "praiseworthy" and "blameworthy" social norms "necessitated" by the emergence of the coronavirus pandemic.[33] The SCS's *Planning Outline* [116] specifically mandates the application of the concept of "credit" to health care, health services, and public health. When we revisited the platform, we found that it displayed three types of narratives that can be translated into "positive role models/正面典型", "exposure of dishonest conducts/失信曝光", and "how wonderful you are/你有多美". Narratives on "positive role models" appeared to portray companies that have produced and distributed epidemic prevention materials to help fight the crisis. In contrast, narratives on the "exposure of dishonest conducts" focused on companies that—in response to the coronavirus—have jacked up their prices, produced and sold poor-quality or counterfeit epidemic prevention products, posted deceptive advertisements, or committed coronavirus-related tax fraud. These coronavirus-related "blameworthy" narratives also showcased protagonists who have sold wild animals illegally, spread rumors related to the pandemic, and hid or lied about their travel histories to avoid quarantine. The third type of coronavirus narrative "how wonderful you are" portrayed protagonists that have responded to the crisis particularly well as professionals (e.g., doctors, nurses, businessmen, etc.) and non-professionals (various types of volunteers). This shows that SCS narratives on `creditchina.gov.cn` can be swiftly adapted to address novel demands for moral "praise" and "blame".

SCS narratives fall back on traditional Chinese narratives that convey ethical values and norms. This can be interpreted as an attempt to disguise novel measures of social control as "old wine in new bottles". To say it in Chinese: 新瓶装旧酒 (roughly translated "using a successful strategy that echoes the past"). At least since the 1950s, however, moral education has never only been about cultivating people's morality in China, but has always been closely intertwined with the political agenda of the CCP [210].

Digital role model narratives keep up with the trend of applying digital technologies as tools of social control; they serve as a political instrument promoting policies, spreading ideology, and shaping public discussion. The familiar format of the narrative contributes to the government's efforts to legitimize a new form of social control through a variety of SCS-specific mechanisms such as blacklisting, public shaming, joint enforcement as well as other means of mass surveillance. Narratives on `creditchina.gov.cn` may seem innocuous to some readers. At the same time, they work as a further building block for the state's increasing surveillance and control over Chinese society.

Acknowledgement

We would like to thank the anonymous reviewers as well as Marianne von Blomberg for their constructive feedback. We further are grateful for funding support from the Bavarian Research Institute for Digital Transformation (bidt). Mo Chen also received funding from the

[33] See `https://www.creditchina.gov.cn/xinxingfeiyanyiqing/`, accessed on May 26, 2022.

Fritz Thyssen Foundation for this research. Responsibility for the contents of this publication rests with the authors.

4 Published Article Part 2: Social Media Classification Procedures

Research Article: Social media profiling continues to partake in the development of formalistic self-concepts. Social media users think so, too. (2022)

Please note that the published articles are slightly modified mainly to allow for unification of format and reference style. References for each research paper appear in the overall bibliography at the end of the doctoral dissertation. Published versions of the research articles are appended to end of the doctoral dissertation in chapter 7.

4.1 Research Article: Social media profiling continues to partake in the development of formalistic self-concepts. Social media users think so, too.

Authors

Severin Engelmann, Valentin Scheibe, Fiorella Battaglia, Jens Grossklags

Publication Outlet

AIES'22: Proceedings of the 2022 AAAI/ACM Conference on AI Ethics, and Society; July 2022; Pages 238–252; https://doi.org/10.1145/3514094.3534192

Abstract

Social media platforms generate user profiles to recommend informational resources including targeted advertisements. The technical possibilities of user profiling methods go beyond the classification of individuals into types of potential customers. They enable the transformation of implicit identity claims of individuals into explicit declarations of identity. As such, a key ethical challenge of social media profiling is that it stands in contrast with people's ability to self-determine autonomously, a core principle of the right to informational self-determination. In this research study, we take a step back and revisit theories of personal identity in philosophy that underline two constitutive meta-principles necessary for individuals to self-interpret autonomously: justification and control. That is, individuals have the ability to justify and control essential aspects of their self-concept. Returning to a philosophical basis for the value of self-determination serves as a reminder that user profiling is essentially normative in that it formalizes a person's self-concept within an algorithmic system. To understand whether social media users would want to justify and control social media's identity declarations, we conducted a vignette survey study (N = 368). First, participants indicate a strong preference for more transparency in social media identity declarations, a core requirement for the justification of a self-concept. Second, respondents state they would correct wrong identity declarations but show no clear motivation to manage them. Finally, our results illustrate that social media users acknowledge the narrative force of social media profiling but do not strongly believe in its capacity to shape their self-concept.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

4.1.1 Introduction

Social media platforms enable advertisers to create and target user audiences based on the identification, processing, and analysis of several thousand user attributes such as likes, interests, beliefs, behaviors, relationships, moral convictions, and political leanings [235, 236, 237, 238, 239, 240]. User profiling techniques infer identity claims of users based on views and clicks, visual data such as images and videos, or the number and types of "followers" or "friends" [241, 242, 243, 244, 245, 246, 237, 247]. There is growing recognition in user profiling and user modeling communities that such profiling techniques create unique ethical challenges [248, 236, 249].

These challenges typically fall back on the inability of users to access, understand, and contest automatically-generated identity claims based on their personal data. Specifically, they arise from the restricted ability of social media users to exercise their right to informational self-determination, a central right of many privacy laws around the world. The right to informational self-determination rests on the fundamental idea that it is critical for individuals to freely and autonomously "self-determine" or "self-develop" [250, 251, 252, 253, 254]. The right to informational self-determination mandates that it is critical for individuals to be able to exercise control over their personal information. In the face of technologies that analyze the sentiment of users based on speech or visual data [255, 256, 257, 237] or that interpret data that users have shared unintentionally [258], the notion of individual control over personal data as a feasible mechanism for informational self-determination is, however, severely challenged.

In this paper, we offer a partly philosophical and a partly empirical account to address this problem field. From a philosophical perspective, we aim to make the following two contributions. First, we return to scholarship on the fundamental value of autonomous self-determination offered by philosophical theories of personal identity. Philosophical theories of personal identity conceptualize necessary *procedural criteria* that enable an individual to form a self-concept. Personal identity is an essentially contested concept and, as such, inherently procedural—disputes on the concept's boundaries are essential to the concept itself [47, 259].[1] In contrast, when essentially contested concepts become subjected to digital formalism, they are fixated by definitions that work optimally only under the constraints of computability. The analysis of theories of personal identity can illustrate to us, perhaps again, the enormous power of social media user profiling in determining all procedural elements that exist between personal data and their analysis as declarations of identity: the power to create user profiles over time, the power to change or correct user profiles when needed, as well as the power to change the rules by which user profiles can be generated, changed, or corrected.

Second, the generation of digital representations of personal identity necessarily creates normative trade-offs. We present one normative trade-off by referring to what we call "model fitness." Here we ask whether the digital representation of an individual's self-concept *should*

[1] Please note that this account focuses exclusively on Western approaches to philosophical theories of personal identity.

align as much as possible with how a person would self-determine in order to respect that person's autonomy. Social media platforms have the power to decide what types of data and what amounts of data are sufficient to justify an identity claim about a user. Social media platforms control "model fitness." We exemplify this phenomenon by referring to the literature on "window sliding" in learning tasks with concept drift adaptation [260, 261, 262, 263] and collaborative filtering [264].

We further take it that the power of social media profiling to make identity claims about billions of users is a strong argument in favor of usable transparency that allows users to view (understand their justification) and correct (exercise control over) such identity claims. Here, we engage in another trade-off: if people could view and correct identity claims of social media profiling, then such identity claims could influence a person's self-concept. Social media identity claims could undermine a person's autonomy to self-determine under conditions of transparency when people see, reflect on, and internalize "how a machine interprets" them. Transparency could empower social media identity claims rather than people's autonomy to self-determine.

Subsequently, we have conducted an empirical vignette study to understand how individuals (N = 368) evaluate social media's identity claims with regard to accuracy, transparency, and control. We find that people believe social media user profiling can make accurate judgments about essential aspects of their personal identity, but that they prefer privacy over accuracy. Moreover, people show a strong desire for transparency defined as the ability to view and understand declarations of identity by social media platforms. While people state that they want to compare whether such identity declarations align with their own self-concept, they believe that these do not influence their self-concept. Our study provides evidence that people assert that social media identity claims do not feed back into their own self-concept when they are made transparent and intelligible.

With this work, we seek to contribute to scholarship on the relation and interaction between humans and their algorithmically generated identity declarations. We provide a philosophical lens on the value of self-determination as the process to justify and control essential aspects of a person's self-concept. The conceptualization of autonomy through personal identity creates a firm foundation for determining the ethical challenges of social media user profiling. With a vignette survey study, we take a tangible step towards understanding how people actually evaluate algorithmic identity declarations by social media platforms.

Before we move on to the next section, we would like to offer a disclaimer: In this work, we do not claim that social media user profiling *generates* personal identity or suggest that the resulting profiles can be considered as equal to a person's self-concept. We do not engage in arguments that draw an ontological comparison between a user profile and the person behind it. In other words, we do not claim that social media user profiling leads to a user profile that *is* the personal identity of the individual. Rather, we observe that social media user profiling procedures possess a unique, technologically-afforded narrative force that computationally fixates the interpretative potential of a person's self-concept. This fixation creates ethical challenges when user profiling algorithms turn a person's personal data into declarations of

identity that a person cannot view, cannot understand, and cannot contest.

4.1.2 Social media user profiling is fundamentally normative

Our analysis considers user profiling procedures for *social media* advertisement. All major social media platforms offer a marketing page with an interface[2] where marketers can select desirable user attributes[3] for targeted advertising.

Previous work on social media profiling has summarized what kind of user attributes social media profiling generates. Such profiles consist of user inferences based on online data (e.g., user-generated content on the platform) as well as offline data (e.g., data integrated from data brokers) [235, 236]. User profiling for social media generates sophisticated representations of users based on demographic information including age or gender as well as information associated with user behaviors, preferences, and intentions [265, 266, 267, 268, 269]. Inferences are in part based on "explicit identity claims" (e.g., explicitly *stated* profession or sexual orientation) as well as on "implicit identity claims." Implicit identity claims are "given off" by an individual rather than consciously communicated [270, 271]. Implicit identity claims are inferences users communicate indirectly, for example, through their affiliations to certain individuals, social or institutional groups, preferences, and interests expressed in a non-specific manner. Explicit and implicit identity claims can comprise behaviors (e.g., clicks or views) and beliefs (expressions of interest, intentions, convictions, etc.) [236, 235].

Social media targeting tools offer marketers the option to select an audience (a group of users) based on whether they "possess" or do not "possess" a desirable attribute. Aimeur has provided a comprehensive list of the types of attributes (i.e., identity claims) analyzed for user profiling including name, age, address, identity of friends, sexual orientation, political views, smoker yes/no, pregnancy/wedding, interests, credit score, home value, and others [258]. To understand the normative dimensions of user profiling on social media, the technological *instantiation* of a user's profile, for example as a feature vector [272], is not significant for this analysis. What is relevant is the algorithmic mapping function implemented to assign attributes to users based on their data. Any mapping process from user data to user inference digitally fixates the interpretative potential of an individual user. We refer to this process as the generation of a *formalistic self-concept*. By essentially determining this interpretative potential within an algorithmic frame, mapping functions become normative, for example, when they prioritize user data to constitute an attribute while failing to consider others.

In philosophy, a person's self-concept is procedural, contextual, and contestable [273, 274, 275]. Recent work in Science and Technology Studies has outlined that profiling socially contested concepts through mathematical formalism without accounting for their full meaning creates so-called abstraction and formalism traps [259]. Abstraction and formalization necessarily involve a process of imperfect translation: no model (or profile) is large enough to include all characteristics of an informational object. Similarly, in philosophy, no single theory

[2] See, for example, Meta audience insights or Instagram audience insights.
[3] We refer to such user attributes as "declarations of identity."

of personal identity contains all constitutive principles that make up personhood. Indeed, it is the disagreement on fundamental conceptual features that creates the essential demarcations of a contested concept such as freedom, privacy, autonomy and so on [47]. In user profiling for social media advertisement, abstraction is constrained by two core conditions: First, by the purpose for which the object is profiled—here for commercial purposes (marketing)—and, second, by the mathematical constraints of computability. Regarding the latter, not all features of an object can be modeled by computational resources; for example, the phenomenological experience of human consciousness cannot— in principle—be captured by computational means.[4] Overall, philosophical theories of personal identity offer a useful conceptual framework to understand the normativity of generating formalistic self-concepts.

4.1.3 Justification and control: two meta-principles of personal identity

In the following section, we detail how three influential theories of personal identity lay out procedural criteria that enable a person to form a self-concept autonomously.[5] Attributable to philosophical scholarship, such procedural requirements are subject to productive dispute. Yet, a body of philosophical scholarship on personal identity [278, 279, 280, 281, 282, 283] agrees on two constitutive meta-principles necessary for individuals to self-interpret autonomously: individuals have the ability to *justify* and *control* essential elements of their self-concept.[6] Some philosophers place the source of individuals' abilities to justify and control essential aspects of their self-concept in the individual only (e.g., [275, 286]); other theorists argue that social agents partake in the formation of a self-concept [282, 278, 280, 283].

Harry Frankfurt's second-order desires

In "Freedom of the Will and Concept of a Person," Harry Frankfurt developed a notion of personal identity grounded in the structure of human will [281]. Humans are capable of evaluating the desirability of their desires. A person also cares about the desirability of their desires. Frankfurt calls such desires "second-order desires" that are desires about desires or wants about wants. The object of a first-order desire is a state of affair, while a second-order desire's state of affair is a first-order desire. The desirability of our desires is ethically significant. For example, a person can want to want to eat in a certain way. Vegetarianism, an ethical principle, governs how a person acts on their first-order desire to eat. Frankfurt argues, "*only humans are capable of reflective self-evaluation manifested in the formation of second-order*

[4] Theories on the phenomenological self by Dan Zahavi [276] develop a notion of personal identity that falls back on phenomenological experience.

[5] Personal identity conceptually differs from theories of *personality*. An account of personality is, for example, the prominent Big-Five (BFM) model of personality [277]. The BFM subscribes to personality theories that suggest personality to consist of context-consistent, quantitatively-assessable, enduring traits. In contrast, personal identity explains *how* individuals come to form a persistent self-concept. While such a self-concept may comprise a set of traits, it is the set of principles by which an individual's self-concept develops that is the focus of philosophical theories of personal identity.

[6] Other conceptualizations of hermeneutic personal identity also highlight—in some way or another—the importance of the two meta-principles of justification and control for a person's self-concept (see, for example, [273, 284, 285]). However, they motivate these principles with a different set of reasons.

desires" [281]. The essence of a person lies in will, however, a person needs to be able to "*become critically aware of their own will*" [281]. Individuals need critical reflection to evaluate which of their desires are desirable. Persons are autonomous in determining which desire they want to be moved by when acting. Repeated identification with a specific second-order desire enables us to truly care for something.

Frankfurt's theory of personal identity clearly presents a strong ideal of what it means to be a person. Individuals are required to engage in reflective justification of their second-order desires to fully qualify as persons. There is little room for ambiguous or even paradoxical desires that clearly constitute human experiences. Frankfurt's conception of personhood is an example of a theory from "within": his principles of personal identity are subjective and can even be criticized as "solipsism." External influences, cultural or social, appear to restrict rather than help strengthen individuals' ability to form a self-concept. Summarizing, Frankfurt's second-order desires stress the need for *justifying* one's self-concept, while the identification with a second-order desire underscores that persons can *control* what principles constitute their self-concept.

Charles Taylor's weak and strong evaluator

The philosopher Charles Taylor deliberately tries to avoid "solipsistic tendencies" and points to the importance of social interaction for the development of a self-concept. Taylor stresses the significance others have for our capacity to evaluate what we desire [282]. Many of our desires, wishes, hopes, attitudes, goals and so on develop only in dialogue with others. Taylor places personal identity between private and public spheres: Privately, a human being is a person because of their reflective self-evaluative capacities that require qualitative articulacy. Publicly, a person necessarily adopts such qualitative articulacy by interaction with other individuals.

Similar to Frankfurt's first-order desires and second-order desires, Taylor distinguishes between so-called "weak" and "strong evaluators"[7]. A weak evaluator simply deliberates different options on the basis of their convenience: their goal is to get the most overall satisfaction. Such an evaluator does not reflect on the qualitative aspects of their choices. Non-qualitative evaluation leads to the selection of a desired object or action because "*of its contingent incompatibility with a more desired alternative*" [283]. A weak evaluator chooses something merely on circumstantial grounds. Their deliberation does not exceed a mere desirability calculation for choices to provide some satisfaction. Taylor claims that persons can evaluate what they are and shape whatever they wish to be on this basis. Different from Frankfurt, however, the freedom to self-interpret takes place between private and social spheres. This freedom (i.e., control) to self-define by evaluation (i.e, by justification) means that persons can be made *responsible* for their self-concept [283].

[7] Arguably, a person that chooses merely on the basis of Frankfurt's first-order desires corresponds to Taylor's weak evaluator.

Maya Schechtman's narrative self-constitution view

The philosopher Marya Schechtman asserts that an autonomous person has the capacity to psychologically organize a stream of events into a culturally accepted form of a narrative "*by which we will come to think of ourselves as persisting individuals with a single life story*" [278]. The elements of a narrative that a person can articulate constitute the person to a higher degree than those elements that a person cannot articulate.

An individual compares, organizes, and relates experiences by culturally-determined standards. It follows that no time-slice—any momentary event that an individual experiences—is in any way definitive for a person's identity. Only when interpreted in the context of the narrative is such a time-slice a meaningful element of a person.[8] Telling a story is only one element of a person's narrative. Individuals form a narrative, but they also enact it and subsequently criticize it: they are not only the authors of their narrative but their protagonists and critics, too. As an author, a person tries to understand the meaning events have by integrating them into their continuous narrative. A person is the critic of their narrative when they come to reflect, evaluate, and criticize the actions they have carried out. While the order in which these steps take place is certainly dynamic, it demonstrates that a person plays different roles within their own narrative—they are not simply describing what they have experienced as a commentator or storyteller in the literal meaning of the term. For Schechtman, a person's narrative is actively *negotiated* between subjective and objective accounts. A person may have their own interpretation of a certain event; however, their identity will be undermined if claims reach a level of incomprehensibility for other people. A person's choices and actions must "*flow intelligibly from (their) intentions, motives, passions, and purposes...*" [278]. Without our narrative context, other individuals cannot make sense of our choices and actions. The narrative view gives individuals freedom to shape (i.e., control) who they wish to be, re-interpret their past and anticipate their future self-concept (i.e., justification). A person's social environment holds a person accountable for the narrative they articulate.

Summary of philosophical theories of personal identity: While differences exist between the theories by Frankfurt, Taylor, and Schechtman, two meta-principles can be discerned: justification and control. First, a self-concept develops through *reflective justification*. Individuals become persons when they justify their self-concept—through reflective capabilities and in a narrative that is negotiated between subjective and objective accounts. Second, individuals can exert some control over their self-concept. While the theories disagree over the degree of control individuals have in forming an understanding of themselves, fundamentally, they all suggest that personhood is grounded in an individual's autonomy to determine essential aspects of their hermeneutic identity. It is for this reason that persons can justifiably be held responsible for their own identity.

[8] *"Whether or not a particular action, experience, or characteristic counts as mine is a question of whether or not it is included in my self-narrative"* [287].

4.1.4 Two normative trade-offs in user profiling for social media marketing

We argue that social media profiling generates digitally formalized identity claims of a person by mechanisms that do not sufficiently allow for justification and control. In the following, we discuss two normative trade-offs that result from the inherent normativity of social media user profiling as discussed in Section 5.1.2.

4.1.5 Normative trade-off 1: The privacy versus model fit trade-off

Concept drift challenges

One normative judgment user profiling is necessarily required to make is to determine when enough data (or evidence) has been collected and analyzed to justify the inference of a person's attribute (i.e., an identity declaration). It is a normative undertaking to decide when the amount of personal data is sufficient to ensure proportionality between the user input and the attribute inference. Is the inference proportional to a single activity or expression of belief? Or is its proportionality dependent on multiple consecutive expressions of the belief? Resolving such questions, user profiling necessarily excludes user input from being considered for drawing user inferences. Schechtman asserts that individuals have the capacity to attribute meaning to a selection of experiences that become part of their own unique narrative. However, it is the narrative that is self-constituting, not the single experience. It follows that no time-slice—any momentary event that an individual experiences—is in any way definitive for a person's identity. Such a time-slice is only a descriptive and meaningful element of a person when interpreted in the *context* of the narrative.

Schechtman's concept of a "time-slice" can be compared to the concept of "window sliding" used in learning tasks with concept drift adaption [260, 261, 262, 263]. Concept drift techniques are deployed to gain knowledge from data stream *changes*. Drifts or changes in a data stream can be either sudden or gradual. The former could be a sudden new interest in a new subject, while the latter could be a growing interest in moving to another country. In user profiling, concept drift belongs to a class of challenges called dynamicity problems [288, 289]. Recommender systems apply dynamic user profiles to offer more value to the user, who sees informational resources they have only recently become interested in, and to the advertiser that can bid for audiences with the most up-to-date profile.

Machine learning (ML) classifiers are able to respond to concept drift—gradual, sudden, or reoccurring changes often in multiple data streams—without "neglecting" the outdated data [263]. For example, sliding windows of fixed and variable sizes of training data are used to build an updated model [260]. Since both fixed and variable windows are definite in their size, some old data will necessarily be "forgotten." What criteria determine which data are to be forgotten and which ones are to be considered in creating an updated profile of a person? The promise of targeted advertisement rests on the belief that more recent user data corresponds to a more accurate profile of the user. However, model fit, a continuously updated model of a user's profile, requires a potentially uninterrupted flow of user data, raising privacy concerns [290]. The more time-slices are created, the more accurate the

representation of the user, but the more user data is needed.

Lookalikes through Neighborhood-Based Collaborative Filtering

Collaborative filtering (CF) is one of the most widely applied user modeling techniques in many recommender systems. For example, as a user profiling technique, *k*-nearest neighbor relies on the assumption of similarity between individuals [264]. Similar profiles presumably react similarly to certain informational items. The advantage of CF is that one only requires a model of one of the two—users or items—to model the other. Consequently, CF uses items to model users and users to model items. The more users evaluate informational resources, the more they help the system for its predictive analysis of other users. Social media (as well as search engines) offer their customers so-called "Lookalike Audiences."[9] With many marketers, Lookalikes are popular since they can use their well-known customer base to target "similar" but potentially new customers. Lookalikes are less privacy-invasive because they use data that is already available to make inferences about a user. Taylor's and Frankfurt's concept of a person, however, stresses the ability of persons to decide what is *desirable for them*. *k*NN-based CF and Lookalikes work in the opposite way. They determine the desirability of one's desires as equal or at least similar to the desirability of other, already "known" individuals' desires, to use Frankfurt's nomenclature.

4.1.6 Normative trade-off 2: The transparency versus autonomy trade-off

A key question is if people would actually care about model fit—an accurate representation of their *formalistic* data narrative. Perhaps individuals do, after all, live in the best of all possible worlds: they draw enormous benefits from using social media and do not worry about how their data is mapped to a spectrum of attribute inferences. One way forward would be to enable individuals to understand and correct inferences they do not agree with. Here, another normative complication emerges. A person could gain autonomy from having access to their social media's identity declarations. However, these identity declarations could in turn influence a person's self-concept.

Should individuals get access in order to understand and contest their "data narrative"? Providing explanations on "how the systems works" has shown to increase users' trust in many different recommender systems [291, 292, 293, 294]. Usable transparency allows users to tell the system when an inference is presumptuous (or even wrong). For example, a system could show users those identity declarations that have been sold to marketers or that were based on implicit identity claims. However, simply revealing—at least in part—the content behind user profiles could support internalization and conformation to the proposed inferences. Perhaps individuals would welcome such a degree of transparency as a mechanism to "offload" the psychological work necessary to attribute meaning to certain life events posted online [278, 279, 280]. Making inferences transparent to the individual means recognizing their semantic power in shaping who individuals are and who they can become. This second

[9] See, for example: `https://www.facebook.com/business/help/164749007013531` accessed May 30, 2022.

normative trade-off arises from the question of whether the autonomy gained from being able to understand such recommended inferences outweighs a potential loss of autonomy when they become part of a person's self-concept. This could mean that, today, a person, their social network (offline and online), *and* social media profiling identity declarations together participate in creating a person's self-concept.

The effect on individuals' self-concept could be enhanced if social media user profiling generates specific identity declarations repeatedly or even permanently. According to Frankfurt's theory of personal identity, a person attempts to form a self-concept that stems from their care for what they desire. Frankfurt recognizes that one can only care about something if it is for extended periods of time. Desires typically last for moments only: if one cared about something for only a moment one could not be distinguished from a person that acted out of impulse. How would users perceive such recommended attribute inferences? Perhaps with little skepticism, since they would acknowledge the algorithmic output as an objective and truthful interpretation of their wishes, wants, and desires?

4.1.7 Methods and Experimental Procedure: Vignette Study

To address the key questions arising from both normative trade-offs, we conducted a vignette study that asked respondents a) whether they believed social media profiling could accurately infer elements of their self-concept, b) whether they considered accuracy of these identity declarations to be desirable, c) whether they had motivation to view and correct identity declarations, and d) whether they believed that social media identity declarations would influence their self-concept if they were made transparent to them. The goal of the vignette study was to take a tangible step towards understanding whether social media users preferred accuracy of social media identity declarations over privacy (trade-off 1) and whether they believed that social media identity declarations would influence their self-concept (trade-off 2). Vignette studies have been extensively used in human computer interaction, psychology, and experimental philosophy to elicit participants' explicit ethical judgments in various hypothetical scenarios [106, 107, 108, 33, 295, 111, 296, 297, 298]. Moreover, with our vignette survey study, we follow calls for more experimentally-informed AI ethics [299].

Our study was a within-subject design, we presented each respondent with the same hypothetical vignette scenario. First, the vignette asked respondents to imagine that they are active users on a social media platform (see the hypothetical vignette scenario in Appendix 4.1.10). As an active user, each respondent was told that they regularly engage in typical actions on the social media platform. Participants read that they publish postings, share postings by other users, and react to other users' postings. Second, the vignette introduced examples of data types each respondent shares with the social media platform (gender, location, relationship status, social contacts, content viewed, content clicked, etc.). Respondents were told that the social media platform uses algorithms to draw conclusions about them based on the data they share in order to show them more suitable content and advertisements. Third, the vignette elaborated on the types of conclusions (i.e., identity declarations) that the social media platforms draws about them. The vignette explained that the platform collects data

that users actively share to draw conclusions about them. For example: *"...when you provide your real birthday, the platform uses this information to show you content that it takes to be suitable for your age group."* Respondents were also told that the platform draws conclusions about users based on data that users may not be aware that they are sharing. For example, *"...since you share your location data, the platform tries to conclude where you work and live. As another example, the platform also tries to conclude what hobbies you have based on your friends' activities."* Respondents were further told that, using their data, the social media platform attempts to conclude their interests, their political orientation, their religious beliefs, and aspects of their personality (among others). Lastly, we asked respondents to imagine that the platform "combines and stores" all conclusions about them in a so-called "user data profile" (UDP). The vignette explained to users that the social media platform uses the content of their UDP to recommend relevant information and advertisements. The hypothetical vignette scenario ended by telling respondents that the social media platform generates all of its revenues from personalized advertisement. We included two attention checks in the vignette. All participants were active social media users.

After respondents had read the vignette and passed the attention checks, they rated questions using a 7-point Likert scale. Questions were divided into 5 categories and shown to respondents in random order within these categories. The first two categories of questions asked respondents whether they believed social media platforms could make accurate judgments about them and whether the social media platform *should* make accurate judgments about them. We defined accuracy as a) general judgments, b) specific judgments, and c) temporal judgments. The third and fourth set of questions asked whether respondents desired to view and understand social media judgments about them and whether they would change incorrect judgments. Questions on respondents' preference for transparency included a) data collection & use, b) preference for understanding conclusions of the social media platform, and c) preference for transparency of their UDP (i.e., all identity declarations). Finally, a fifth set of questions asked respondents whether social media judgments would have an influence on their self-concept given that respondents could view their UDP. We defined "influence" as respondents' willingness to a) compare elements of their UDP with their self-concept, b) their willingness to reevaluate their self-concept in light of the identity declarations in their UDP, and c) their willingness to integrate elements of their UDP into their self-concept that they would not have associated with their self-concept. All questions are listed in Appendix 4.1.10.

We recruited participants with Prolific. Based on pretests, we set the expected completion time at 20 minutes, with a payout of USD 3.75 (above US minimum wage of 2021). Data collection started on July 26, 2021 and ended on August 8, 2021. We recruited 458 respondents from the United States user base. 59 submissions were excluded for failing one of two attention checks, 10 for duplicate submissions, 9 for an unusually short response time, and 11 for being invalid (e.g., no prolific ID). This resulted in a final sample of 368 respondents (see demographics in the Appendix 4.1.10). The mean time of completion was 15.3 minutes.

Our home institution does not require an ethics approval for questionnaire-based online studies. When conducting the study and analyzing the data, we followed standard practices

for ethical research: presenting detailed study procedures, obtaining consent, not collecting identifiable information or device data, and using a survey service[10] that guaranteed compliance with the European Union's General Data Protection Regulation. The study did not include any deceptive practices. Subjects could drop out of the study at any point. All data were fully anonymized, and the privacy of all subjects was maintained at all times during the study.

4.1.8 Results

Respondents' beliefs on the ability of social media platforms to make accurate judgments about them (Fig. 4.1a). A majority of respondents believed that social media algorithms could make accurate and correct judgments about them *in general* (78.2%). While 66.2% of respondents were convinced that social media algorithms could correctly judge "what is valuable to them," just over half of respondents said that social media algorithms can accurately reflect who they are (51.1%). Most respondents believed that their UDP was unique in comparison to other social media users (72.6%). However, only a minority of respondents said that family and close friends would be able to identify them by their UDP (45.5%).

Respondents' beliefs on the ability of social media platforms to make accurate judgments about them on specific attributes (Fig. 4.1b). Respondents believed that social media algorithms can accurately infer their interests (89.9%), their past (81.3%) and future purchasing behaviors (64.5%), as well as their location (77.4%). Just over half of those surveyed stated that social media algorithms could accurately conclude who they meet (54.8%).

Respondents also said that social media algorithms are able to accurately conclude their political stance (80.5%) and, albeit with less agreement, their religious beliefs (59.5%). Most respondents agreed that social media algorithms can correctly infer their attitudes towards the COVID-19 vaccine (77.9%), climate change (74.6%), and immigration (64.8%). However, respondents did not think that social media profiling was able to differentiate between their private and social self both online (35.5%) and offline (30.7%).

Respondents' beliefs on the ability of social media platforms to make accurate temporal judgments about them (Fig. 4.1c). Respondents believed that social media algorithms are able to keep their UDP up to date (71.4%). Respondents stated that their UDP from a month ago still included accurate conclusions (69.1%). However, just over half of respondents thought that their UDP from a year ago was still accurate (51.4%). A majority of respondents said that the social media platform would be able to conclude whether they had changed as a person after several years of being a user (68.9%). In contrast, only a minority of respondents believed that their *entire UDP* would tell an accurate story of their life since they started using the platform (37.9%).

Respondents' beliefs on the normativity of accurate social media judgments (Fig. 4.2). Most respondents stated that they wanted social media platform operators to ensure that their UDP was accurate (72.4%). Just more than half of respondents wanted social media operators to invest extra resources to make sure their UDP was accurate (56.9%). However,

[10] SoSci Survey: `https://www.soscisurvey.de/`

Figure 4.1: (a) Respondents believe social media platforms (SMP) can make accurate judgements about them. UDP—user data profile. (b) Respondents believe social media (SM) algorithms are able to accurately infer a variety of attributes including their interests, purchases, location, political stance, or religious beliefs. Respondents do not believe SMP is able to distinguish who they are in private vs. who they are in social contexts. (c) Respondents believe SMP is able to keep their UDP up to date, but that their UDP does not tell an accurate story of their life. Note for all figures: results for "strongly agree" and "agree" are shown as "agree," results for "strongly disagree" and "disagree" are shown as "disagree."

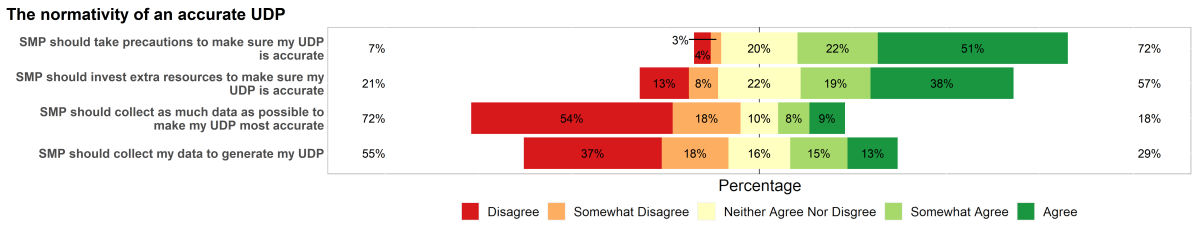


Figure 4.2: Respondents prefer an accurate UDP but not at the expense of their privacy.

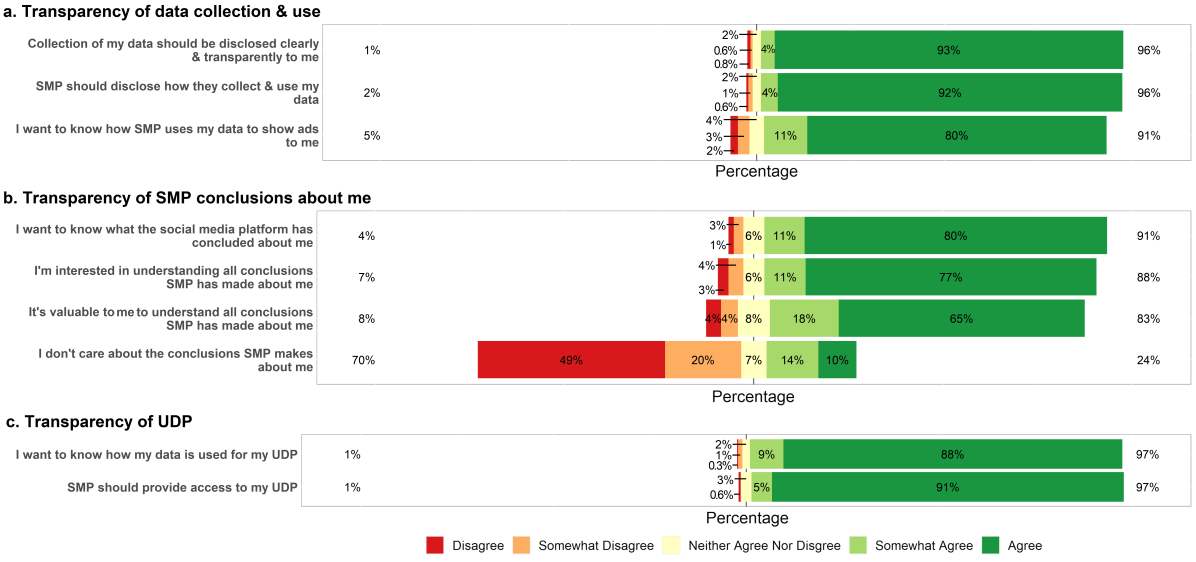


Figure 4.3: Respondents show great preference for transparency of (a) personal data collection & use, (b) conclusions SMP has made about them, and (c) of their UDP.

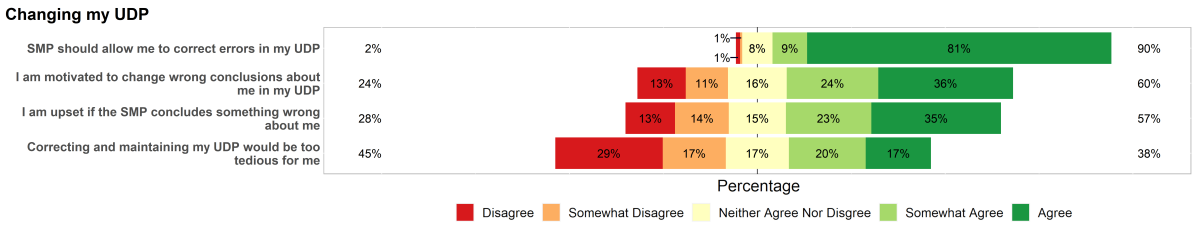


Figure 4.4: Respondents state that the SMP should allow them to correct errors in their UDP but provide no clear preference on whether they would be willing to correct and maintain their UDP.

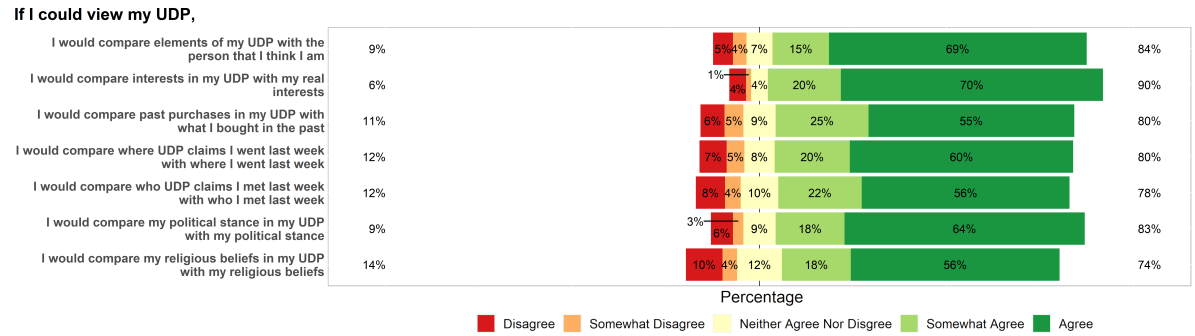


Figure 4.5: Provided their UDP was transparent, respondents would compare elements of their UDP with a range of personal attributes.

only a minority of 28.8% of respondents were in favor of trading their personal data for the creation of their UDP. Importantly, respondents did not want to trade their personal data for an accurate UDP: only 17.9% agreed that the social media platform should collect as much personal data as possible to ensure that their UDP was as accurate as possible.

Respondents' preference for transparency of data collection & use (Fig. 4.3a). Respondents expressed their desire for transparency of personal data collection on social media, transparency of conclusions the social media platform made about them based on their data, and transparency of their UDP. Regarding data collection, most respondents stated that procedures of data collection should be disclosed clearly and transparently to them (96.4%) and that the social media platform should disclose how they collected and used their personal data in general (96.1%) and for showing advertisements (91.1%).

Respondents' preference for transparency of conclusions (Fig. 4.3b & c). Similarly, respondents showed a strong preference to understand what the social media platform has concluded about them (90.1%). Of the respondents, 87.7% stated that they were interested in understanding all conclusions the social media platform had made about them and 83.2% believed that such an understanding would be valuable to them. Only 24% of respondents stated that they do not care about conclusions the social media platform draws about them. Finally, similarly large majorities of respondents expressed their desire to understand how their personal data was used to create their UDP (96.7%). Of the respondents, 96.6% said that they wanted access to their UDP in general (Fig. 4.3c).

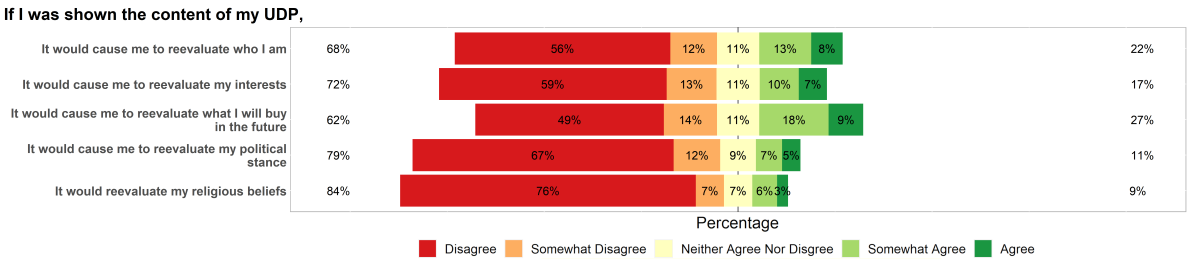


Figure 4.6: Respondents strongly believe that viewing the content of their UDP would not cause them to reevaluate elements of their self-concept.

Respondents' preference for control over their UDP (Fig. 4.4). While respondents showed a clear preference for transparency, their desire to control (i.e., change or otherwise influence) their UDP was mixed. A majority stated that the social media platform should allow them to correct errors in their UDP (90.2%). However, only a small majority said they would be motivated to change wrong conclusions in their UDP (60.2%). When we asked whether correcting and maintaining their UDP would be "too tedious," respondents showed no clear preference (agree: 37.8% vs. disagree: 45.4%, neither: 16.8%). Approximately half of respondents (57.3%) believed they would be upset if the social media platform concluded something about them that they thought was incorrect.

Respondents' beliefs on the influence of the UDP on their self-concept (comparison UDP vs. self-concept, Fig. 4.5). Provided they had access to their UDP, the majority of respondents maintained that they would compare elements of their UDP with the person they thought they were (84.1%). Most respondents said they would compare interests in their UDP with their real interests (89.7%). Respondents further stated they would compare past purchases (79.9%), past locations (79.9%, "last week"), and past social meetings (77.9%, "last week") with those in their UDP. Among the respondents, 82.7% would compare their political stance with the one registered in their UDP and 74.3% of respondents would compare their religious beliefs with those in their UDP.

Respondents' beliefs on the influence of the UDP on their self-concept (reevaluation of self-concept, Fig. 4.6). Only a minority of respondents believed that viewing their user data profile would result in a reevaluation of their self-concept (agree: 21.5%). Few respondents stated that they would reevaluate their interests (agree: 17.3%), their future purchases (agree: 26.8%), their political stance (agree: 11.5%) or their religious beliefs (agree: 9.22%) after viewing their UDP.

Respondents' beliefs on the influence of the UDP on their self-concept (meaning of unaware identity declarations in the UDP, Appendix Fig. 4.7). Respondents were undecided whether social media conclusions were meaningful to them (agree: 47.3% vs. disagree: 35.6%). A small majority of respondents disagreed that conclusions about them in their UDP—that they did not know about—would be meaningful to them (disagree: 53.8%). A small majority of respondents also objected to statements saying conclusions about their political stance (disagree: 55.0%) or religious beliefs (disagree: 59.9%) in their UDP—that they did not know about—would be meaningful to them. Finally, we asked respondents whether

their UDP would be a source of inspiration when looking for a *new* interest. Only 41.1% of respondents said that they would look into their UDP for suggestions on new interests. Likewise, respondents believed that when they saw an interest in their UDP that they would not have believed to be their interest, then this "recommended" interest would not become a new interest for them (agree: 36.6%). An even smaller minority of respondents said that predicted purchases in their UDP would influence actual future purchases (agree: 34.6%).

4.1.9 Discussion of results and concluding remarks

In this work, we argued that the computability of digital representations of personal identity creates normative trade-offs when social media profiling generates identity claims that work only under the constraints of computability and that people cannot understand, view, or contest. Consequently, one of the key ethical challenges of social media profiling is that it stands in contrast with people's ability to self-determine freely and autonomously. To illustrate the inherently procedural nature of autonomous self-determination, we revisited theories of personal identity in philosophy that underline two constitutive meta-principles: justification and control. That is, individuals have the ability to *justify* and *control* essential elements of their self-concept. The return to the philosophical basis for the value of self-determination serves as a reminder that social media profiling represents an inherently normative formalization process of a person's self-concept. Within the interpretative space between data and declaration, social media platforms determine the meaning of views, clicks, posts, and social relationships without offering usable means for understanding or correcting essential parts of this process. As such, social media identity declarations are radically different from the procedural criteria laid out by theories of personal identity in philosophy.

Taking a step toward understanding how "ordinary" social media users view social media identity declarations, we conducted a vignette survey study. We found that people believe that social media platforms can make a variety of accurate judgements about them but that they cannot represent their entire self-concept. For example, respondents thought that social media profiling is able to accurately infer whether they have changed as a person over time, but that it cannot tell an accurate story of their life since signing up to the platform. Thus, respondents defined limits for the ability of social media identity declarations to represent certain aspects of their self-concept. Interestingly, respondents did claim that their own user data profile (UDP) was unique and that other users had a different UDP.

Respondents showed a strong preference for more transparency and stated that they would compare their own self-concept with a variety of social media identity declarations. However, the respondents in our study did not believe that social media identity declarations would be meaningful to them. Respondents also stated they would correct wrong identity declarations but showed no clear motivation to manage them. Taken together, we believe that it is reasonable to assume that social media users have at least some motivation to control essential aspects of their social media identity declarations. Providing such identity controls does present technological as well as design challenges for social media platform operators. However, social media platforms go to great lengths to offer advertisers usable controls to

specify which user attributes exactly they wish to include in their custom audiences. In providing usable justification and control, social media platforms give priority to advertisers determining detailed custom audiences for targeted advertisement over giving users the possibility to understand, control, and rectify potential inaccuracies in their user profiles.

Finally, respondents did not believe that social media identity declarations would *influence* their self-concept. Respondents stated that previously unknown identity declarations would be unlikely to become part of their self-concept and they strongly objected that viewing social media identity declarations would cause them to reevaluate their self-concept. Future studies should try to understand whether people's self-concept is resilient to social media identity declarations as participants stated in our study. Perhaps people are overconfident in the immunity of their self-concept against social media declarations? Also, a majority of respondents expressed the desire to compare components of their UDP with their self-concept. Considering our results, we take it that people are, at least, curious to understand how social media platforms interpret them based on their personal information. They acknowledge the narrative force of social media profiling but do not strongly believe in its capacity to shape their self-concept. We encourage future studies to explore whether our findings extend to social media users in other cultures.

To conclude, we have focused on the *process* by which social media generate identity declarations based on personal information through user profiling. In comparison to the large corpus of studies that have focused on the *consequences* of user profiling (e.g., filter bubbles, misinformation), philosophical accounts on the procedural aspects of social media user profiling remain scarce. While our vignette study produces an initial understanding of the relationship between social media users and their identity declarations, we expect that this account provides ample opportunity for follow-up studies on the ethical challenges of social media profiling. Social media will continue to exercise its power to partake in the formation and development of formalistic self-concepts. We provide evidence that social media users think so, too.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive feedback and comments. This research was supported by a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in any part of the research or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

4.1.10 Appendix Study Materials

Hypothetical Vignette Scenario

Please read the following scenario carefully:

Imagine that you are an active member of a global social media platform. Think of a social media platform that is similar to a handful of prominent examples such as Facebook, Twitter, or Instagram. Imagine that, on this platform, you are an active member and regularly post content. For example, you frequently upload images to the platform. When your friends publish similar posts, you commonly "react" to their posts. Generally, you often consume the content that the platform presents to you in its so-called "news feed".

More specifically, the data you share with the platform includes your real name, your age, and gender. You also share your current location with the platform, your social contacts, your relationship status, the type of device you use, and your activity data: what content you view and click on when you use the platform and at what time you do so. You are aware that the social media platform has developed algorithms that attempt to draw a variety of conclusions about you based on the types of data you share. The social media platform states that it uses such "conclusions about you" in order to show you more suitable content and product advertisement.

Some conclusions may be based on the data you share actively and consciously. For example, when you provide your real birthday, the platform uses this information to show you content that it takes to be suitable for your age group. Some conclusions about you are based on data that you share implicitly, so you may not be aware that you have shared such data about you.

For example, since you share your location data, the platform tries to conclude where you work and live. As another example, the platform also tries to conclude what hobbies you have based on your friends' activities. The platform attempts to conclude your interests (e.g., movies, music, or books you might like) and your behaviors (e.g., what you buy, who you meet). It tries to conclude your religious beliefs (e.g., whether you are part of a religion or an atheist) and your political stance (e.g., whether you consider yourself liberal or a conservative). The social media platform also tries to draw conclusions about who you are as a person more generally; for example, how you might react to certain content, how introverted or extroverted you are, or how sociable you are.

Now, please imagine that the social media platform combines and stores all conclusions about you in your user data profile. Again, the social media platform claims that it needs the content of your user data profile to know what content and advertisement you find suitable. The social media platform generates all of its revenues by showing you advertisements.

To recap, there are two different user profiles on social media: One profile that you use to share posts or share messages, your profile on the social media platform. The other one is generated by the social media platform about you, which will be referred to as your "user data profile" for the rest of the survey. All survey questions relate to your user data profile, not your social media profile.

You were shown a description of a social media platform. You will now be asked questions regarding your personal perception of social media platforms like the one described previously. All questions relate to a social media platform that was introduced to you in the opening text.

Please answer these questions from your own point of view.

Manipulation Checks (in-text)

1. Asked prior to vignette text: It is important that you pay attention to this study. Please read the scenario described below carefully.

    - *Please confirm this by selecting "Strongly disagree."*

2. Asked at the end of the vignette text: Please indicate which of the following is true.

    My user data profile is:

    - *My social media profile that I use to socialize when I log on to the social media platform.*
    - *My profile that the social media platform's algorithms generate about me based on the data I share explicitly and implicitly.*
    - *I don't know.*

Survey Questions

7-point-scale, 1 = "Strongly disagree" to 7 = "Strongly Agree," and "I don't want to answer."

Questions were divided into 5 categories and shown to respondents in random order within these categories. Participants did not see headlines of question categories.

Accurate judgments (general)

- *The social media platform is able to draw correct conclusions about me.*
- *I believe that the social media platform is able to know what is valuable to me.*
- *I believe that my user data profile is unique. Other users have a different user data profile.*
- *If close friends and family saw my user data profile, they would be able to identify that it's me.*

Accurate judgments (specifics)

- *I believe that social media algorithms are able to accurately conclude what my interests are (e.g., movies, music, or books I like).*
- *I believe that social media algorithms are able to accurately conclude what I have bought in the past.*

- *I believe that social media algorithms are able to accurately conclude what I will buy in the future.*
- *I believe that social media algorithms are able to accurately conclude where I go.*
- *I believe that social media algorithms are able to accurately conclude who I meet.*
- *I believe that social media algorithms are able to accurately conclude my political stance.*
- *I believe that social media algorithms are able to accurately conclude my religious beliefs.*
- *I believe that the social media platform is able to know where I stand on important issues such as my acceptance of the Covid-19 vaccination.*
- *I believe that the social media platform is able to know where I stand on important issues such as climate change.*
- *I believe that the social media platform is able to know where I stand on important issues such as immigration.*
- *The social media platform is able to distinguish between who I am in private and who I am in social contexts on the social media platform.*
- *The social media platform is able to distinguish between who I am in private and who I am in social contexts when I am not online.*

Accurate judgments (temporal)

- *The social media platform is able to keep my user data profile up to date with my interests, behaviors, and beliefs as they change over time.*
- *My user data profile from a month ago includes conclusions about me that are still accurate today.*
- *My user data profile from a year ago includes conclusions about me that are still accurate today.*
- *After having been an active user on the social media platform for several years, the platform can conclude whether I have changed as a person since I started using the platform.*
- *My entire user data profile tells an accurate story of the life that I have lived since I started using the platform.*

The normativity of an accurate UDP

- *The social media platform should take precautions to make sure that my user data profile is accurate.*
- *The social media platform should double-check my user data profile for accuracy, even if it takes them time or possibly other resources (e.g., money or additional employees) to do so.*

- *The social media platform should collect as much of my data as possible to ensure my user data profile is as correct as possible.*
- *The social media platform should collect my data to generate my user data profile.*

Transparency of data collection & use

- *The collection of my data should be disclosed to me clearly and transparently.*
- *The social media platform should disclose the way they collect and use my data.*
- *I want to know what data the social media platform has used to show advertisements to me.*

Transparency of SMP conclusions about me

- *I want to know what the social media platform has concluded about me.*
- *I am interested in understanding all the conclusions the social media platform has made about me.*
- *It is valuable to me to understand all the conclusions the social media platform has made about me.*
- *I do not care about the conclusions that the social media platform makes about me.*

Transparency of UDP

- *It is important to me that I am aware and knowledgeable about how my personal data will be used for my user data profile.*
- *The social media platform should allow me to access my user data profile.*

Changing my UDP

- *The social media platform should allow me to correct errors in my user data profile.*
- *I am motivated to change conclusions that I think are wrong in my user data profile.*
- *I am upset if the social media platform concludes something about me that I think is wrong.*
- *If my user data profile was made transparent to me, then correcting and maintaining my user data profile would be too tedious for me.*

If I could view my UDP

- *If I had the ability to view my user data profile, I would compare elements of the user data profile to the person that I think I am.*

- *If I had the ability to view my interests (i.e., movies, music, or books that I like) in my user data profile, I would compare them to my own real interests.*
- *If I had the ability to view what the social media platform claims I have bought in the past, I would compare it to what I have actually bought.*
- *If I had the ability to view where the social media platform claims I went in the past week, I would compare it to where I really went last week.*
- *If I had the ability to view who the social media platform claims I have met in the past week, I would compare it to who I met in the past week.*
- *If I had the ability to view my political stance in my user data profile, I would compare it to my own real political stance.*
- *If I had the ability to view my religious beliefs in my user data profile, I would compare them to my own real religious beliefs.*

If I was shown the content of my UDP

- *If I was shown the content of my user data profile, it would cause me to reevaluate who I am.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my interests (i.e., movies, music, or books that I like).*
- *If I was shown the content of my user data profile, it would cause me to reevaluate what I will buy in the future.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my political stance.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my religious beliefs.*

Influence of self-concept (unaware elements)

- *If I could view the content of my user data profile, then the conclusions the social media platform has made about me would have meaning to me.*
- *If my user data profile contains conclusions about who I am that I did not know about, then these conclusions don't have meaning to me.*
- *If my user data profile contains conclusions about my political stance that I did not know about, then these conclusions don't have meaning to me.*
- *If my user data profile contains conclusions about my religious beliefs that I did not know about, then these conclusions don't have meaning to me.*
- *If I was looking for a new interest, I would look into my user data profile for a suggestion.*

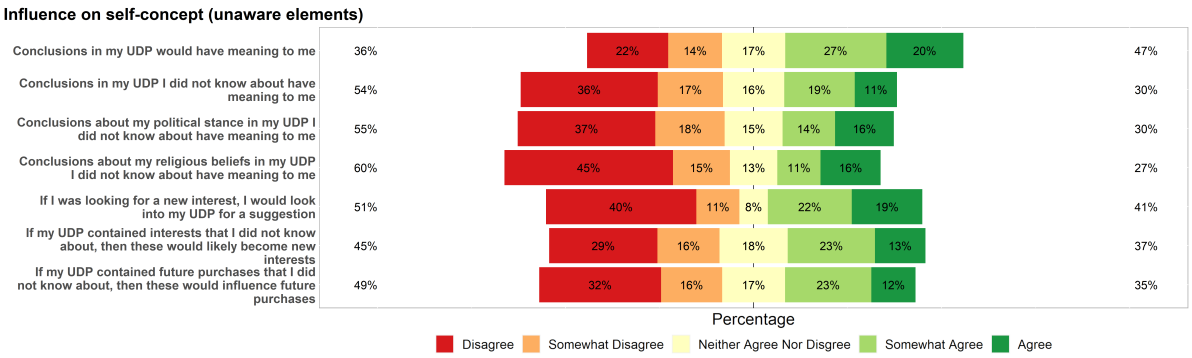


Figure 4.7: Respondents' ratings were largely divided over the question whether UDP conclusions would be meaningful to them and whether unknown identity declarations would carry meaning for them.

- *If my user data profile contains conclusions about my interests (e.g., movies, music, or books that I like) that I do not know about, then these conclusions will likely become new interests of mine.*
- *If my user data profile contains conclusions about what I will likely buy in the future, that I didn't know about, then these conclusions will likely influence what I buy in the future.*

Demographics

54.3% of participants were female, 43.8% male, and 1.9% defined themselves as other. 69% of participants were between 18 and 35 years old. 56.8% of participants had some form of university education, 33.4% had at least a high school diploma. 50.8% of participants were employees, 18.2% were students. Finally, 92.9% of participants listed their current country of residence as the United States.

Appendix Figure 7

Figure 4.7 shows respondents' beliefs on the influence of the UDP on their self-concept. In particular, we wanted to understand whether participants would attribute meaning to identity declarations in their user data profile (UDP) that they were not aware of. Figure 4.7 is shown on the following page.

5 Published Articles Part 3: Facial Analysis AI

Research Article 1: Setting the Stage: Towards Principles for Reasonable Image Inferences (2019)

Research Article 2: What People Think AI Should Infer From Faces (2022)

Research Article 3: AI-competent Individuals and Laypeople Tend to Oppose Facial Analysis AI (2022)

Please note that the published articles are slightly modified mainly to allow for unification of format and reference style. References for each research paper appear in the overall bibliography at the end of the doctoral dissertation. Published versions of the research articles are appended to end of the doctoral dissertation in chapter 7.

5.1 Research Article 1: Setting the Stage: Towards Principles for Reasonable Image Inferences

Authors

Severin Engelmann, Jens Grossklags

Publication Outlet

UMAP'19 Adjunct: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization; June 2019; Pages 301–307; https://doi.org/10.1145/3314183.3323846

Abstract

User modeling has become an indispensable feature of a plethora of different digital services such as search engines, social media or e-commerce. Indeed, decision procedures of online algorithmic systems apply various methods including machine learning (ML) to generate virtual models of billions of human beings based on large amounts of personal and other data. Recently, there has been a call for a "Right to Reasonable Inferences" for Europe's General Data Protection Regulation (GDPR). Here, we explore a conceptualization of reasonable inference in the context of image analytics that refers to the notion of evidence in theoretical reasoning. The main goal of this paper is to start defining principles for reasonable image inferences, in particular, portraits of individuals. Based on an image analytics case study, we use the notions of first- and second-order inferences to determine the reasonableness of predicted concepts. Finally, we highlight three key challenges for the future of this research space: first, we argue for the potential value of hidden quasi-semantics. Second, we indicate that automatic inferences can create a fundamental trade-off between privacy preservation and "model fit" and, third, we end with the question whether human reasoning can serve as a normative benchmark for reasonable automatic inferences.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

5.1.1 Introduction

Recently, user modeling techniques have been used to infer aesthetic (e.g., beauty), mental (e.g., beliefs, intentions), emotional (e.g., happiness, depression), and social (e.g., group affiliation) features about individuals based on their personal data as well as their digital footprints. The possibilities of user modeling techniques go far beyond the mere classification of individuals into types of customers: they create virtual models of individuals at an industrial scale based on personal and other data. This data is commonly associated with implicit mental characteristics and social situational factors often unknown to the corresponding individual. Thereby, many big data companies produce billions of virtual models of people to connect a particular informational resource (e.g., an advertising material) to the individual with the most "appropriate" model.

This signifies what we refer to as a hermeneutic shift: parts of the interpretative potential of the person is realized not by the person itself but by the "quasi-semantic power"[1] of textual extraction, image understanding, emotion and speech analysis, location analysis or even inaction interpretation (among others) [239, 240, 244, 245, 246]. Assigning quasi-semantic values to implicit identity claims stands in contrast to The Enlightenment's core idea that humans have the ability to freely and autonomously assign meaning to what they have experienced. From this perspective, user modeling techniques can create tensions with the autonomy of individuals to form a hermeneutic self-concept. Moreover, the quasi-semantic power of user modeling techniques can lead to consequential discriminatory biases, for example, when credit decisions are based on the collection and analysis of digital footprints unknown to the corresponding individual. The opacity of user modeling processes makes it generally difficult to detect, understand and correct such biases.

Recently, there has been a call for a "Right to Reasonable Inferences" to set legally-binding standards with the purpose to protect individuals against inferences that are privacy-invasive, reputation-damaging, and difficult to verify [301]. Yet, the decisive question is what *reasonable* ought to mean in the context of an automatic inference about a person based on some published media content.

Here, we wish to set the stage for a productive discussion between the computer and social sciences in determining standards for reasonable inferences in image analytics.[2] Based on an image analytics case study using the Clarifai concept prediction prototype[3], we show that inferences about human portraits can be unreasonable when they predict concepts with underlying beliefs that cannot be revised in light of further evidence of the same type. Our claims are based on an empiricist view of reasonableness[4] that considers a knowledge-object's quality of evidence for a particular inference to qualify as reasonable or unreasonable.

We proceed as follows. In Section 5.1.2, we discuss why image analytics result in epistemic

[1] Since humans are the only semantic engines in nature, see, for example, [300].
[2] Specifically, images that depict human beings.
[3] Available at: https://www.clarifai.com/demo.
[4] The terms "reasonableness" and "rationality" are considered synonymous in this work.

and ethical challenges and review related work in Section 5.1.3. In Section 5.1.4, we introduce an empiricist conceptualization of reasonableness that demands that what one is justified in believing is determined exclusively by evidence. We then upload two portraits to the Clarifai web interface image prediction prototype and analyze the reasonableness of the concepts the engine returns (see Section 5.1.7). Finally, in Section 5.1.8, we consider the potential autonomy-enabling value of hidden quasi-semantics and discuss a fundamental trade-off between privacy and model fit.

5.1.2 Background

Social media users engage in both explicit[5] and implicit identity claims. Generally, images are among the most prevalent forms of self-presentation techniques on social media. Given their inherent semantic ambiguity, images are considered implicit identity claims. Implicit identity claims are "given off" in various indirect manners. Typical examples of implicit identity claims are showing one's affiliation to certain individuals, social or institutional groups, or expressing preferences and interests in an indirect manner [302, 303]. Indeed, there is evidence that "showing rather than telling" has become the most common self-presentation strategy on social media platforms [304, 270].

Consequently, marketers value images more than other media content. According to Socialbakers, images posted on Instagram[6] create four times more user engagement than other user content on Facebook[7]. Another reason is that image understanding further closes the gap between organic and commercial media content since objects in an image can be classified as products. Overall, there have been significant efforts made in the advancement of image-understanding technologies to model users based on pictorial identity claims in both academia and industry.[8]

When modeling an individual, image-understanding technologies do not simply draw semantics from the content of images but assign, add, and possibly produce their meaning in the first place. Despite their quasiness, user modeling techniques model features of individuals that are likely inaccessible for the individual herself. Thereby, user modeling techniques presumably attempt to transfer what is radically subjective (and therefore difficult if not impossible to falsify) into the realm of objective evaluation. They, therefore, try to explain something that is essentially first-person in third-person terms.

The majority of contemporary philosophical theories on personal identity support the idea that being free in interpreting one's self is a constitutive element of the conceptual boundaries of personal identity [281, 305, 278, 282]. Importantly, a moral status comprising

[5] For example, when individuals communicate specific self-relevant information in written form, they usually engage in explicit identity claims: "I am 20 years of age and I like reading biographies of great scientists".

[6] Advertising campaigns on Instagram are run via the Facebook advertising platform including the choice of custom audiences and lookalike audiences: see https://business.instagram.com/advertising/.

[7] https://www.socialbakers.com/blog/instagram-engagement

[8] For example, Amazon: https://aws.amazon.com/de/rekognition/, Microsoft: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/, Facebook: https://code.fb.com/ai-research/fair-fifth-anniversary/, Google: https://cloud.google.com/vision/.

moral rights and duties presupposes autonomy over one's self-concept. In other words, it is *because* individuals can evaluate what they are, shape whatever they wish to be on this basis, that they can be made responsible for what they become [283]. Moral accountability would, therefore, be impossible if individuals did not have the freedom and autonomy to form and negotiate such a hermeneutic self-concept.

Furthermore, empirical studies in psychology have demonstrated that individuals have the ability to attribute meaning to their experiences as a processes of hermeneutic identity formation [306, 307, 308]. Studies by [309] show that individuals interact with other individuals strategically in order to verify their self-concept: self-concept negotiation denotes the verification attempt of a person's self-concept through the interaction with other individuals. Whether individuals perceive user modeling outcomes as a means of technologically-mediated self-verification or self-discontinuity remains to be studied. Yet, hiding a person's quasi-semantic self-concept, i.e. disallowing user modeling techniques to partake in a self-verification process, could have some benefits (see Section 5.1.8).

Taken together, an autonomous self-concept emerges when an individual carries out *the psychological work required to attribute meaning to certain experiences*. Image analytics signify a hermeneutic shift because they transform implicit identity claims into explicit declarations of identity. Image analytics are not solely epistemic tools but quasi-semantic engines that potentially interfere with a person's autonomy to freely form a self-concept.

5.1.3 Related Work

With the rise of search engines in the early 2000s, automatizing the attribution of semantics to images returned high accuracy on object identification [310]. In the context of search tasks, object identification proved to be an efficient strategy.[9] In social media's people-based marketing mere object identification does not suffice for advertisement delivery based on implicit identity claims. Today, learning from content and structure of social network sites as well as correlating aspects about natural persons and groups to online content is a fast-growing research field. In the following, we briefly discuss main trends as they pertain to image data analyses.

Popularity prediction of image data: Several projects focus on determining the likelihood that certain image postings will achieve high view counts and high positive approval. Using a variety of machine learning approaches the context of a user and posting is taken into consideration to predict the future attention given to a newly posted image (e.g., [311, 312, 313, 314]).

Self-presentation: Various papers explore how (and under what circumstances) individuals strategically manage their social network accounts to aim for more favorable reception by the intended audience (e.g., [315, 316]). In the context of image data, for example, researchers have

[9] Object inferences can be semantically ambiguous. For example, while distinct colors and shapes can be mapped to mathematical vectors with relative ease, the same is more difficult with objects containing continuous features [249].

begun exploring users' management of multiple accounts on Instagram to present themselves to different audiences in strategically altered ways. On a "Rinsta" (Real Instagram) account, a curated self is presented to a wider audience; whereas on a "Finsta" (Fake Instagram) account, less perfect material is presented to a hand-selected group of individuals for feedback and banter [304]. Interestingly, research has shown that users perceive their carefully styled images on the Finsta accounts to capture their real self more accurately in comparison to their Rinsta accounts with presumably more "genuine" material [304].

Inferring personality traits and user characteristics from image data: Partly triggered by the Gaydar research study [317] in 2009, significant attention has been given by the research community to finding associations between aspects of user profiles, user relationships, and posts, on the one hand, and traits/characteristics of the user or groups of users, on the other hand. In the context of image data, recent research suggests a relationship between personality traits and style aspects of posted pictures (e.g., hue, brightness and saturation); likewise, the content of pictures can be associated with personality characteristics [318, 319, 320]. Previous work also aims to find image characteristics that match specific user groups [321]. Likewise, analyses focus on automatically detecting gender and age from posted image content [322, 323]. Behavioral research has also explored how different personality characteristics (e.g., narcissistic tendencies [324]) impact the perception of image data.

Relationship of mental health and image data: Numerous research projects have focused on uncovering correlations between the usage of social network sites and mental health aspects such as addiction, anxiety, depression or body image (see, for example, a recent review [325]). Similar work can be found that is focused on image data. For example, perusal of attractive pictures of celebrities and peers has been found to be associated with a more negative body image by women [326, 327]. Likewise, uploaded image data can also be revealing of mental health indicators such as related to depression [328]. While there is a plethora of technical research and behavioral studies to understand social network site usage and its impact on users, also in the context of image data, we are unaware of any work that explores principles to develop reasonable standards for image inferences made by automated systems.

5.1.4 First Steps Towards Principles for Reasonable Image Inferences

5.1.5 An empiricist view of reasonable inferences

Fundamentally, there are two types of reasoning: practical and theoretical reasoning also sometimes referred to as instrumental and epistemic reasoning, respectively (see for example [329]). Practical reasoning is concerned with the question "What to do?". Theoretical reasoning asks "What to believe?". Practical and theoretical reasoning are not mutually exclusive. When choosing a reasonable action for a desirable outcome an individual relies on a theoretically reasonable belief. Thus, practical or instrumental reasoning usually follows theoretical reasoning.

In this work, we assume an empiricist view that considers a knowledge-object's quality

PREDICTED CONCEPT	PROBABILITY
woman	0.980
portrait	0.965
facial expression	0.964
fashion	0.930
pretty	0.930
multicultural	0.925
one	0.911
wear	0.908
promotion	0.899
contemporary	0.858
indoors	0.856
friendly	0.847
arrival	0.846
people	0.842
elegant	0.822
intelligence	0.786
dentition	0.742
casual	0.728
business	0.720
charming	0.714

(a) Female portrait

(b) Predicted concepts

Figure 5.1: Concept results using the Clarifai image prediction demo for a female portrait. The engine returns predictions on gender "woman", ethnicity-related features "multicultural", cognitive skills "intelligence", and presumably aesthetic features "pretty", "elegant", "friendly", "charming" (among others). For copyright purposes, we artistically rendered the original picture. Original picture ©https://thispersondoesnotexist.com/.

PREDICTED CONCEPT	PROBABILITY
portrait	0.993
eye	0.986
face	0.974
guy	0.974
man	0.971
fine-looking	0.968
young	0.961
hair	0.938
boy	0.933
people	0.932
blood	0.930
dark	0.916
freckle	0.909
serious	0.909
model	0.909
crazy	0.909
fashion	0.895
funny	0.893
smile	0.890
skin	0.888

(a) Male portrait

(b) Predicted concepts

Figure 5.2: Concept results using the Clarifai image prediction demo for a male portrait. The engine returns predictions on gender "man", age "young"/"boy", mental "crazy"/"funny", and presumably aesthetic features "fine-looking", "serious" (among others). For copyright purposes, we artistically rendered the original picture. Original picture ©Bruce Gilden.

of evidence to decide whether a particular inference qualifies as reasonable or unreasonable. The empiricist view of a reasonable inference considers whether the belief about a proposition is *proportional* to the evidence available. Generally, the empiricist view on being reasonable in the theoretical sense considers the "goodness" or "fitness" of reasons provided that favors the truth of a proposition. While this conceptualization of reasonableness perhaps seems simple or even trivial, empirical research has demonstrated that individuals exhibit many information-processing biases pursuant to this empiricist account of reasonableness [330, 331].[10]

The goal of this work is to start developing principles for *portrait* image inferences that are eligible to be called reasonable. To do this, we need an example output from an image analytics engine. Here, we use the Clarifai web interface image prediction demo, which is based on deep convolutional neural networks (CNNs). We upload two portraits (see Figure 5.1 and Figure 5.2) to this image prediction demo and analyze the reasonableness of the concepts the engine returns. Corresponding to the literature reviewed in Section 5.1.3, we view a single image as a stand-alone knowledge-object whereby a predicted concept (i.e., the predicted outcome) is based only on the content of that single image.

5.1.6 Case study: Reasonableness and correctness of predicted concepts for two portraits

Reasonable and correct inferences

Consider the two images in Figure 5.1 and Figure 5.2. Is the content of these two images eligible to serve as evidence for the inferences made (see "predicted concepts" top right corner on both images)?

Figure 5.1 displays the face of a woman. The first three predicted concepts "woman", "portrait", and "facial expression" cannot be argued against, just like the first five predicted concepts in Figure 5.2. Here, the given beliefs about these propositions are *proportional* to the evidence available and therefore these inferences can be said to be reasonable. All of these features can be reasonably inferred from the evidence given. Note that we do not evaluate the potential discriminatory or unfair *consequences* of specific labels, rather we are first and foremost interested in their epistemic justification. For example, returning the label "gender" may lead to consequential discrimination independent from whether it is a (epistemically) reasonable inference. Additionally, considering our two portraits, the features "woman", "portrait" and "facial expression" (Figure 5.1) and "portrait", "eye", "face", "guy", "man" (Figure 5.2) have been classified correctly.[11] Overall, these inferences are – to a large enough degree – reasonable and correct.

Reasonable inferences with incorrect predictions

Other predicted concepts can in principle be reasonable but seem to have been classified incorrectly for the specific portraits given. In Figure 5.2, for example, the CNNs predict the concept "smile", which is incorrect since the person depicted does not seem to smile. Note

[10] For example, category mistakes, anchoring, representative bias, ignoring the context, framing effects etc.

[11] For Figure 5.2, the predicted concepts "hair", "model", "skin" seem to be reasonable and correct as well.

that this would not be an unreasonable inference since a face can potentially bear a smile. Rather, the accuracy of the training set's classification (i.e., the ground truth) is insufficient in returning an otherwise reasonable inference correctly. In this specific case, the prediction seems to be incorrect but only in relation to an otherwise reasonable assumption made when annotating the training set.

Unreasonable inferences due to non-falsifiability There seem to be inferences that are unreasonable due to their non-falsifiability. For example, both images contain predicted concepts of aesthetic evaluations or judgments. For a judgment to be an aesthetic judgment it necessarily needs to be subjective, making it the exact opposite of an empirical judgment. More generally, judgments on beauty and ugliness are commonly taken to be core examples of aesthetic judgments. In Figure 5.1, an example of an aesthetic judgment is "pretty" and in Figure 5.2 "fine-looking". Other, perhaps more indirect, aesthetic evaluations seem to be "elegant", "friendly", and "charming" (Figure 5.1) as well as "serious" (Figure 5.2). Overall, such aesthetic judgments of taste are unreasonable since they cannot be falsified by additional evidence of the same type. For such inferences, additional image evidence cannot *in principle* verify or falsify, in other words, change the proposition.[12]

Similarly to aesthetic inferences, another class of inferences are unreasonable due to their non-falsifiability. These inferences contain category mistakes because they take a physical or anatomical property to be evidence for a mental feature. In Figure 5.1, the facial proportions of the woman are taken to be evidence for her "intelligence" while the face in Figure 5.2 is taken to be evidence for the person to be "crazy". Portraits seem to be inadequate evidence for a person's mental capabilities or, generally, their mental characteristics. This inference cannot be made more reasonable by providing more portraits of the two people shown in Figure 5.1 and Figure 5.2. In other words, the proposition that the person in Figure 5.2 is actually crazy does not become more likely the more pictures of that person are analyzed. Again, the prediction for such labels can be correct but only in relation to the unreasonable assumptions made when annotating the training set.

5.1.7 Analysis of the Case Study

There is an epistemic difference between descriptively identifying the objects "basketball" and "person" and conclusively inferring "Interest person x = basketball", merely because these objects have been identified. In a similar vein, there is a difference between measuring the physical property "wide space between eyes"and the object "glasses" and inferring some measure of intelligence based on these features. In our case study, we generally judged inferences that could be "directly" read off the portrait as reasonable. Such first-order inferences, as one might want to call them, seem epistemically valid and are henceforth difficult to object morally. They are reasonable independent of the predictive strength of the model.

Unreasonable inferences, on the other hand, seem to be predominantly constructed infer-

[12]There are, however, reasonable physical or anatomical inferences, for example, "freckle" in Figure 5.2.

ences. In our case study, they included claims about the person that could not be observed or accessed through the evidence given. Such second-order inferences presuppose a selection (and naturally a disregard) of specific first-order inferences that – combined – produce a new proposition. Second-order inferences must not necessarily be unreasonable. Consider, for example, the predicted concept "indoors" for the portrait in Figure 5.1. Predicting whether a depicted scenery is indoors or outdoors is a second-order inference because a single object is unlikely to produce a definite conclusion. The difference is that this second-order inference is responsive to additional evidence of the same type resulting in belief revision. Thereby, an inference is unreasonable in the case that novel or additional evidence becomes available that defeats the previous justification to believe in a proposition. In case of better evidence one ought to change the previously held belief in light of this new evidence. For example, another image of this scenery could in principle provide what Pollock refers to as "rebutting evidence" [332]. The new image is the same type or source of evidence. But because it is a reasonable second-order inference it is responsive to belief revision, which in this case is equivalent to the principles of Bayesian inference.

This claim does not hold for unreasonable second-order inferences. Bayesian inference (or belief revision) cannot convert an unreasonable second-order inference into a reasonable inference (e.g., predicted concept "intelligence" in Figure 5.1). Such category mistakes can only be reverted by changing the underlying assumption or by gathering different types of evidence but not by considering more evidence with the same category mistake.

5.1.8 Discussion & Concluding Remarks

In this discussion paper, we applied an empiricist account of reasoning to determine the reasonableness of predicted concepts in the context of an image analytics case study. This is only one of many possible accounts of reasoning each of which comes with specific trade-offs. Arguably, an empiricist account is autonomy-preserving but limited to first-order inferences about individuals. Regardless of the account of reasonableness, an inference may be reasonable and correct but still be rejected by the individual. Here, one could argue that an inference becomes reasonable only when the data subject agrees with its proposition.

The recent call for a "Right to Reasonable Inferences" proposes a "Right to know about Inferences" and a "Right to rectify Inferences" (among others) [301]. However, hiding the quasi-semantic power of user modeling techniques does have its benefits. By revealing the logic involved in making hermeneutic inferences, the system directly recommends these hermeneutics to the user. It remains to be explored how individuals would perceive information on inferences as given in our two image examples. Revealing at least in part the manner and content of user modeling processes and outcomes enables internalization and conformation to the proposed inferences. Perhaps individuals would welcome such a degree of transparency as a mechanism to "offload" the psychological work necessary to attribute meaning to certain life events. Revealing such inferences to the individual means recognizing their quasi-semantic power in shaping who we are and who we can become – we accept that they have their own narrative capacity. Thus, transparency of user modeling inferences could

even exacerbate the polarization effect observed in social media personalization.

Another key challenge is privacy. Image inferences tend to become more reasonable the more personal data is collected and analyzed. This creates a privacy trade-off. The trade-off consists in the observation that a representative model of an individual is possible only at the expense of privacy. For example, ML classifiers must be able to respond to concept drift without "neglecting" the outdated data when learning a model of personal identity [263]. For example, sliding windows of fixed and variable sizes of training data are used to build an updated model [260]. Since both fixed and variable windows are definite in their size, some old data will necessarily be forgotten. What criteria determine which data are to be forgotten and which ones are to be considered in creating an updated representative model of a person? Model fit requires a potentially uninterrupted flow of data possibly resulting in significant privacy challenges [290].

Finally, a key question is whether we should take human reasoning as a benchmark for reasonable automatic inferences. In the empirical literature on human reasoning *..."the ordinary person is claimed to be prone to serious and systematic error in deductive reasoning, in judging probabilities, in correcting his biases, and in many other activities"* [333]. For example, humans make judgments about cognitive capabilities based on physical properties [334, 335]. Following our image analytics case study, we conclude that inferences about individuals' cognitive and mental features are unreasonable since an image does not provide the kind of evidence needed to justify such claims. This also counts for inferences made about individuals' intentions or goals based on image evidence (see [336]).

Overall, it will remain a pressing ethical challenge to define normative standards of reasonableness that automatic image inferences should comply with.

Acknowledgements

We thank the reviewers for their insightful comments that helped to improve our work. The paper is based on research conducted as part of a Volkswagen Foundation planning grant project.

5.2 Research Article 2: What People Think AI Should Infer From Faces

Authors

Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, Jens Grossklags

Publication Outlet

FAccT'22: 2022 ACM Conference on Fairness, Accountability, and Transparency; June 2022; Pages 128–141; https://doi.org/10.1145/3531146.3533080

Abstract

Faces play an indispensable role in human social life. At present, computer vision artificial intelligence (AI) captures and interprets human faces for a variety of digital applications and services. The ambiguity of facial information has recently led to a debate among scholars in different fields about the types of inferences AI should make about people based on their facial looks. AI research often justifies facial AI inference-making by referring to how people form impressions in first-encounter scenarios. Critics raise concerns about bias and discrimination and warn that facial analysis AI resembles an automated version of physiognomy. What has been missing from this debate, however, is an understanding of how "non-experts" in AI ethically evaluate facial AI inference-making. In a two-scenario vignette study with 24 treatment groups, we show that non-experts (N = 3745) reject facial AI inferences such as trustworthiness and likability from portrait images in a low-stake advertising and a high-stake hiring context. In contrast, non-experts agree with facial AI inferences such as skin color or gender in the advertising but not the hiring decision context. For each AI inference, we ask non-experts to justify their evaluation in a written response. Analyzing 29,760 written justifications, we find that non-experts are either "evidentialists" or "pragmatists": they assess the ethical status of a facial AI inference based on whether they think faces warrant sufficient or insufficient evidence for an inference (evidentialist justification) or whether making the inference results in beneficial or detrimental outcomes (pragmatist justification). Non-experts' justifications underscore the normative complexity behind facial AI inference-making. AI inferences with insufficient evidence can be rationalized by considerations of relevance while irrelevant inferences can be justified by reference to sufficient evidence. We argue that participatory approaches contribute valuable insights for the development of ethical AI in an increasingly *visual* data culture.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

5.2.1 Introduction

Human faces and the information they convey are essential in human interaction. When seeing a person for the first time, humans rapidly and automatically make a variety of judgments, such as whether a person looks trustworthy or likable [44, 42, 337, 338]. People's faces can play a significant role in some of society's most important decision-making scenarios: first facial impressions can determine hiring choices [339, 42], election outcomes [43, 340, 341], or jail sentences [342, 343, 344]. Yet, we are often told not to judge a book by its cover, an imperative that it is morally wrong to form beliefs about a person based on insufficient evidence. Indeed, inferring inner character traits based on looks had been foundational for once lauded physiognomic and phrenological practices in organizations and institutions [345, 346, 347, 255, 348].

Today, research in psychology and evolutionary anthropology shows that first facial impressions have an "irresistible" force, but are nonetheless largely inaccurate [349, 44, 350, 337, 351]. This line of research provides ample evidence that there is no relationship between how we look and how trustworthy or intelligent we actually are. Surprisingly, another body of research studies continues to suggest that first facial impressions are accurate or, at least, not completely invalid [45, 352, 353, 354, 355, 356]. Commonly recognizing this latter body of literature, computer vision artificial intelligence (AI) – the computerization of visual perception – has recently developed datasets, algorithms, and models to automate social perception tasks in fields such as affective computing (e.g., [357]) and social robotics [358, 359]. Using computer vision AI, studies have claimed to successfully infer emotion expression and intensity [360, 361], sexual [362, 363] and political orientation [364, 365], as well as a variety of latent traits in personality assessments based on people's faces in images [366, 367, 368, 369, 318, 319, 320, 370, 371]. AI research has established tools for feature extractions from faces (e.g., Face++ [13], EmoVu[14]) as well as for open training datasets (ImageNet[15], First Impression V2[16], PsychoFlickr dataset[17]) and models [372, 373] for facial analysis AI.

Computer vision AI drives software that helps "make sense" of user images on social media for advertising purposes, video interviews in hiring software, or mood detection in car systems. The AI emotion recognition industry alone is said to be worth US37$ billion by 2026 [374]. AI systems play an increasingly important role in the semantic interpretation of our world, and because faces have an indispensable social signaling function, they are taken to be particularly revealing of who we are. But how should AI interpret people's faces? All imagery is semantically ambiguous and computer vision AI inference-making necessarily follows from the semantic annotation of visual data by humans, in most cases, by crowd-sourced platform workers [375, 376, 377]. This complicated ethical question has led to debates between policymakers, researchers in computational and social sciences, and

[13] https://www.faceplusplus.com/
[14] https://www.programmableweb.com/api/emovu
[15] https://www.image-net.org/
[16] http://chalearnlap.cvc.uab.es/dataset/24/description/
[17] https://figshare.com/articles/dataset/zahra_plos_data_zip/6469577

companies that develop or use such AI. A number of research papers, including from the FAccT research community, have pointed out ethical challenges with regard to computer vision AI inferences [378, 345, 379, 380, 346, 237, 381, 382, 383, 255, 348, 347, 256]. However, we believe that such an effort must at least be cognizant of how "ordinary" people, i.e., non-experts in AI, evaluate the normativity of computer vision inferences.

In this work, we follow calls for more empirically-informed AI ethics [299, 41] and investigate what non-experts (N = 3745) think AI should and should not infer from portrait images – images that only show a person's face. Using a two-scenario vignette study with 24 treatment groups, we show that non-experts find AI latent trait inferences (e.g., intelligence) morally impermissible regardless of the decision context for which the inference is used for (advertising & hiring). A majority of subjects evaluates inferences such as gender, skin color, and emotion expression as morally permissible in the low-stake decision context (advertising) but impermissible in the high-stake decision context (hiring). None of our framing effects influenced subjects' evaluations indicating a strong value disposition toward AI facial analysis. We use the transformer-based model RoBERTa [78] to analyze subjects' 29,760 written justifications for each AI inference. We find that subjects raise ethical concerns about all AI inferences in both contexts. When justifying the normativity of an AI inference, subjects use one of two meta-principles: an AI facial inference is permissible when facial information warrants sufficient evidence or when making the inference results in beneficial outcomes. Our analysis illustrates the normative complexity behind facial AI inferences, and provides guidance for forthcoming technology policy debates.

5.2.2 Related Work: The imposition of meaning in a visual data culture

5.2.3 Power dynamics between requesters and data annotators

Recently, several authors have raised ethical questions regarding the creation, management, and application of computer vision datasets. Computer vision companies (also known as "requesters") hire data processing companies, most often located in "less developed" countries, to perform efficient and cost-effective dataset creation, including data annotation. The emergence of a visual data culture – across Facebook's services alone, 2 billion images are shared every day[18] – together with the need for manual, human semantic labeling has led to the establishment of a data annotation industry[19] [377, 376]. Critical data science (broadly speaking) highlights challenges related to accountability and transparency gaps resulting from the near-unbounded power of computer vision AI companies and AI research institutes (i.e, requesters) to determine the interpretative potential of visual content [379, 380, 41, 382, 384, 385].

Studies find that requesters face little pressure to justify data labeling projects when hiring data processing companies for dataset labeling [380, 379, 377, 41]. In a field study on two

[18] *Using Artificial Intelligence to Help Blind People 'See' Facebook*: `https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/`

[19] For a contribution by investigative journalists on the data annotation industry, see: *A.I. Is Learning From Humans. Many Humans.* `https://www.nytimes.com/2019/08/16/technology/ai-humans.html`

data processing companies, Miceli et al. concluded that the work of image annotators is largely guided by the interests of the requester organization [379]. The authors report that this power dynamic does not allow image annotators to voice ethical concerns during the data labeling process. The hierarchical managerial structure at data processing companies restricts the possibility for the deliberative input by annotators [380]. In [379], the authors assert that "the one who is paying has the right to the imposition of meaning". To increase transparency and accountability of dataset creation, researchers have developed proposals to standardize documentation. For example, Gebru et al. suggest that each dataset should have a corresponding datasheet, explaining, among others, the purpose for which the dataset was created, the description of the images (or other data types), procedural aspects such as data cleaning and labeling, as well as the tasks and their unique contexts that the dataset is intended to be used for [384]. Holland et al. propose a "Dataset Nutrition Label" that specifies different modules, including the data origin, dataset variables, and ground truth correlations [386]. These and other standardized documentation practices [e.g., 385] can help AI developers to select more suitable datasets for their model development. However, such documentation practices are currently voluntary and rely entirely on the initiative and implementation of dataset creators.

Faces as sources of meaning and means for classification?

Authors have raised critical questions regarding a second key ethical challenge that is the subject of this work: What kind of inferences should a computer vision AI make about people based on visual data? Moreover, how do we justify what differentiates permissible from impermissible facial inferences when the context application changes? Given the inherently semantic ambiguity of visual data, fixing the large space of interpretive possibilities to a selection of target variables is an act of classification that inevitably demands an ethical justification [345, 387, 388, 347, 347, 383, 382]. This particularly applies to inferences about people based on their facial looks. Human faces are among the most frequently used "objects of interpretation" in computer vision AI. A recent review of nearly 500 prominent computer vision AI datasets found that 205 were "face-based": no other object was represented more often in computer vision datasets than human faces [41]. Social psychologists assert that humans are "obsessed" with faces and that they "cannot help but form impressions based on facial appearances" [350, 389, 44]. On first encounter, faces influence first impressions and shape whether we think someone *appears* trustworthy, intelligent, assertive, or attractive (among other traits) [42, 389, 337]. In many ancient cultures, and still today, there are persistent beliefs that faces are "a window to a person's true nature" [389], the idea that there is a reliable relationship between facial appearance and character[20]. The "irresistible influence" of faces can be consequential: first impressions can determine to whom we speak at a social gathering, whether we perceive a politician to be trustworthy, or whether we judge a job applicant as intelligent [350, 389, 391].

[20] In evolutionary psychology, current research debates whether facial attributes (first impressions) are solely innate, evolutionary adaptive heuristics [44] or whether they also have a learned, cultural dimension [337, 390].

Recently, computer vision AI has purportedly inferred such first facial impressions for a variety of different contexts, for example in social media and for automatic hiring software [366, 367, 392, 393, 394, 318, 319, 320, 373, 368]. In the United States alone, millions of job applicants have participated in automatic hiring procedures that assess, among others, candidates' faces to produce an employability score [383, 256]. Sensitive categories such as gender and race are often treated as "commonsense categories" in computer vision datasets [380, 345, 382, 383]. However, a recent comparison between computer vision datasets presents findings that some racial categories show more variance than others across datasets despite nominally equivalent categorization [387]. Buolamwini and Gebru show that facial analysis AI produces the highest error rate for darker-skinned women and the lowest error rate for lighter-skinned males [53]. Critical perspectives warn that gender and skin color classification by facial analysis AI echoes colonial acts of "reading race onto the body" [395]. Facial analysis AI tends to rely on binary, cis-normative gender classifications [396, 395], thereby neglecting a trans-inclusive view of gender. Emotion recognition and sentiment analysis based on facial expressions have been the subject of multiple AI research projects and a plethora of digital companies – from large corporations to startups – use AI to infer facial emotion expression for social media, hiring, education, health, or security [347]. Other studies present facial analysis AI that is "better" at inferring sexual and political orientation from facial features than people [364, 362]. Others have organized yearly "first impression challenges" – competitions to create benchmark vision models for automatic first impression inferences in job candidate screening[21]. Computer vision AI studies often embrace research studies that underscore the apparent validity of first impressions or that, at least, assert that the invalidity of first impressions is inconclusive [397, 45, 352, 353, 354, 355, 356]. However, there is strong evidence that first facial impressions do not go beyond a "kernel of truth" [349, 350, 389, 44, 390, 337].

The conviction that facial configurations are indicative of a person's character inevitably rests on the pseudoscientific ideas of physiognomy and phrenology. Once celebrated scientific theories, prominent figures in the field of physiognomy such as Caspar Lavatar, Ceseare Lombroso, and Francis Galton developed entire taxonomies of facial configurations with what they believed to be corresponding character interpretations (for a historic account on physiognomy, see [44]). Critical data science research points to several ethical concerns resulting from the AI classification of people based on their facial appearance. Hanley et al. criticize that inferences about people based on visual data necessarily represent only those factors of an inference concept that are visibly discernible [388]. Similarly, Stark & Hoey underscore a "fixation on the visible" in their conceptual analysis on the ethics of emotion recognition AI [347]. Computer vision AI inference-making can be presumptuous when designed to predict aims or intentions of people in images [398]. Such systems are morally objectionable because they treat individuals as objects of categorization [399, 388]. Studying the influential ImageNet dataset, Crawford & Paglen find "highly questionable semiotic assumptions [that] echoe(s) of nineteenth-century phrenology" [345]. Other authors call for a

[21] ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results: `https://hal.archives-ouvertes.fr/hal-01381149`, 2017 Looking at People CVPR/IJCNN Competition: `https://chalearnlap.cvc.uab.cat/challenge/23/description/`

ban on "Physiognomic AI" altogether [256].

Research in fairness, accountability, and transparency has successfully produced different formalizations of fairness metrics and approaches for de-biased datasets. However, when it comes to fair visual data inferences it is the selection of target variables that requires careful ethical consideration. If such ethical evaluations are "subjective" and "inescapably political", then how can we make progress in justifying a line between permissible and impermissible inferences? Contributing to this metaethical challenge, we analyze non-experts' ethical evaluations of specific computer vision AI inferences in a low-stake advertising and a high-stake hiring context. We argue that the input of non-experts (i.e., their moral intuitions) can help us critically advance the debate concerning fair computer vision inferences. We consider a participatory approach to be *at least* complementary to conceptual ethical analyses. For example, much of AI ethics in companies and research institutes is guided by "principlism": efforts of expert groups defining often vague ethical principles for algorithmic systems such as transparency, justice or responsibility [61]. Principlism has recently received criticism (e.g., [59]) arguing that abstract ethical principles too often leave room for interpretation and are therefore particularly susceptible to forms of "ethics washing" [62]. Relying on ethical principles alone critically fails to account for the influence of unique contextual factors on the ethical status of AI inference-making. Moreover, by democratic principle, whenever power hierarchies lead to an accountability vacuum, non-expert "users" should have – minimally – a voice in formulating values for the interpretative potential of visual data, including their own. We see this as one element of a holistic approach to advance computer vision AI ethics. For the purpose of the current study, we developed a factorial vignette study that we describe in more detail in the next section. Experimental vignette studies have been extensively used in different fields (including human computer interaction, psychology, experimental philosophy, business ethics) to elicit participants' explicit ethical judgments in a variety of hypothetical scenarios [106, 107, 108, 33, 295, 111, 296, 297, 298]. Our study follows calls for more survey-based AI computer vision ethics [41] and more experimentally-informed AI ethics in general [299]. For a review on the value of studying the "moral intuitions" of non-experts in ethics and philosophy more generally, see [108].

5.2.4 Methods and Experimental Procedure

Data Collection

3745 subjects (male = 50.7%, female = 48.9%, other = 0.4%) participated in our study. Subjects were recruited via Amazon Mechanical Turk. Only "Turkers" with an approval rating above 95% were selected for the study. We deliberately chose to conduct our study via this platform because Turkers have been indispensable for the labeling of some of the most important datasets in computer vision [400, 401]. Besides the large subject pool required for our study, we were interested to understand how a community involved in the labeling of computer vision datasets would ethically evaluate AI facial inference-making.

Our home institution does not require an ethics approval for questionnaire-based online

studies. When conducting the study and analyzing the data, we followed standard practices for ethical research: presenting detailed study procedures, obtaining consent, not collecting identifiable information or device data, and using a survey service[22] that guaranteed compliance with the European Union's General Data Protection Regulation. The study did not include any deceptive practices. Subjects could drop out of the study at any point. All data were fully anonymized, the privacy of all subjects was maintained at all times during the study. Following recommended principles of ethical crowdsourced research [402], we first ran a pre-study with 120 Turkers to determine the average time it would take to complete the survey and used this reference time to determine a payout above the US minimum wage (*mean*= 8.03 min). In our study (N = 3745), the *mean* was 10.4 min (min = 3.35 min, max = 31.55 min).

Vignette Study

The experiment was a between-subject design; each participant was randomly assigned to one of 24 groups. The 24 groups were composed of three experimentally altered variables: two decision contexts (advertising vs. hiring), six evaluative adjective terms (reasonable, fair, justifiable, acceptable, responsible, appropriate), and the presentation or absence of a dictionary definition of the evaluative adjective term. The use of different evaluative adjective terms with or without a dictionary definition accounted for framing effects and tested the robustness of subjects' conception of a normative AI inference [403, 108, 299].

First, subjects were randomly assigned to one of two hypothetical decision contexts: either a low-stake advertisement scenario (n = 1869; mean per group = 155) or a high-stake hiring scenario (n = 1876; mean per group = 156). In the hypothetical advertisement scenario, participants were told that an advertising company deployed computer vision AI to make a variety of judgments about social media users based on their portrait image. Participants were told that the inferences were used to show users more suitable *product advertisements*. We explicitly referred to product advertisements to avoid associations with political advertisements that could have raised the stakes of the decision context. In the hypothetical hiring scenario, a declared high-stake decision context by other studies on algorithmic perception [404, 405], participants were told that a company used computer vision AI to make a variety of judgments about applicants based on their application photo. Subjects were told that portrait inferences were used, together with other assessment metrics, to determine whether or not a candidate is suitable for a job. These scenarios presented curated, *hypothetical* decision contexts typical in vignette research on moral phenomena [108, 297, 106] and fulfilled one of our study's main purposes: to understand whether non-experts evaluate the same set of AI facial inferences differently across low-stake and high-stake contexts. The vignettes can be found in the Appendix in Figs. 1 and 2.

Second, past research has shown that vignettes can be prone to framing effects and that such effects can indicate weak value dispositions in morally-laden scenarios [108, 406]. In our vignettes, the *evaluative adjective term* that prompted subjects' normative deliberation prior

[22]SoSci Survey: `https://www.soscisurvey.de/`

to the primary rating task could have exerted a framing effect. To control for this potential framing effect, each participant was assigned *one* of six evaluative adjective terms – reasonable, fair, justifiable, acceptable, responsible, or appropriate – when performing the rating task: "Do you agree or disagree that this sort of inference made by a software using artificial intelligence is [evaluative adjective term]?". This increased the external validity of our vignette. Using only the evaluative term "fair" could have biased subjects' ratings and justifications. Some people (and in fact cultures) associate the term "reasonable" more descriptively with logical thinking and deliberation while other cultures associate it more prescriptively, such as being honest and responsible [407]. The same was found for people's intuitions about perceptions of normality (also part descriptive, part prescriptive) [408].

Third, studies in experimental philosophy have used "definition vs. no definition" conditions to understand whether subjects use their own intuitive concept when they evaluate essentially contested concepts (such as: what is a reasonable inference?) [297, 403, 108, 299]. Accordingly, half of subjects were presented with a generic dictionary definition of the evaluative adjective term assigned to them, the other half was not. For example: "What do we mean by fair? Something is fair if it's based on equality without favoritism or discrimination." All definitions were taken from the Cambridge Dictionary and were slightly adjusted for our context (see Appendix Table 1). The "definition vs. no definition" treatment allowed us to further test the robustness of subjects' normative evaluations for specific AI inferences: If non-experts' normative judgments were arbitrary to the extent that they could be manipulated by the presentation of a different evaluative adjective term (fair vs. reasonable, for example) or absence of a generic definition of that term, then this would indicate subjects' concept of a normative AI inference to lack robustness. Subjects would then have a low value disposition toward AI facial analysis inferences (studies in experimental philosophy typically use such and similar framing conditions see, for example, [403, 108, 409, 406]).

Facial inferences

To allow for comparison across contexts, inferences needed to have an acceptable degree of appropriateness for two very different decision contexts: advertising and hiring. To keep the cognitive load of our subjects at an acceptable level, we restricted the number of inferences rated and justified by each subject. We decided to present subjects with a total of eight inferences, first asking them to rate their agreement/disagreement and then to provide a short, written justification for each inference rating. We selected the inference "emotion expression" due to its prevalence in emotion detection AI [347, 374]. Similarly, the two inferences "skin color" and "gender" are common attributes in AI inference-making [53, 396]. Four inferences – "trustworthiness", "assertiveness", "intelligence", "likability" – were selected for their importance in studies on *human* first impression-making [44, 337, 42, 390, 350, 389]. Finally, we wanted to understand how subjects would evaluate a facial accessory. We chose "glasses" instead of piercings or tattoos, for example, because the latter two objects exist in more diverse forms. We constructed an 8-item scale to measure agreement with these eight facial inferences made by an AI on a 7-point Likert scale (1 = "strongly agree" to 7

= "strongly disagree", "can't answer"). We did not present subjects with sample portraits, since the impression they would have formed based on the face in the portrait would have likely influenced their normative judgments [350, 44]. The goal of the study was to explore non-experts' ethical evaluations of facial AI inferences *in principle*.

Classification of subjects' justifications

After rating each inference, subjects were asked to justify their evaluation in a written statement. This allowed us to understand the rationale behind subjects' inference ratings and increased data quality (e.g., understanding the plausibility and validity of evaluations, see, [410, 411]). While there is an entire research field dedicated to studying first impressions (e.g., [350, 389, 44]), we could not identify studies investigating people's *ethical evaluations* of such first impressions. This meant that we could not draw from an existing coding scheme for the classification of the 29,760 written justifications. Therefore, we derived the codes directly from the textual corpus. The manual coding process consisted of two iterative cycles. First, one researcher labelled 500 comments to discover major recurring types of reasoning. Another researcher labelled 250 of these comments with the same intent. The researchers then met to discuss and refine the set of identified "justification labels". In a second coding cycle, we randomly sampled 1,250 comments. Two researchers independently added a justification label to each comment. The intercoder reliability was high (Krippendorff's alpha = 0.953). In case of disagreement between the two coders, the comment was discussed with and reviewed by a third researcher. The final set of justification types consisted of the following: 1. "AI can tell", 2. "AI cannot tell", 3. "Inference relevant for decision", 4. "Inference not relevant for decision", 5. "Inference creates harm", 6. "AI has human biases", and 7. "Incomprehensible responses".

Based on this developed coding scheme, we used the language model RoBERTa [78] to analyze the remaining comments. RoBERTa is a more efficiently trained version of BERT [103], an NLP architecture designed for general-purpose language understanding. This required collecting 100 example comments for each justification type (i.e., code). One researcher collected 100 example comments for each justification type. A second researcher then verified classifications. Disagreement was resolved by a third researcher. We split our labeled dataset in 1,001 training and 250 test samples, and performed over-sampling of the smaller classes to create a balanced training dataset. The final optimized model had an overall accuracy on the test set of 95% and each label's F-1 score was higher than 0.94. For the optimization process, we used a learning rate of 3e-5, a maximum sequence length of 32 tokens, and warm-up initialization. We then predicted the labels of the remaining justifications based on the trained model. For the class overview with F-1 scores, see Appendix Table 7.

Our analysis strategy comprised statistical testing of subjects' inference ratings, an exploratory factor analysis, automated text classifications, and a multivariate analysis of variance with follow-up tests. Given the large number of subjects in our sample, we calculated the effect sizes for all significant ($p<0.01$) test results on subjects' ratings.

5.2.5 Results

The consequentiality of the scenario influences non-experts' ethical evaluations of AI facial inferences

We first compared mean aggregate ratings of all inferences between the advertisement and the hiring scenario. A two-sided Welch two-sample t-test found subjects showed greater preference for the same set of inferences in the advertisement scenario (*mean*=3.85; *SE*=1.06) than in the hiring scenario (*mean*=4.41; *SE*=1.2). The difference was significant (t(3687.3)=-15.30; $P<0.001$; 95% *CI*: (-0.64, -0.49)) and represented a small to medium effect (d=0.50) (Fig. 5.3a).

We then compared mean ratings for each inference in the advertisement and hiring scenarios using a two-sided Welch two-sample t-test with Bonferroni corrections for eight tests (Fig. 5.3b). Subjects rated the inferences gender, emotion expression, wearing glasses, and skin color (e.g., skin color, *mean* AD=2.88, mean HR=4.19; d=0.60; $P<0.001$; 95% *CI*: (-1.44, -1.17)) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. In contrast, the inference ratings for intelligence, trustworthiness, and likability (e.g., likability, *mean* AD=5.04, *mean* HR=5.16; d=0.06; P=0.31; 95% *CI*: (-0.24, -0.006)) did not show a significant difference between the two scenarios. Ratings for the assertiveness inference were significantly different between the two scenarios, but the effect size was negligible (*mean* AD=4.69, *mean* HR=4.89; d=0.10; P=0.01; 95% *CI*: (-0.32, -0.078)).

To summarize, comparing the inference ratings solely based on the grouping variable *context*, the consequentiality of the decision context influenced subjects' ratings: in the hiring context, subjects showed significantly more disagreement with the AI inferences gender, skin color, emotion expression, and glasses than in the advertising context. Cohen's d was particularly large for ratings on gender, skin color, and wearing glasses between the two contexts. This difference did not replicate to ratings for the inferences trustworthiness, intelligence, assertiveness, and likability (Fig. 5.3).

Subjects differentiate between "first-order" and "second-order" inferences

To explore underlying constructs in our set of eight inferences, we conducted an exploratory factor analysis (EFA) (Appendix 6). Parallel analysis, scree plot, and the MAP criterion all suggested two factors. One factor included the inferences gender, skin color, wearing glasses, and emotion expression. To use this group of inferences for further statistical comparison, we termed this construct *first-order inferences*. The other factor included the four latent trait inferences intelligence, trustworthiness, assertiveness, and likability. We termed this construct *second-order inferences*. We used these terms (first-order/second-order) as linguistic categories to reflect the statistical reality of subjects' ratings and less as an initial semantic interpretation of subjects' ethical evaluations. Both sub-scales had high reliability, the overall α was 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences* (Fig. 5.4; see Appendix 6.6 for distribution of EFA factor scores).

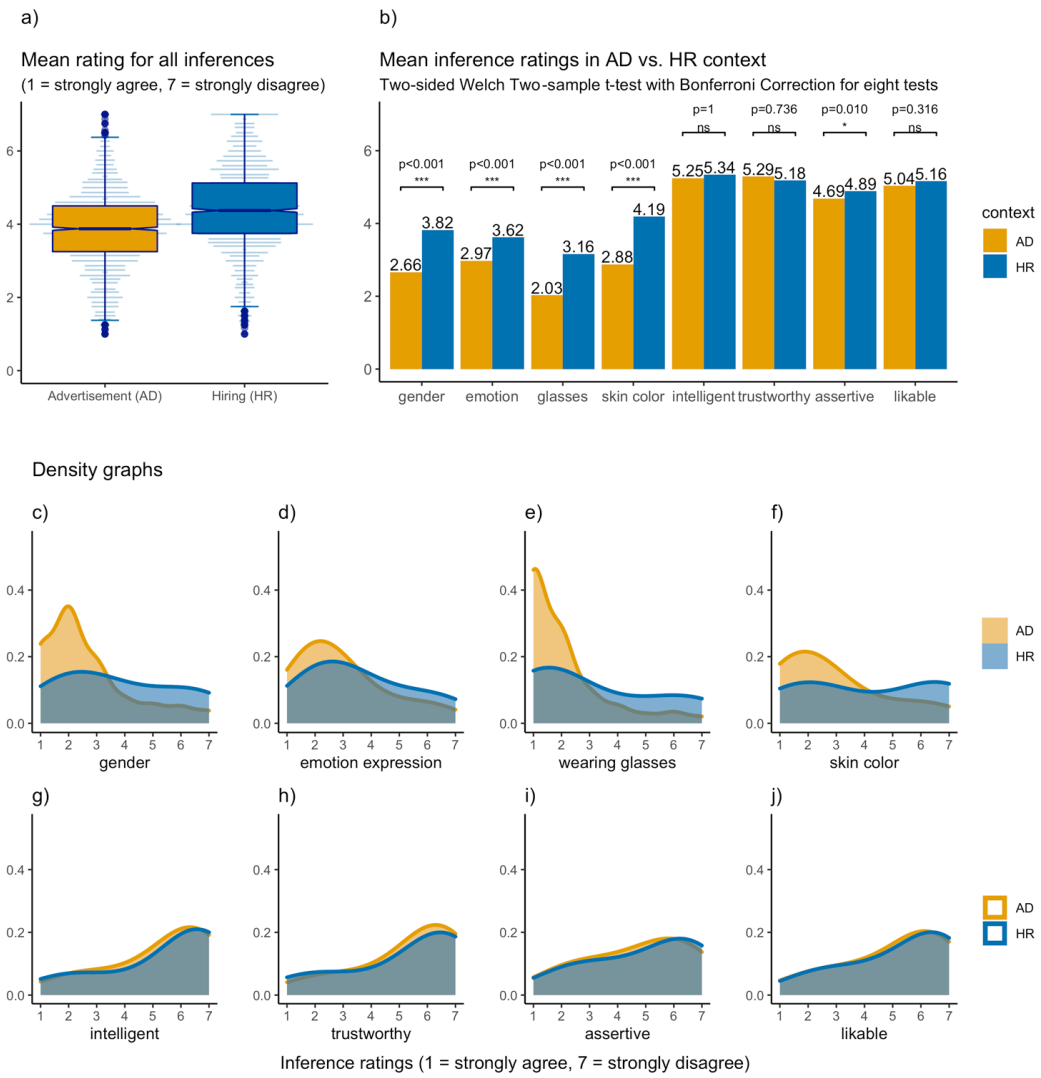


Figure 5.3: (a) Mean aggregate ratings for inferences were more positive in the advertising context than in the hiring context. (b) Participants rated the inferences gender, skin color, emotion expression, and wearing glasses significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent, trustworthy, and likable did not show a significant difference between the two scenarios. Only ratings for the inference assertive were significantly different between the two scenarios, but the effect was negligible (see Appendix 5 for statistics). (c-j) Density plots of inference ratings. 1 = strongly agree; 7 = strongly disagree; 4 = neutral.

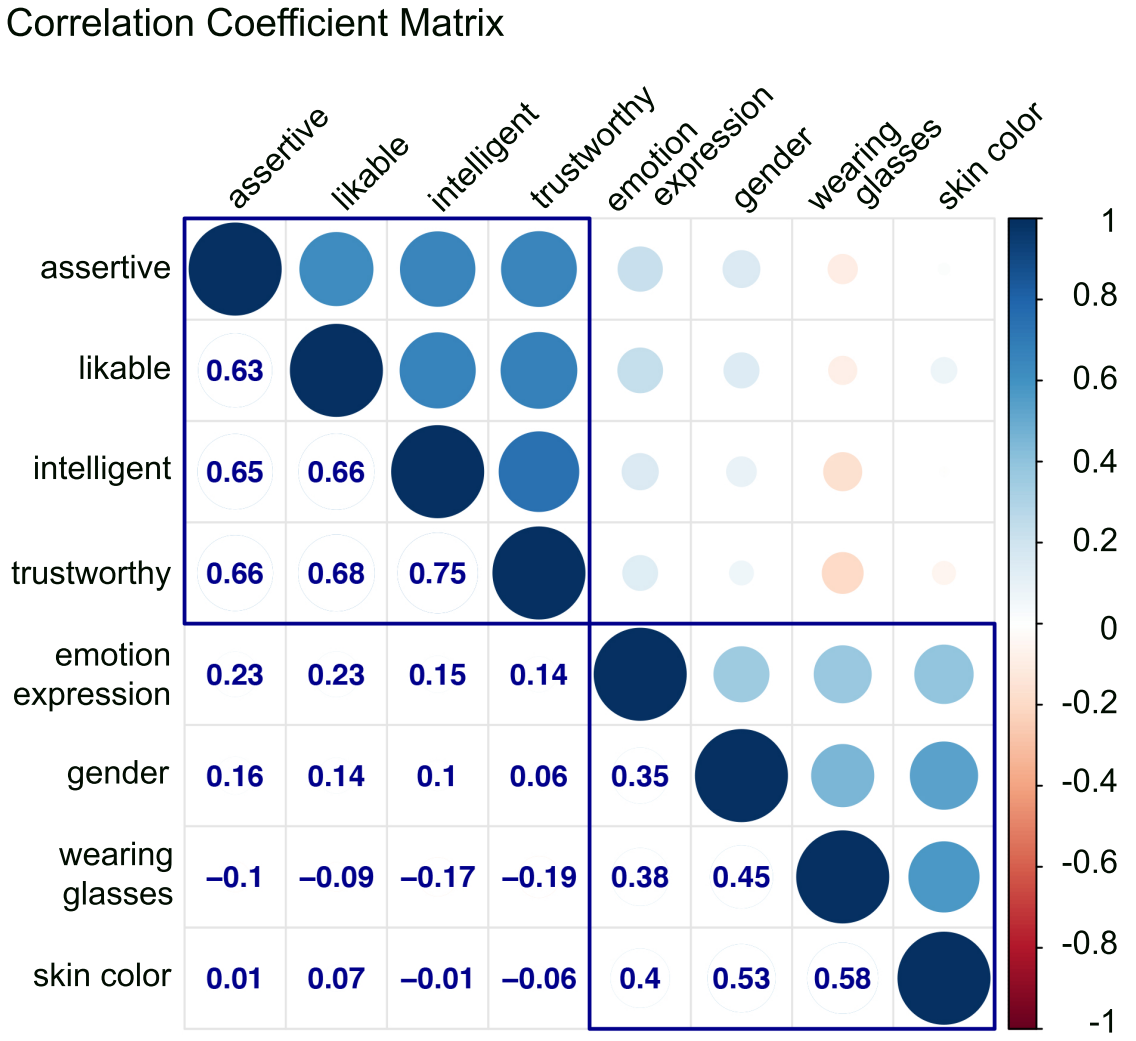


Figure 5.4: Exploratory factor analysis (EFA) resulted in two underlying constructs for subjects' ratings. One factor included the emotion expression, gender, wearing glasses, and skin color inferences. We termed this set of inferences *first-order inferences*. The other factor included the latent trait inferences assertive, likable, intelligent, and trustworthy. We termed this set of inferences *second-order inferences*.

Table 5.1: Follow-up ANOVAs for factor scores from exploratory factor analysis (EFA)

	ANOVA for first-order					ANOVA for second-order				
	SS	df	*F*	Bonferroni	part. η^2	SS	df	*F*	Bonferroni	part. η^2
(Intercept)	7.32	1	22.22	**0.000**	0.006	5.135	1	15.399	**0.001**	0.004
Justifications										
first-order justifications	946.163	6	478.774	**0.000**	**0.455**	46.331	6	23.157	**0.000**	**0.039**
second-order justifications	18.785	6	9.506	**0.000**	**0.016**	844.717	6	422.212	**0.000**	**0.424**
Control Variables										
AI knowledge	14.069	4	10.679	**0.000**	**0.012**	26.058	4	19.537	**0.000**	**0.022**
age	9.939	5	6.035	**0.000**	0.009	5.648	5	3.387	0.052	0.005
gender	0.272	2	0.414	1.000	0.000	2.463	2	3.693	0.275	0.002
occupation	7.834	8	2.973	0.028	0.007	5.720	8	2.144	0.317	0.005
education	1.553	7	0.674	1.000	0.001	2.749	7	1.178	1.000	0.002
Experimental Variables										
context	48.115	1	146.081	**0.000**	**0.041**	2.325	1	6.972	0.092	0.002
terms	6.502	5	3.948	0.016	0.006	5.140	5	3.083	0.097	0.004
definition	0.161	1	0.487	1.000	0.000	0.293	1	0.880	1.000	0.000
Residuals	1135.010	3446				1149.065	3446			

Note:
All Bonferroni-corrected *P*-values are compared to a Bonferroni-corrected α = 0.005 for the computation of two ANOVAs.
Significant *P*-values and partial η^2 values of relevant size are marked in **bold**.
Partial η^2 = 0.01 small effect; partial η^2 = 0.06 medium effect; partial η^2 = 0.14 large effect.

Decision context only influences agreement with first-order inferences

We then extended our analysis to the entire set of treatment conditions. To test significant group differences among the 24 treatment groups on a combination of *first-order* and *second-order* factor scores from the EFA as a dependent variable, we computed a 2 (context: advertisement, hiring) x 6 (evaluative adjective terms) x 2 (definition, no definition) multivariate analysis of variance (MANOVA; Appendix 7). We controlled for main first-order justification theme, main second-order justification theme, AI knowledge, age, gender, occupation and education. Using Pillai's trace, there were significant main effects at an α-level of 0.01 for first-order justification (V=0.50, F(12, 6892)=190.76, P <.001, partial η^2 = 0.249), second-order justification (V=0.45, F(12, 6892)=164.60, P <.001, partial η^2 = 0.223), AI knowledge (V=0.03, F(8, 6892)=13.43, P <.001, partial η^2 = 0.015), and context (V=0.04, F(2, 3445)=73.68, P <.001, partial η^2 = 0.041) (Appendix Table 5).

Finally, univariate analysis with two separate ANOVAs on the *first-order* factor scores and on the *second-order* factor scores from the EFA revealed varying effect structures (Table 5.1; Appendix 7.2). With respect to the experimentally altered variables, *context* was the only significant treatment effect found, but only had an effect on ratings of first-order inferences (F(1, 3446) = 146.08, P <0.001, partial η^2 = 0.04). This finding supported the results from the two-sided Welch two-sample t-test. The experimental treatments *evaluative terms* and *definition vs. no definition* had no significant effect on subjects' ratings. This indicated that the subjects in our sample had a robust concept of a normative facial AI inference. AI knowledge had a small but significant effect on both inference ratings, whereas age had

only a small effect on first-order ratings. Gender, occupation, and education did not have a statistically significant effect on subjects' ratings. Pairwise comparisons confirmed the results by identifying significant group differences between the advertisement and hiring context (Appendix 7.3).

Subjects find AI cannot tell second-order inferences in both contexts. Gender, skin color, and emotion expression produce more complex justifications.

Subjects evaluate the normativity of an AI inference according to two meta-principles

In their written evaluations, subjects considered whether or not an inference was proportional to the evidence (i.e., an epistemic justification) or whether making the inference resulted in positive or negative outcomes (i.e., a pragmatic justification). Representing epistemic principles, we introduced two codes: "AI can tell" and its opposite "AI cannot tell". For example, the comment *"I believe that someone's facial expressions can easily tell if they are assertive. I feel like facial expressions are easy to read and a computer could do that even better."* (assertiveness, HR) was classified as "AI can tell". The comment *"A person's intelligence is internal and based on learning, education, and other experiences. This can't be reflected in someone's looks."* was classified as "AI cannot tell" (intelligence, HR).

With the second meta-principle, subjects considered pragmatic reasons: we identified two contrary justification types "Inference relevant for decision" and "Inference not relevant for decision". The justification *"The reason I believe it is appropriate...is because this will help to select the potential candidate that possesses the assertiveness that could be useful for the job."* was classified as "Inference relevant for decision". The comment "*I don't think assertiveness makes or breaks a job applicant*" was classified as "Inference not relevant for decision" (both assertiveness, HR). A third justification type "Inference creates harm" classified comments stating AI inference-making could be harmful if used as part of the decision-making process (e.g., discrimination due to racism or sexism). For example, the justifications "*Seems like phrenology where intelligence and other traits were determined by the shape of someones head.*" (intelligence, AD) or *"Color should not matter in job hiring. This would be discrimination."* (skin color, HR) were classified as "Inference creates harm". Finally, a justification type that we called "AI has human biases" classified comments stating AI inference-making was flawed by biased human inference-making. Justifications in "AI has human biases" contained epistemic reasons (e.g., *"The software could be implanted with the bias of its creator"*; trustworthy, HR) or pragmatic reasons (e.g., *"The inference is unfair as the AI may be programmed to favor one sex over the other without context."*; gender, HR).

The classification results of subjects' written responses underline the semantic ambiguity of facial portraits: for each inference, we found a corpus of diverse explanations that fell back on epistemic and pragmatic accounts (the two meta-principles). We show the general line of subjects' justifications in Fig.5.5, where we map ratings (agreement/disagreement) to justification types. We complement subjects' general line of justifications with example comments. More example comments can be found in our "code book" in Appendix Table 8.

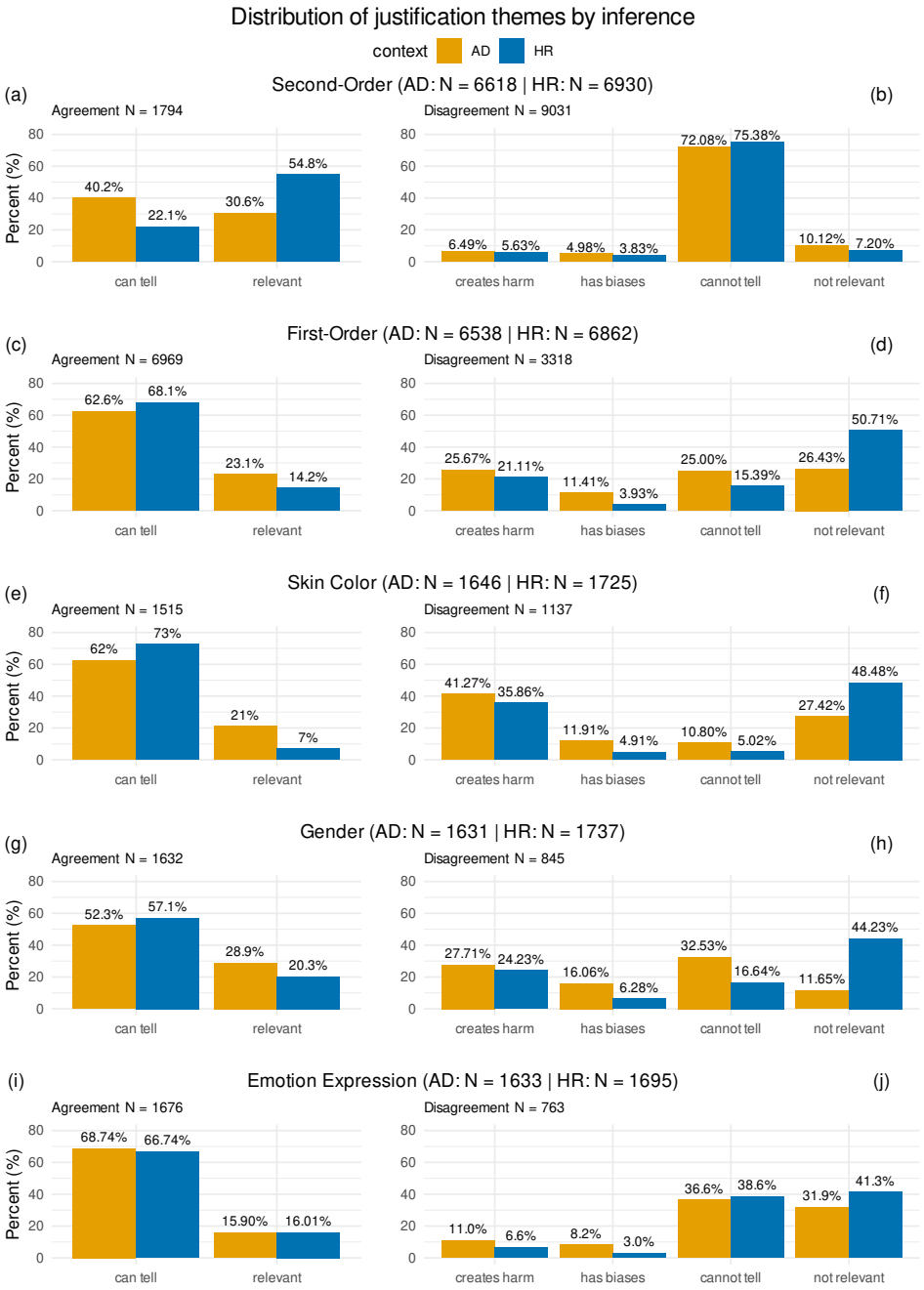


Figure 5.5: Distribution of justification types. Plots a) to o) present the proportions of the justification types used per context. E.g., for first-order ratings, 62.6% of participants in the AD context justified their agreement with an explanation allocated to the justification type "AI can tell" and 50.71% of respondents in the HR context justified their disagreement with an explanation related to the justification type "not relevant". The sum of N for AD and HR for an inference does not amount to the total N because the plot does not include individuals who neither agreed or disagreed. Percentages by context and agreement/disagreement do not sum up to 100%, since the visualization does not include a minority of individuals who provided a counter-intuitive justification based on their score.

Subjects believe AI second-order inferences are invalid inferences regardless of the decision-making context

The majority of subjects believed that faces do not provide sufficient evidence ("AI cannot tell") for inferences intelligence, trustworthiness, likability, and assertiveness (i.e., all second-order inferences) – regardless of the decision context. *"If you're just looking at a person and trying to determine if they're assertive, you're going to score no better than a random guess, I don't care how sophisticated this AI is."* (assertiveness, HR). Some subjects believed second-order inferences to be epistemically valid. *"Assertive people tend to have a set in their jaw, and eyes that is a bit more severe in the angles at the corners than those who are more passive...It might be possible to quantify those angles and measurements to have an AI program analyze the likelihood that they match those of assertive people...If you can come up with a mathematical formula to determine this, then the AI would be capable of measuring it."* (assertiveness, HR). The largest group of subjects agreeing with second-order inferences argued for their *relevance* in the hiring context (54.8%, "Inference relevant for decision"). Here, subjects did not express any epistemic reasoning, but asserted that such inferences were desirable qualities for employers. *"Almost always when you are working, you will work in teams and have to get along with others. You have to be likable to be successful on these teams - I would want the AI to try and assess this as best they could."* (likability, HR).

Subjects believe first-order inferences are epistemically valid, but irrelevant and harmful in hiring

For the inferences emotion expression, wearing glasses, skin color, and gender, subjects' justification profile was more complex (Fig.5.5 e-j). The majority of subjects that agreed with these inferences believed in their epistemic validity in both contexts ("AI can tell"; AD: 62.6%, HR: 68.1%). However, in comparison to second-order inferences, the justification patterns differed between the advertising and hiring context: in the hiring context, considerations of relevance became more important reasons to reject an inference in comparison to the advertising context (Fig.5.5 c). The majority of subjects agreeing with skin color and gender in both contexts believed an "AI can tell" such inferences from facial information (Fig.5.5 e-h): *"Photos reveal this pretty easily assuming the photo is reasonably high rez. I would probably trust a computer to get this right more than some people."* (skin color, HR) or *"This is something that we, as humans can perceive with our sights, so an AI is definitely capable of inferring this.*" (gender, AD). However, subjects that believed "AI can tell" skin color and gender still raised concerns in their written responses even when agreeing with these inferences. For example, subjects noted that accurately inferring skin color may be constrained by photo quality and lighting and may not be an indication of race or ethnicity as the following two comments illustrate: *"I believe a properly calibrated AI could estimate a person's skin color, but lighting, photo quality etc., would have to be accounted for. Also, skin color doesn't necessarily inform us about race."* (skin color, HR). *"Mixed feelings about this one – although skin color is something that can be visually seen in a photo, there is lots of room for error here depending on lighting in photo. Also, whether it's morally right is a whole different subject."* (skin color, AD). Likewise, for gender, subjects pointed to

classification problems of non-binary gender identities: *"For the most part, male/female is an easy question, but there are many people that defy these binary categories that would be excluded."* (gender, HR).

Among the subjects rejecting skin color and gender in hiring, the most common justifications were "Inference not relevant for decision" (skin color: 48.48%; gender: 44.23%) and "Inference creates harm" (skin color: 35.86%; gender: 24.23%). With regard to skin color, most comments stated that skin color does not matter in hiring, while a few added that the inference was justifiable if it resulted in a more diverse workplace: *"This does not matter unless this information is being used to ensure a diverse workplace."* (skin color, HR). Subjects generally agreed that gender does not matter in hiring, however, some subjects asserted that some jobs may be more suitable for certain genders: *"Gender has nothing to do with how capable a person is to do a job unless the job itself requires a specific gender (which is very rare)."* (gender, HR). In contrast, subjects believed that both skin color (21%) and gender (28.9%) are a relevant AI inference in advertising: *"People with different skin colors need different products, and tend to shop for different styles, colors, and patterns."* (skin color, AD) or *"I think this is a 50/50 subject, but I believe personally that this is fair...Perhaps men wouldn't like to see advertisements for bras which would be avoided with this scan."* (gender, AD).

A majority of subjects believe emotion expression indicates emotion sensation

For emotion expression (Fig.5.5 i-j), subjects' agreement or disagreement mainly depended on whether or not they believed facial expressions to be a valid indicator for emotion sensation. Comments classified as "AI can tell" (agreement, AD: 68.74%, HR: 66.74%) claimed internal emotional states could be expressed via the face: *"It is reasonable to judge emotions by looking at a person's face, humans do it all the time. Though some faces can be more expressive than others."* (emotion expression, HR). Given that many Turkers have engaged in portrait image labelling tasks, we also found comments that highlighted the possibility of AI emotion expression inference based on previously conducted labelling tasks: *"A person's emotion can be seen pretty well by looking at a picture as I have done surveys in the past deciding emotion through facial expressions"* (emotion expression, AD). Comments classified as "AI cannot tell" (disagreement, HR: 38.6%, AD: 36.6%) stated the opposite. *"An emotion could be expressed, but the person may not actually be expressing it. In other words, the emotion viewed externally could be one of joy, but, inside the actual person, they may have a different emotion from what is outwardly being expressed."* (emotion expression, HR). The difficult relationship between emotion expression and emotion inference was also evident in comments with the justification types "Inference relevant for decision" (agreement, AD: 15.9%, HR: 16.01%) and "Inference not relevant for decision" (disagreement, AD: 31.9%, HR: 41.3%). To give one example, in comments classified as "Inference relevant for decision" in hiring, subjects claimed that employers may seek employees that need to be friendly, particularly in jobs involving customer interaction: *"Depending on the job emotional expressiveness may be a requirement, you don't want a person in a customer service position who's monotonous and robotic."* (emotion expression, HR).

5.2.6 Key observations & final discussion

The vast abundance of digital imagery together with recent advances in computer vision analysis have raised concerns about the kinds of conclusions AI should make about people based on their face. How do we design computer vision AI in such a way that it will incorporate those preferences and values that are ethically desirable? We explored non-experts' normative preferences of AI portrait inferences in a two-scenario vignette study with 24 treatment groups. One MANOVA and two ANOVAs found that none of our framing effects influenced subjects' ratings, indicating that subjects have a robust, intuitive concept of a normative AI inference for both contexts. Future studies need to further explore how strong this normative concept is in light of other trade-offs such as cost-efficiency, narratives of bias-free technology, or success of the decision outcome, for example.

Conducting an exploratory factor analysis on subjects' evaluations of eight AI facial inferences, two inference categories emerge: we term one category of inferences first-order inferences and the other second-order inferences. Factor loadings of emotion expression as a first-order inference together with subjects' justifications suggest that a majority of the subjects in our sample subscribe to the so-called "Basic View" of emotions [412], which proposes that facial expressions (or "facial action units") are reliable indicators of emotion. Note that this perspective has recently been challenged by emotion researchers arguing that contextual and social factors lead to variability in facial emotion expression that make such inferences unreliable and unspecific [31, 347]. Nonetheless, subjects are aware of the volatility of AI emotion inference from facial expression. They assert that emotion expression as social signaling can be different from the internal phenomenological experience.

Finally, independent of the decision context, subjects believe AI should not draw inferences common in human first facial impression-making due to their epistemic invalidity, i.e., intelligence, likability, assertiveness, and trustworthiness [350, 389, 44]. Subjects raised concerns about all AI inferences in both contexts, even for the – perhaps intuitively – non-problematic "glasses" inference in the low-stake advertising context (Appendix Fig. 7). This leads us to assume that other facial AI inferences, such as beauty, sexual orientation, or political stance, that all have been inferred from faces using AI will likely draw their own justification profiles.

Our analysis highlights the normative complexity behind facial AI inferences. We find that some subjects use a *pragmatic* rationalization of AI facial inferences when they believe that an AI inference is relevant for (i.e., has a supposedly positive effect on) a decision's outcome. However, why should the normativity of a *vision*-based inference be evaluated by criteria other than evidence? The decision context does not have any bearing on the relationship between evidence and inference and therefore should not lead to a different normative evaluation. Thus, our results show that epistemically invalid AI vision inferences can be rationalized by considerations of relevance. The fact that AI research organizations, academic and commercial, commission data annotation companies to label visual data relevant for a specific application purpose necessarily creates a conflicting negotiation between epistemic and pragmatic considerations. Taken together, over-reliance on AI capabilities, narratives

of bias-free technological decision-making, and beliefs in the relevance of an inference for the decision context may form a line of reasoning that supports justification of epistemically invalid AI inference-making. The ongoing publication of research studies that purportedly find a significant correlation between second-order inferences and facial information produces a quasi-epistemic legitimization of first-impression AI. Our study provides evidence that a vast majority of non-expert subjects do not form a justification of AI inference-making along these lines of reasoning.

Finally, how would experts differ in their justification of AI inference-making in comparison to non-experts? Indeed, critical data scientists argue that facial inferences are not reasonable because of their lack of scientific validity (evidentialists) [374, 255], while some AI experts deploying computer vision AI point to positive outcomes in terms of efficiency, cost-reduction, and flexibility that AI inference-making will facilitate [413, 414, 415, 416, 417]. Future studies will need to provide evidence for a unique ethical justification profile of AI vision inferences among AI expert groups. Other future studies should explore to what extent cultural factors play a role in evaluating the normativity of AI inferences based on visual data. We also believe it would be valuable to understand whether subjects evaluate AI video analysis inferences differently than AI image inferences. In fact, AI video analysis interprets visual content at the level of individual frames (i.e., decomposed as a collection of single images) [418].

We hope that the present study underlines the importance of including non-experts in the process of arguing for and against ethically permissible and non-permissible computer vision inferences. We expect norms regarding AI inference-making to shift over time. Allowing non-experts to engage in the formulation of goals and values for AI helps identify such shifts in sociocultural norms. Our study lays an important foundation for determining what types of inferences machines should and should not make about one of the most significant characteristics of us and our place in the social world: our faces.

Acknowledgements

We thank the reviewers for their insightful comments that improved the paper. For their valuable feedback we thank the participants of the 2021 CEPE/International Association of Computing and Philosophy Conference, the participants of the 2021 Ethics and Technology Lecture Series of the Munich Center for Technology in Society, and the participants of the Venice 2019 Metaethics of AI & Self-learning Robots Workshop.

Funding & Support

This research was conducted with the help of a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

5.2.7 Appendix

Vignette scenarios

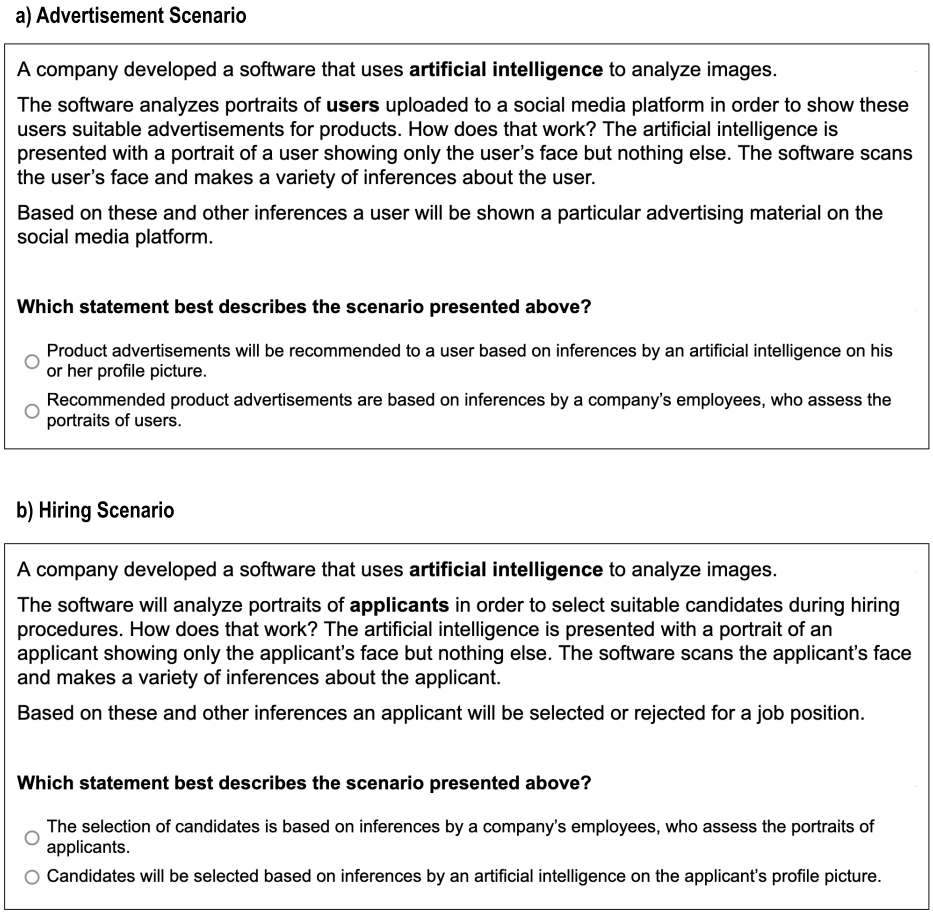


Figure 5.6: Vignette description of the hypothetical advertising scenario a) and hiring scenario b).

Primary task

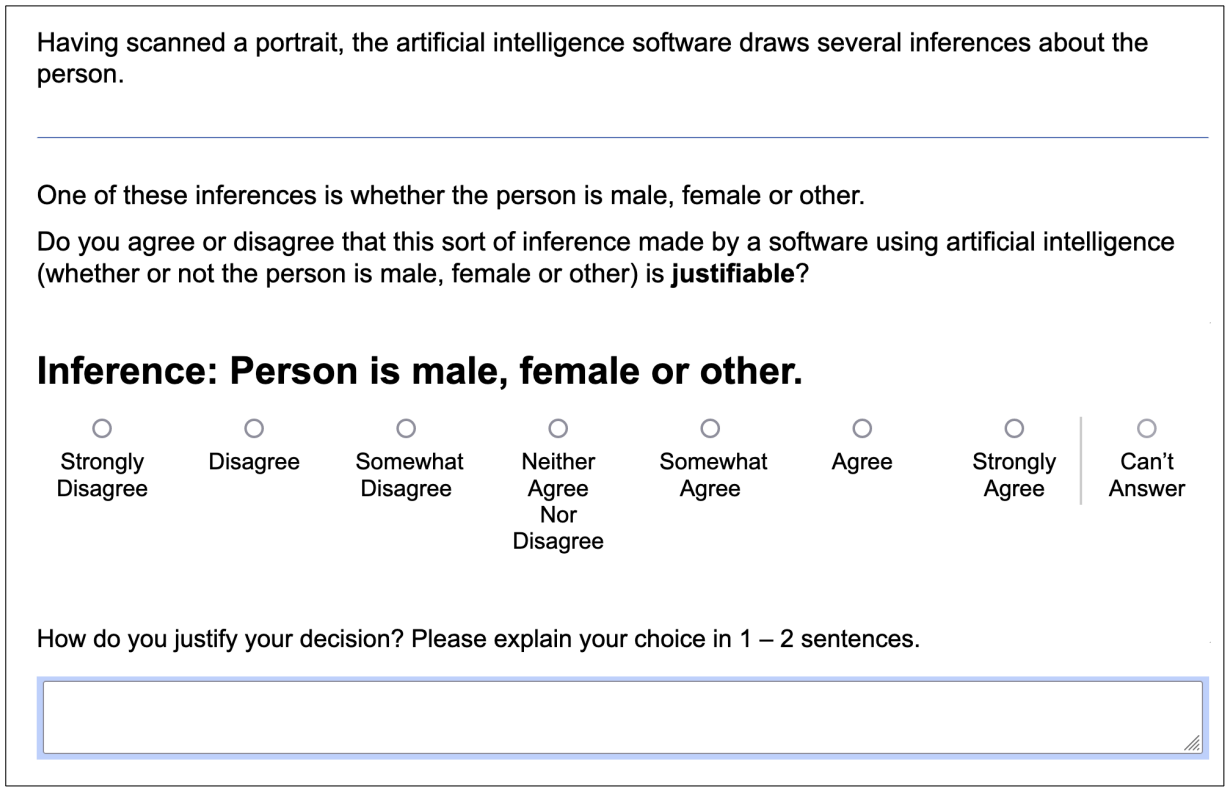


Figure 5.7: Example interface of the primary rating task and the prompt to provide a written response. Example does not show treatment with the presentation of a definition of the evaluative term.

Generic definitions of evaluative terms

Table 5.2: Generic definitions of the six evaluative adjectives presented to half of the participants. All definitions were based on the Cambridge Dictionary, some formulations were slightly adapted to fit our context.

inference	definition
reasonable	What do we mean by **reasonable**? Something is reasonable if it's based on good sense and/or in accordance with reason.
fair	What do we mean by **fair**? Something is fair if it's based on equality without favoritism or discrimination.
justifiable	What do we mean by **justifiable**? Something is justifiable if it can be marked by a good or legitimate reason.
responsible	What do we mean by **responsible**? Something is responsible if it can answer for its conduct and obligations.
appropriate	What do we mean by **appropriate**? Something is appropriate if it's suitable or compatible in the circumstances.
acceptable	What do we mean by **acceptable**? Something is acceptable if it can be agreed on and is worthy of being accepted.

Data Cleaning

The data was cleaned based on the criteria presented in Table 5.3, which gives an overview on the measures taken and a count of identified cases per measure. The SoSci Survey online survey tool provides a relative speed index (RSI) that identifies fast responding participants. This index indicates how much faster a participant has completed the experiment than the typical participant (median). As recommended by SoSci, all respondents with an RSI >= 2 (n = 418) are removed. All samples with duration time between 2 minutes and 4 minutes, cases that rated all inferences with the same rating, and cases with a RSI value above 1.75 were manually checked. Cases identified as problematical were discussed with a second researcher and removed in case of agreement.

Two-sided Welch two-sample t-test

Participants rated the inferences gender (*mean* AD=2.66, *mean* HR=3.82; t(3513.1)=-18.536; $P<0.001$; 95% *CI*: (-1.28, -1.04); d=0.62), skin color (*mean* AD=2.88, mean HR=4.19; t(3513.1)=-18.536; $P<0.001$; 95% *CI*: (-1.44, -1.17); d=0.61), emotion expression (*mean* AD=2.97, *mean* HR=3.62; t(3654.7)=-11.079; $P<0.001$; 95% *CI*: (-0.75, -0.52); d=0.36), and wearing glasses (*mean* AD=2.03, *mean* HR=3.16; t(3147.2)=-18.082; $P<0.001$; 95% *CI*: (-1.26, -1.01); d=0.59) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario.

Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent (*mean* AD=5.25, *mean* HR=5.34; t(3662.2)=-1.425; P=1; 95% *CI*: (-0.21, 0.03); d=0.05), trustworthy (*mean* AD=5.29, *mean* HR=5.18; t(3637.5) = 1.685; P=0.74; 95% *CI*: (-0.02, 0.23); d=0.06), and likable (*mean* AD=5.04, *mean*

Table 5.3: Summary of measures to clean data and number of removed cases

description	removed cases	N
Original N		4752
Time_RSI $>$ 2	418	4334
$<$ 18 years old	1	4333
Attention Check AD	245	4088
Attention Check HR	208	3880
Duration $<$ 120	0	3880
Duration $>$ 120 & $<$ 240	9	3871
Straightliners	52	3819
TIME_RSI $>$ 1.75 & $<$ 2	67	3752
Double Turkers	4	3748
Nonsense Samples	3	3745

HR=5.16; t(3695.7)=-2.059; *P*=0.32; 95% *CI*: (-0.24, -0.006); *d*=0.06) did not show a significant difference between the two scenarios. Only ratings for the inference assertive (*mean* AD=4.69, *mean* HR=4.89; t(3668.3) = -3.219; *P*=0.01; 95% *CI*: (-0.32, -0.078); *d*=0.11) were significantly different between the two scenarios, but the effect was negligible.

Exploratory Factor Analysis (EFA)

Prior to the computation of the exploratory factor analysis (EFA), several assumptions were tested.

Assumptions

Missing Data for Inference Ratings. Missing values appeared to be random and were less than 2% per variable (max. n=71 for the variable *assertive*, accounting for 1.9%; min n=31 for the variable *wearing glasses*, accounting for 0.83%). For EFA, all samples with missing values for the inference ratings were removed (in total 208). The sample size was reduced to 3537.

Normality and Linearity. Table 5.4 lists statistics for each of the dependent inference variables, including skewness and kurtosis. The deviations from normal skewness and kurtosis are within an acceptable range. Additionally, given the large sample size, the impact of departures from normal skewness and kurtosis is negligible.

Table 5.4: Statistics for each dependent variable

	mean	sd	median	trimmed	skew	kurtosis	se
gender	3.26	1.96	3.00	3.07	0.68	-0.80	0.03
emotion expression	3.30	1.80	3.00	3.16	0.67	-0.64	0.03
wearing glasses	2.59	2.00	2.00	2.26	1.13	-0.12	0.03
skin color	3.53	2.25	3.00	3.41	0.46	-1.36	0.04
intelligent	5.32	1.92	6.00	5.58	-0.95	-0.46	0.03
trustworthy	5.25	1.93	6.00	5.52	-0.95	-0.44	0.03
assertive	4.80	1.88	5.00	4.94	-0.46	-1.06	0.03
likable	5.12	1.85	6.00	5.33	-0.73	-0.72	0.03

Absence of Multicollinearity and Singularity. None of the correlation coefficients displayed in Fig. 2 of the main article are greater than .8. This suggested there is no multicollinearity or singularity. Additionally, the determinant of the R-matrix was 0.031 and greater than the heuristic of 0.00001. [69, p. 771]

Factorability of the Correlation Matrix. The correlation coefficient matrix in Fig. 2 of the main article displayed several correlations above .3. An alternative measure is the Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy [419]. A factor analysis is said to yield reliable and distinct factors, if values are close to 1, which suggests that correlation patterns are relatively compact [69, p. 769]. We used the KMO criteria based on [420]. The KMO values for all inference ratings were above .71 and fell within the range of middling values. The overall MSA value was .82, falling in the range of meritorious values [421, 419].

Number of Factors

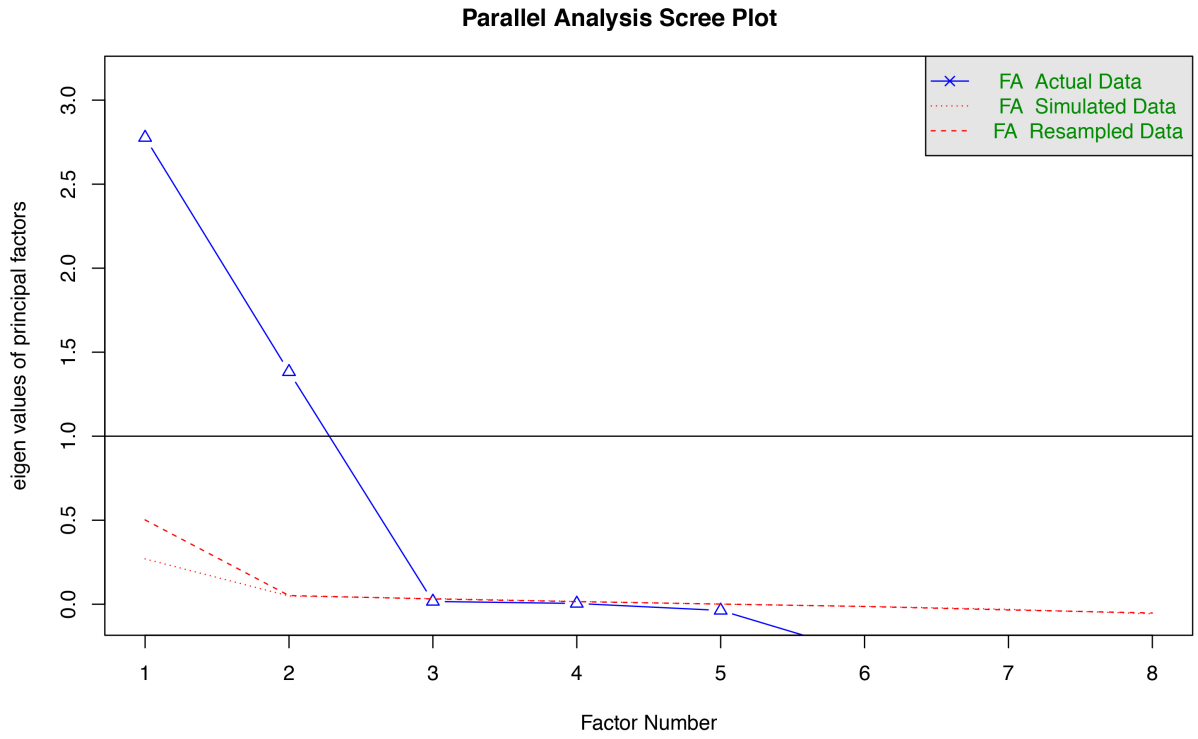


Figure 5.8: Graphical analysis for the number of factors using parallel analysis scree plot.

Given the result from the parallel analysis and scree plot in Fig. 5.8 and other criteria such as the Velicer's MAP test, Very Simple Structure test of complexity 1, and Kaiser's criterion, first a two-factor solution was computed and compared to the results of a three-factor solution and a four-factor solution.

Test Specifications

It was reasonable to assume that the constructs underlying the measured dependent variables correlated, because we measured the agreement to inferences made from the facial region. Therefore, we first applied oblimin as oblique rotation and estimated factor scores using tenBerge for preserving correlations. Supporting this decision, [422, 69] points out that in practice there are many reasons to believe that orthogonal rotation is not appropriate for data involving people, because any construct of psychological nature is correlated in some way with another psychological construct. However, for two factors, oblique rotation resulted in two factors with no correlation. This indicates that the two factors were independent. For correlations of factors below 0.32, [423] suggest orthogonal rotation. Therefore, we applied varimax for orthogonal rotation. Minimum residual (minres) was retained as factoring

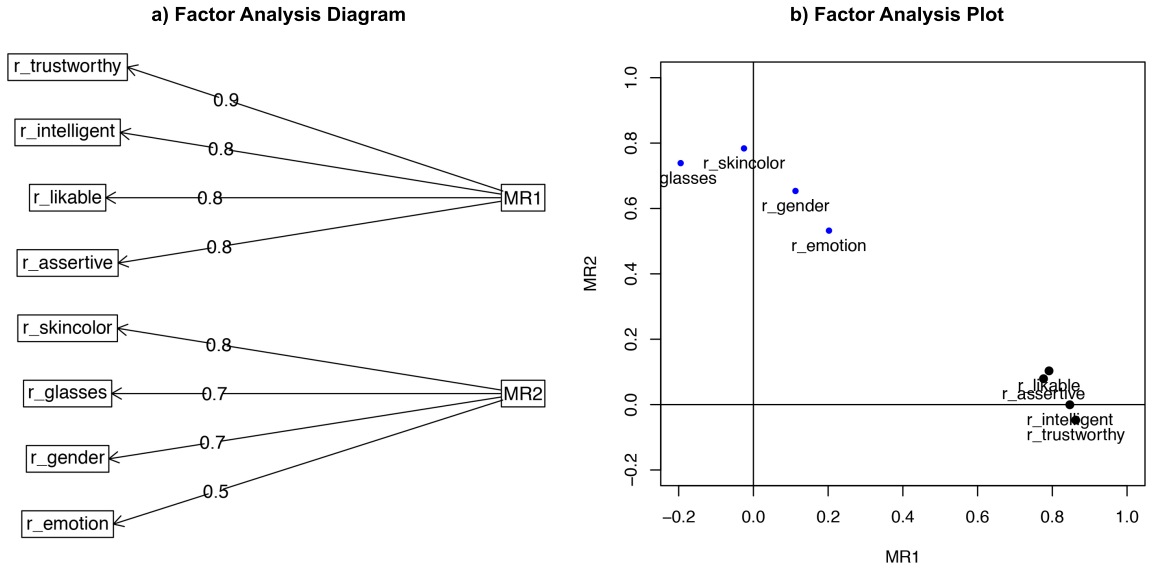


Figure 5.9: Summary of two-factor solution with factor diagram and factor plots.

method, because multivariate normality does not have to be assumed [424]. Factor scores were estimated using regression. To compute the exploratory factor analysis, the R psych package and the GPArotation package were used.

Factor analysis model with 2 factors

Fig. 5.9 a) displays the structure of the factor analysis with two factors and indicates the rounded loadings. MR1 represents the first factor labeled *second-order inferences* and MR2 the second factor labeled *first-order inferences*. Fig. 5.9 b) is a graphical representation of the item's grouping based on their loadings on both of the factors.

There were no residuals > 0.05. The root-mean-square residual was 0.014. The residuals appeared to be approximately normally distributed. Regarding the factor scores, no outliers were identified.

We validated the results by randomly splitting the data in half and running the factor analysis on both subsets. This procedure was repeated three times. For each validation procedure, both factor analyses on the two subsets of the data set resulted in the variables having the same patterns of the factor loadings as with the complete sample. Additionally, the communalities were similar. This validated the factor solution previously obtained on the full dataset.

Both sub-scales had high reliability, the overall α is 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences*.

Table 5.5 displays all solutions with two, three and four factors.

Table 5.5: Overview of Exploratory Factor Analysis Solutions with 2, 3 and 4 Factors.

	Two Factors		Three Factors			Four Factors			
	MR1	MR2	MR1	MR2	MR3	MR1	MR2	MR3	MR4
gender	0.11	**0.65**	0.14	**0.65**	0.01	0.07	**0.66**	-0.01	0.09
emotion expression	0.20	**0.53**	0.08	0.09	**0.62**	0.01	-0.00	**1.00**	-0.00
wearing glasses	-0.19	**0.74**	-0.21	**0.60**	0.17	-0.19	**0.67**	0.07	0.01
skin color	-0.03	**0.78**	0.01	**0.83**	-0.03	0.06	**0.82**	-0.01	-0.05
intelligent	**0.85**	-0.00	**0.87**	0.05	-0.08	**0.86**	0.01	-0.02	0.00
trustworthy	**0.86**	-0.05	**0.87**	-0.04	-0.03	**0.87**	-0.05	0.00	-0.00
assertive	**0.78**	0.08	**0.75**	-0.04	0.14	0.01	-0.00	-0.00	**0.99**
likable	**0.79**	0.10	**0.77**	0.03	0.08	**0.73**	0.06	0.05	0.06
eigenvalues	2.78	1.89	2.73	1.48	0.45	2.07	1.57	1.00	1.00
proportion variance	0.35	0.24	0.34	0.17	0.06	0.26	0.20	0.13	0.13
cumulative variance	0.35	0.58	0.34	0.53	0.58	0.26	0.46	0.58	0.71
α	0.89	0.77	0.89	–	0.76	0.87	0.76	–	–

Factor analysis for 3 and 4 factor solutions

The factor analyses with three and four factors resulted in one and two factors with only one indicator variable respectively (see Table 5.5). This is opposed to the general idea of a factor analysis identifying latent constructs by forming factors out of a combination of at least two variables [425]. Additionally, for the three-factor solution, the cumulative variance was equal to the cumulative variance for a two-factor solution. The third factor had an eigenvalue of $<$ 1. The composition of the three factors was not robust when computing the factor analysis on randomly sampled subsets of the complete data. While the cumulative variance explained by a factor analysis for four factors was the greatest among all tested factor analysis models, this solution was also not robust. Running the factor analysis on two randomly sampled subsets resulted in different patterns of the loadings on the factors. Altering the random sampling produced different patterns of loadings once again.

Although the fit based upon off diagonal values equaled 1 in each of the models, the solutions with three and four factors were neither appropriate in terms of variables per factor nor robust across subsets of the data. Hence, exploratory factor analysis of the eight items measured in this study revealed that two factors were sufficient to explain the underlying structure of common inferences from faces.

Distribution of EFA factor scores and original ratings

The global means for all variables that load on the first factor and all variables that load on the second factor are highlighted by the horizontal lines in Fig. 5.10 a) and b). The bold lines in panels a) and b) indicate the means for the individual groups. By using the factor scores as dependent variables for further analysis, the interpretation of the dependent variables

depicted in panels c) and d) changes compared to the original inference ratings. A factor score of approximately 0 indicates that a participant's mean rating of all variables that load on this factor is close to the global mean of these variables (horizontal lines in panels a) and b)). A negative factor score indicates this subject gave lower than average ratings. A factor score close to 1 indicates that the subject's ratings for the variables loading on this specific factor are about one standard deviation above the average rating.

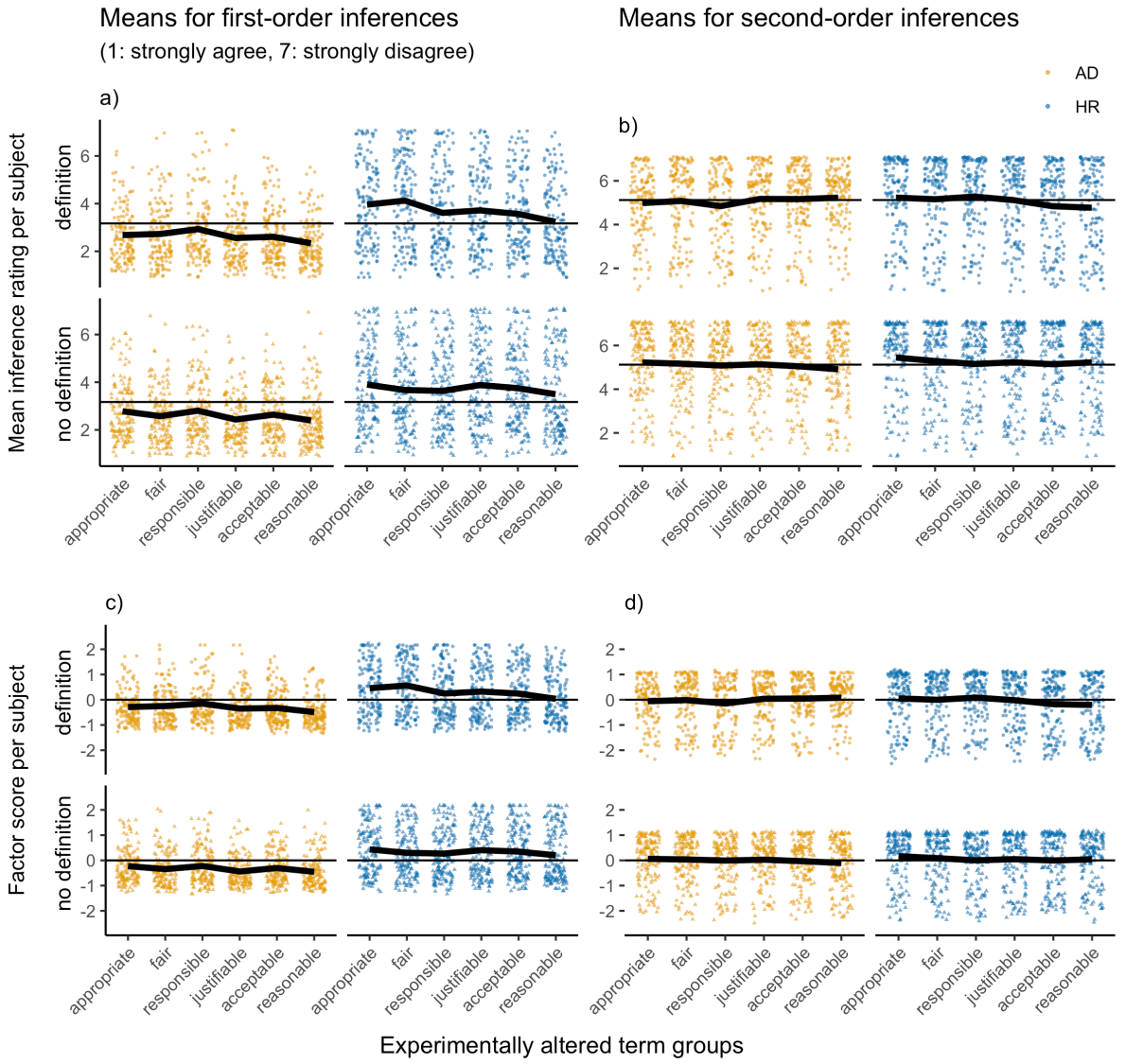


Figure 5.10: Distribution of participants' ratings and distribution of the factor scores extracted from the exploratory factor analysis.

MANOVA

We performed a multi-factorial MANOVA to statistically test the differences in group means. The two factors identified by performing exploratory factor analysis served as dependent variables. We included three experimentally altered independent variables (context, adjective terms, definition), all measured control variables (AI knowledge, gender, age, education and occupation) and the main justification types for first-order and second-order inferences from the classification. All predictors were included as categorical variables. For the MANOVA and ANOVA analysis, the R car package was used.

Assumption tests and fitting the model

Assumption tests prior to fitting the model

Although the exploratory factor analysis produced uncorrelated factor scores, we first computed a MANOVA to obtain an overview of patterns between first-order ratings and second-order ratings as dependent variables. Given the lack of correlation and thus no further information from the correlation structure of the dependent variables, we expected a diffused structure of results. Running the MANOVA based on factor scores from the factor analysis with oblique rotation did not change the results. Nine further cases with missing data, i.e., no justification provided for their ratings, were additionally removed.

The following assumptions were tested prior to computing the MANOVA. **Adequate Sample Size**. We applied the one-in-ten-rule for adequate sample size. Our sample size of 3,528 with at least 133 subjects per group based on the experimentally altered independent variables exceeded the threshold of 100 subjects (ten times the number of independent variables: Context, Adjective Terms, Definition, AI Knowledge, Age, Gender, Education, Occupation, Main Justification First-Order, Main Justification Second-Order).

Independent Observations. Given the randomization, all observations were independent. **Outliers Based on Raw Data**. Neither univariate extreme outliers based on the boxplot method with observations being three interquartile ranges far from the first or third quartile nor multivariate outliers based on Mahalanobis distance were identified. **No Multicollinearity**. There was no multicollinearity.

Model Fitting 1: Testing for Interaction Effects

To test the other assumptions based on residual analysis, we fitted a model with interaction terms first. There were no significant interaction effects. All partial η^2 were calculated using the etasq function from the R heplots package.

Model Fitting 2: Residual Analyses

Because none of the interaction effects were significant at α =0.01, they were removed and a new model without interaction effects was fitted. Residual analyses were conducted on the linear model of this MANOVA.

The following assumptions were tested after fitting the MANOVA. **Linearity of Data**. The residuals vs. fitted values plot indicates that the linearity assumption is met. The line is approximately horizontal at zero. **Homogeneity of Variances of Residuals**. The spread-location plot shows that the residuals have an equal variance above and below the line, which is approximately horizontal across the plot. This indicates that the spread of the residuals is approximately equal at all fitted values and that the assumption of homoscedasticity is satisfied.

Normality of Residuals. The histogram of residuals indicates that the residuals are approximately normally distributed. However, in the Q-Q plot of residuals, the points in the lower left and upper right corner of the plot deviate somewhat from the reference line. A further analysis of outliers and influential cases could help identify cases that might cause the deviations.

Observations having extreme residuals (> 3.5, < -3.5), extreme Cook's Distance values (> 0.0056), extreme hat values (> 0.062, < -0.062), or extreme dffits values (> 0.5, < -0.5) were identified and inspected. These thresholds are based on graphical analysis and are all less strict than common thresholds such as the $> 2(p+1)/n$ for hat values (with p being the number of predictors and n the sample size). Model results for the removal of varying sets of outliers and influential cases were compared. Finally, 36 cases having either extreme residuals (> 3.5, < -3.5) or extreme Cook's Distance values (> 0.0057) were removed. Removing more of the previously identified cases did not improve the results.

Model Fitting 3: Final Multivariate Assumption Check

Table 5.6 presents the output for the model after removing the identified 36 cases. Significant effects are highlighted in bold. The panels in Fig. 5.11 indicate that linearity of data, homogeneity of variances of residuals as well as normality of residuals are now met.

Table 5.6: Final MANOVA without interaction effects and with outliers and influential cases removed

	Df	test stat	approx F	num Df	den Df	Pr(>F)	Bonferroni	partial η^2
(Intercept)	1	0.01	21.43	2	3445	0.000	0.000	0.012
first-order justification	6	0.50	190.76	12	6892	0.000	**0.000**	**0.249**
second-order justification	6	0.45	164.60	12	6892	0.000	**0.000**	**0.223**
AI knowledge	4	0.03	13.43	8	6892	0.000	**0.000**	**0.015**
age	5	0.01	4.50	10	6892	0.000	0.000	0.006
gender	2	0.00	1.97	4	6892	0.097	1.000	0.001
occupation	8	0.01	2.38	16	6892	0.001	0.016	0.006
education	7	0.00	0.94	14	6892	0.519	1.000	0.002
context	1	0.04	73.68	2	3445	0.000	**0.000**	**0.041**
terms	5	0.01	3.58	10	6892	0.000	0.001	0.005
definition	1	0.00	0.61	2	3445	0.543	1.000	0.000

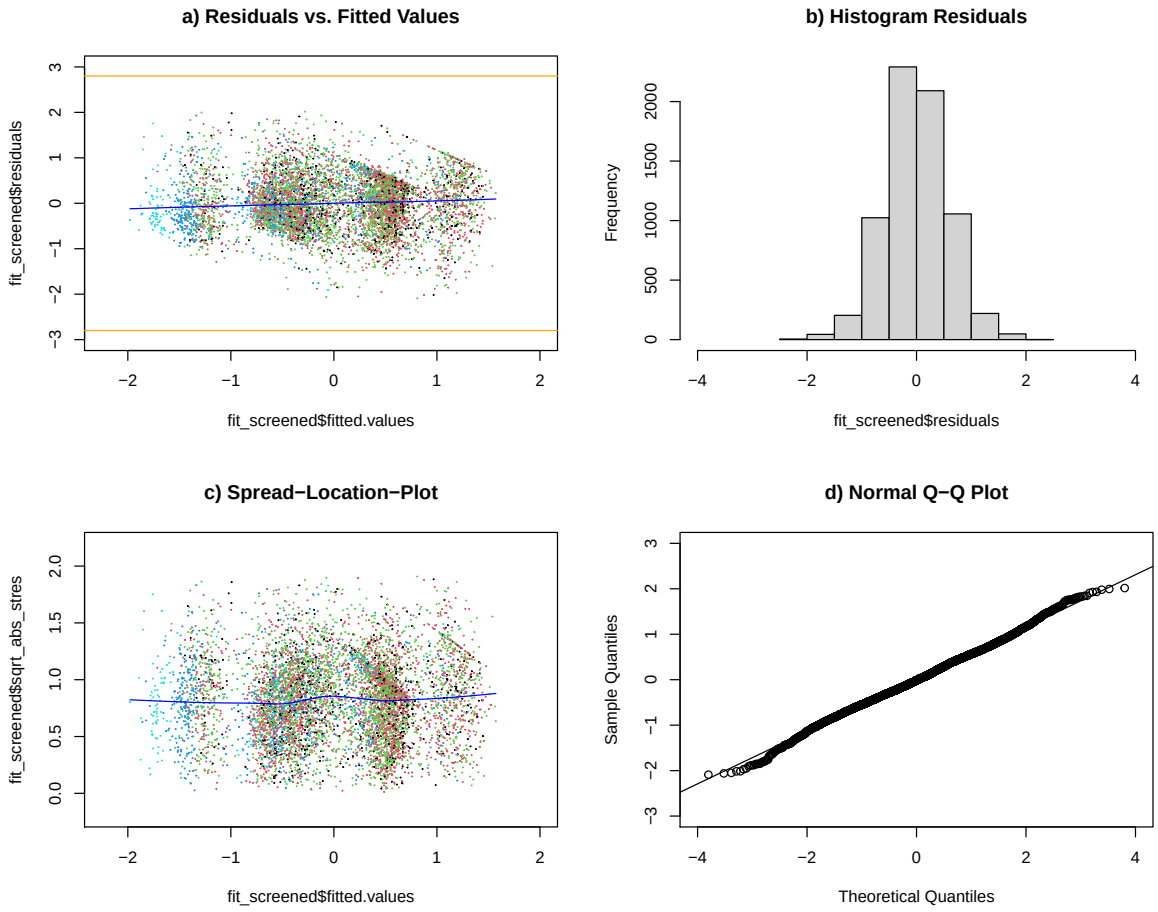


Figure 5.11: Graphical analysis of MANOVA test assumptions after removing 36 identified cases.

Comparison of final model with model based on an equalized dataset

The results of the final model from Table 5.6 were compared to the results of a model for an equalized dataset based on the three experimentally altered independent variables (context, adjective terms, definition). The same outliers and influential cases as in the previous model were removed. After equalization, this dataset contained 3,168 subjects. Because the assumptions based on the graphical analysis did not differ and the results were similar to the previous results of Table 5.6, this model was discarded in favor of retaining more observations in a sample without equalized groups.

Follow-up analysis

To identify which individual predictors had a significant effect on which dependent variable, we conducted univariate analyses.

Univariate Analysis: ANOVA for First-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of 13 identified extreme outliers and influential cases did not improve the homogeneity of variances. To control for the family-wise error rate, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. Additionally, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

Univariate Analysis: ANOVA for Second-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of twelve extreme outliers and influential cases did not improve homogeneity of variances. As we did for the ANOVA for the first-order dependent variable, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. In addition, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

Pairwise comparisons

For first-order inferences, pairwise comparisons for the variable *adjective terms* and the significant experimental variable *context* based on estimated marginal means revealed significant group differences between the advertisement and the hiring context at each level of the variable *adjective terms* (see Table 5.7, rows 1-6). These differences could not be observed for second-order inferences. All groups differed significantly between first-order and second-order inferences (see Table 5.7, rows 7-18). These results are in line with the rating behavior depicted in Fig. 5.10 and the ANOVA results (see Appendix 5.2.7 and for ANOVA outputs

Table 1 of the main text), i.e., the assignment to a context, either advertisement or hiring, had a significant effect on the rating behaviors of participants for first-order inferences. Also, the rating behaviors on first- and second-order inferences within one context differed significantly.

Table 5.7: All significant pairwise tests for context and adjective terms based on estimated marginal means for the complete model

terms	variety	context	contrast	estimate	SE	df	t.ratio	p.value
acceptable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
appropriate	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
fair	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
justifiable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
reasonable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
responsible	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
acceptable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.08	0.00
acceptable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.89	0.00
appropriate	.	AD	factor2nd - factor1st	-0.54	0.11	3454.00	-5.06	0.00
appropriate	.	HR	factor2nd - factor1st	-0.74	0.11	3454.00	-6.88	0.00
fair	.	AD	factor2nd - factor1st	-0.64	0.11	3454.00	-5.87	0.00
fair	.	HR	factor2nd - factor1st	-0.84	0.11	3454.00	-7.67	0.00
justifiable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.01	0.00
justifiable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.81	0.00
reasonable	.	AD	factor2nd - factor1st	-0.58	0.11	3454.00	-5.30	0.00
reasonable	.	HR	factor2nd - factor1st	-0.78	0.11	3454.00	-7.12	0.00
responsible	.	AD	factor2nd - factor1st	-0.71	0.11	3454.00	-6.55	0.00
responsible	.	HR	factor2nd - factor1st	-0.91	0.11	3454.00	-8.36	0.00

The influence of the *justification* variables becomes apparent when computing estimated marginal means for a model without the *justification* variables. When controlling for the *justifications*, the effect of the variable *context* decreases. Nevertheless, the same significant differences of main interest are identified between the AD and HR context.

Subjects' justifications

Documentation of category classes and F1 scores

Table 5.8: Generated category classes for participants' justifications, together with example comments of classified observations per class and test set F-1 score for each class.

	Category classes	Examples	F1 score
1	AI can tell	"You should be able to determine the race of a person with a picture of their face."	0.94
2	AI cannot tell	"You can not tell if a person is likable or not in a photo."	0.96
3	Inference is relevant for the decision making	"Some positions require emotion, or at least sympathy or empathy."	0.96
4	Inference is not relevant for the decision making	"it does not matter if a person is black or white when the AI is recommending products and services"	0.95
5	Inference creates harm (e.g., illegal, discrimination).	"This is unacceptable, as it may be discriminatory against the transgender population."	0.97
6	AI has human biases	"Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.."	0.97
7	Incomprehensible & nonsensical responses	"this person is not fully trustworthy", "Not very like"	0.95

Categories

Table 5.9 defines all categories, provides application descriptions, and differentiates the category to related ones. More examples comments are provided.

Table 5.9: Definition of categories and examples (Code book).

Category	Description	Example
AI can tell (e.g. "easy to tell")	**Definition:** The AI/software is able to/can make an inference because the portrait image provides sufficient evidence for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are suitable/good/sufficient for the AI to make the inference. **Application:** The category is assigned when someone *agrees* that an AI is able to make the inference based on sufficient evidence. Sometimes a *specific reference* to the photograph, portrait, image, picture, or visual data type is made. The word "obvious" can be an indicator to use this category.	*Very easy to tell. All you need is a picture and a database.* (P635/2575) *Can always tell this from a color pic.* (P1329/4565) *AI can determine this easily. It can see if you wear glasses or not.* (P557/2327) *Also extremely obvious and superficial.* (P1257/4338)
AI cannot tell (e.g. "not easy to tell")	**Definition:** The AI/software is not able to/cannot make an inference because the evidence in the portrait image is insufficient for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are not suitable/good/sufficient for the AI to make the inference. **Application:** The category is assigned when someone *disagrees* that an AI is able to make the inference. In some cases, it is *specifically highlighted* that a facial image or visual data type is not correct/insufficient to make a certain inference.	*AI cannot determine whether a person is trustworthy or not.* (P333/1605) *Intelligence is not a physical trait and cannot be determined from a photograph by an AI.* (P220/1207) *You cannot determine whether someone is intelligent based on the way that they look.* (P1362/4610)
Inference is relevant for the decision making	**Definition:** The inference is relevant/important and/or useful for the purpose of application. **Application:** This category is assigned if someone explains why/that a certain inference is relevant for making a decision for a specific application.	*[...] this piece of information is needed for better predictions.* (P260/1339) *[...] I think having emotions is a crucial part of an interview.* (P3515/5661)

Inference is not relevant for the decision making	**Definition:** The inference is not relevant/important/appropriate and/or not useful for the purpose of the application. **Application:** This category is assigned if someone explains why/that a certain inference is not relevant for making a decision for a specific application.	*It does not matter whether a person is assertive or not.* (P46/550) *A sex does not define a person.* (P1109/3856)
Inference creates harm (e.g. illegal, discrimination)	**Definition:** An AI inference is considered discriminatory and/or violates personal rights. **Application:** This category is assigned when drawing an inference would lead to a discriminatory outcome or harm a person in any other way.	*this form of racism should be unacceptable. you cannot infer such a thing on skin color alone.* (610/2491) *Trying to determine a user's personality and trustworthiness is a pretty massive breach of privacy.* (P133/894)
AI has human bias	**Definition:** Inference is affected by human bias; the inference cannot be made without human bias. **Application:** This category is assigned if someone highlights the dependency of AI on humans and hence the implicit integration of human bias, for example, into the data and ultimately into the decision made by an AI.	*I do not see how an AI could make such a determination without relying on human biases to be programmed into it. [...]* (P1862/1966) *Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.* (P1708/1272)
Incomprehensible responses	**Definition:** The comment is unrelated to the task and/or contains text copied from the instructions or nonsensical text. **Application:** This category is assigned if the comment is not a justification for the rating. Additionally, this category is applied if it becomes apparent from the comments that a participant did not understand the task. If one comment of a respondent can clearly be assigned to this category, all comments by this same respondent have to be assigned to this category, because it cannot be assumed that the person trustfully filled out the questionnaire.	*ok a so like in* (P1419/4830) *they are intelligent* (P607/2486) *I agree that person is or is not wearing glasses. because it is useful to portrait a person.* (P928/3352)

Justifications results for the "Glasses" inference

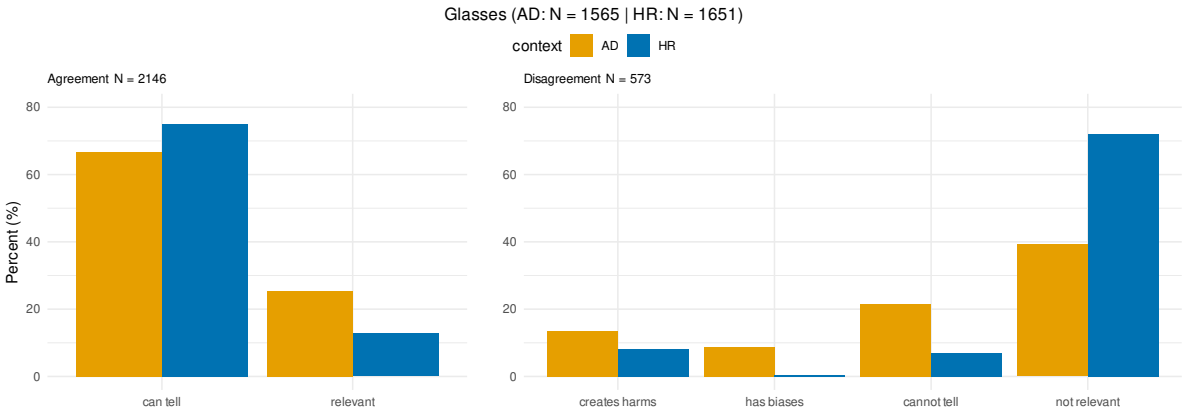


Figure 5.12: Justifications results for the "Glasses" inference.

5.3 Research Article 3: AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI.

Authors

Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Michel Hohendanner, Jens Grossklags

Publication Outlet

EAAMO'22: Equity and Access in Algorithms, Mechanisms, and Optimization; October 2022; https://doi.org/10.1145/3551624.3555294

Abstract

Recent advances in computer vision analysis have led to a debate about the kinds of conclusions artificial intelligence (AI) should make about people based on their faces. Some scholars have argued for supposedly "common sense" facial inferences that can be reliably drawn from faces using AI. Other scholars have raised concerns about an automated version of "physiognomic practices" that facial analysis AI could entail. We contribute to this multidisciplinary discussion by exploring how individuals with AI competence and laypeople evaluate facial analysis AI inference-making. Ethical considerations of both groups should inform the design of ethical computer vision AI. In a two-scenario vignette study, we explore how ethical evaluations of both groups differ across a low-stake advertisement and a high-stake hiring context. Next to a statistical analysis of AI inference ratings, we apply a mixed methods approach to evaluate the justification themes identified by a qualitative content analysis of participants' 2768 justifications. We find that people with AI competence (N=122) and laypeople (N=122; validation N=102) share many ethical perceptions about facial analysis AI. The application context has an effect on how AI inference-making from faces is perceived. While differences in AI competence did not have an effect on inference ratings, specific differences were observable for the ethical justifications. A validation laypeople dataset confirms these results. Our work offers a participatory AI ethics approach to the ongoing policy discussions on the normative dimensions and implications of computer vision AI. Our research seeks to inform, challenge, and complement conceptual and theoretical perspectives on computer vision AI ethics.

Contribution of the Doctoral Candidate

Conceptualization, methodology, investigation, writing - original draft, writing - review & editing

5.3.1 Introduction

Companies and research institutes increasingly produce and release artificial intelligence (AI) applications that draw conclusions about individuals from human faces [426, 427, 428]. One task of such facial processing technologies is facial analysis (hereafter called *facial analysis AI*), which classifies facial characteristics as demographic or physical traits [346] and even personality traits from portrait images. Driven by scientific advances in the areas of face-based inferences on intelligence, trustworthiness, likability and other personality traits [429, 430, 371], as well as sexual orientation [363, 362], such AI products find application in various domains including human resources and advertising. In response, a community of critical data scientists has raised ethical concerns regarding the development of such facial analysis AI [378, 379, 380, 431, 382].

In policy-making, researchers from various disciplines have argued that the veracity of inferences from faces is not significant enough to counterbalance negative consequences [349], and have pointed out the unreliability of human inferences from faces, such as trustworthiness or intelligence [432, 389]. Others have highlighted the variability and context-dependency of emotions depicted in pictures and videos showing faces [31]. Members of the European Parliament recently called "for a ban on the use of private facial recognition databases" [433]. Moreover, serious misclassifications have been uncovered in commercial gender detection tools [53] and job candidate selection software [383, 256]. Nonetheless, many industry actors see an enormous market potential – the AI emotion recognition industry alone is predicted to become worth multiple billion dollars in the coming years [374].

Fundamental questions are how to draw a line between ethically permissible and impermissible AI facial inferences as well as who should be involved in making these decisions. These two questions are central to understand how AI systems and their regulatory frameworks can be developed in a socially-sustainable manner. We contribute to this research debate by exploring how laypeople and individuals with AI competence evaluate facial analysis AI inference-making. We believe that both groups, potential future designers of AI systems and subjects of facial analysis AI, should play a more critical role in the development of ethical computer vision AI.

Prior work has illustrated that the general population (i.e., laypeople) may be aware that facial analysis AI applications exist but that it has little knowledge of their technological characteristics [434]. Mainstream media and science fiction contribute to the propagation of AI narratives that create unrealistic expectations of AI capabilities [435, 434, 436, 437, 438, 439, 440], and pay little attention to their feasibility [441]. Hopes and fears are part of AI narratives [436] and although some argue that current perceptions are skewed or extreme [437] such perceptions can influence the acceptance and adoption of AI systems by the general public [435, 436, 437, 439, 440, 441]. How popular narratives on technology, including the role of AI, can influence the imagination of future societies has, for instance, been explored using research through design and narrative analysis [e.g., 442, 443].

It has become increasingly clear that challenges arising from AI systems do not have purely

technical solutions. For example, the decision to use one fairness metric over another requires value judgments that cannot be solved by formalistic approaches. Normative decisions *always* attract support, skepticism or rejection by different groups in society. Achieving consensus on topics such as "algorithmic fairness will be difficult unless we understand why people disagree in the first place" [444, p.1]. In the context of facial analysis AI, we believe it is important to understand how individuals with AI competence perceive AI inference-making and how their perception differs from the perception of AI inference-making by laypeople. Overall, we ask the following research question:

How do ethical justifications of AI inference-making from faces differ between individuals with AI competence and laypeople?

We build this research on our prior work in which we explored a conceptualization of reasonable inference [237] and asked laypeople how they evaluate such inferences [445]. In this study, we extend this work and compare evaluations of AI inference-making of laypeople with those of individuals with AI competence. We first survey researchers and students studying AI or computer vision AI (N=122) for our sample of "individuals with AI competence". We then compare their ratings and open-text justifications to a laypeople dataset (N=122). Furthermore, we analyze whether a range of demographic factors correlates with differences in the ethical evaluation of AI inference-making from portrait pictures. We confirm the results using a validation laypeople dataset.

5.3.2 Related work

Research on AI inferences of social constructs and character traits from faces

Many companies have developed facial analysis products used for market research, customer targeting, health care or education. For instance, Face++ sells services that infer "face related attributes including age, gender, smile intensity, [...] emotion, beauty" [428]. EmoVu [446] and FaceReader by Noldus perform facial expression analysis and infer, amongst others, personal characteristics and the six basic emotions [447] "happy, sad, angry, surprised, scared, and disgusted" [448]. Betaface and SkyBiometry classify glasses, beard, mustache, mood, or ethnicity [449, 450]. Faception claims to be able to identify people with high IQ [426].

The foundation for these analyses stems from research on inferences from human faces by humans. Research in evolutionary anthropology and psychology presents findings that humans "cannot help" but form first facial impressions despite their proven inaccuracy [349, 44, 350, 337, 351]. In the past, organizational and institutional physiognomic practices relied on making inferences about character traits from visual appearance [345, 346, 347, 256, 348]. Well-known for their contributions to physiognomy, Francis Galton, Caspar Lavatar or Cesare Lombroso, amongst others, developed taxonomies of character interpretations and corresponding facial configurations (see [44] for physiognomy's history). Today, a line of research persists that advocates the accuracy of first facial impressions [45, 353, 354, 355]. Research in computer vision datasets, algorithms, and models is clearly aware of this line of research. Projects in computer vision AI have asserted to successfully infer sexual [362, 363]

and political orientation [364, 365] or emotion intensity and emotion expression [360, 361] based on people's faces in images. Others claim to be able to infer a variety of latent traits in personality assessment, such as trustworthiness [429] or the big 5 personality traits [369, 366, 367, 370, 451, 318, 319, 320] from profile images. However, considerable evidence suggests that first facial impressions do not surpass a "kernel of truth" [349, 350, 389, 44, 390, 337].

Researchers in the field of critical data science highlight ethical concerns arising from classifying individuals with AI on the basis of their facial appearance. Image-based inferences about people can only represent visibly apparent factors of an inferred concept [388]. However, as such inferences are used today, they may be based on bold or questionable semiotic assumptions when predicting intentions, aims, and capabilities or characters of individuals based on their facial characteristics found in portrait images [398, 345]. Judgments of this kind are epistemologically unreliable [256, 237]. Some researchers have argued that such systems are morally objectionable because they treat individuals as categorized objects [399, 388], and others have proposed to abolish physiognomic AI [256].

Does knowledge of AI correlate with ethical perceptions of AI?

While prior research has investigated users' perceptions of AI-based systems, only a handful of research studies exist that investigate experts' ethical perceptions of AI systems [452, 444, 453]. Here, measuring AI knowledge has proven to be difficult. Approaches vary from attempts to identify actual AI knowledge over the recruitment of specific subject pools to measures involving programming and numeracy skills (see Appendix A.1 for an overview). Another difficulty in comparing the studies arises from the diversity of application contexts and the diversity of AI systems, e.g., "automated decision-making by AI" [454], "expert systems" [455], "algorithms" [456], "artificial intelligence" [457] , or "algorithmic decision-making" [452].

Some positive associations were observed: [454] found that both higher levels of education and technical knowledge, including AI knowledge, have a positive association with perceived usefulness, but no significant association with perceived risk of AI decision-making. Higher technical knowledge levels show a positive association with AI fairness perceptions. Similarly, [455] reported that teachers with knowledge on expert systems perceive higher utility of advice from these systems compared to teachers lacking such knowledge; there was no relation between numeracy and acceptance of algorithmic advice. [456] found that less numerate people appreciate advice from algorithms less in the context of forecasting and estimation tasks.

In contrast, [457] found that AI expertise and perceptions on AI adoption were not related. [458] found that greater levels of computer programming knowledge decreased the perceived fairness of algorithmic decisions in the context of dividing household chores. The authors assumed that participants with higher levels of knowledge were either confronted with unexpected algorithmic decision-making results and/or had greater knowledge about the limitations of such systems. Generally, discussion-based decision outcomes were perceived as fairer than outcomes produced by algorithms. Audio-recorded interviews highlighted the importance of participation in decision-making – i.e., the ability to choose and to agree or

disagree – as well as enhanced social transparency of decision outcomes via discussion of the perceptions of whether an outcome was fair or not. [456] observed that greater familiarity with algorithms led to less acceptance of advice from automated forecasting tasks.

[453] found AI researchers to favor a prioritization of research on AI safety, to support pre-publication reviews to evaluate potential harms, to strongly disagree with AI research on lethal autonomous weapons, and, finally, to highly trust scientific and international organizations in shaping the development of AI applications for the public interest. Across three different scenarios (dynamically-priced premium of car insurance, re-routing of flight passengers, automatic loan allocation), [452] did not find students' AI knowledge to influence ethical perceptions of AI. Instead, individual differences were observed between undergraduate and postgraduate participants. For the context of criminal justice, undergraduate computer science students changed their perceptions of algorithmic fairness after one discussion-intensive class [444]: After the intervention, students preferred adding the gender feature to the algorithms, which may be explained by weaknesses of the concept "fairness through blindness". They also preferred algorithms, as opposed to human judges, and favored algorithmic transparency as a general principle. However, consensus did not increase. Rather, opinions were more varied regarding some topics.

The literature reviewed above reveals mixed results regarding the influence of AI knowledge on AI perception. The present study contributes to this line of research by comparing how ethical perceptions of facial analysis in two different contexts vary between laypeople and individuals with AI competence.

5.3.3 Study procedure and methods

Recruitment process and participants

We recruited 346 survey participants across three samples, one of which served validation purposes. We sampled AI-competent individuals at the end of 2021 and beginning of 2022 (N=122, female=27.05%, male=69.67%, other=3.28%). We targeted graduate and PhD students focusing on AI at two large European universities and one large European research institute via social media and news channels of computer science and data science study programs. We describe the exact filtering criteria to determine AI competence in Section 5.3.4 (and provide further data such as course experience in Appendix A.3.4). Each participant was compensated with a fixed payment of 5€. The mean duration was 16.31 minutes (min: 6.50, max: 32.25). The age distribution was: 46.72% with age 18-24, 49.18% with age 25-34, 2.46% with age 35-44, 0.82% with age 45-54, and 0.82% with age 55 or above (see Appendix A.4 for data cleaning).

We collected a laypeople sample at the end of 2019 and at the beginning of 2020 via Amazon Mechanical Turk (MT) in the course of another study [445]. Participation was limited to those regisered in the United States. We produced a final sample of 3102 participants. For the present study, we randomly selected 122 laypeople (female=46.09%, male=48.36%, other=0%) from all participants who indicated to have either very little or novice AI knowledge (46.09% of the entire dataset). The mean duration was 9.98 minutes (min: 3.87, max: 25.08). The age

distribution was: 8.20% with age 18-24, 36.07% with age 25-34, 23.77% with age 35-44, 13.93% with age 45-54, 9.02% with age 55-65, and 9.02% with age 65 or above.

We collected a validation laypeople sample in June of 2022 in a second semester undergraduate lecture at a large European university (N=102, female=18.63%, male=81.37%, other=0%). We excluded respondents with high AI competence from the sample. The mean duration was 21.88 minutes (min: 5.16, max: 37.4). We assume that the higher average duration was due to the perceived complexity of the AI knowledge quiz by participants who were not competent in AI. 99.02% were aged between 18-24, 0.98% were aged between 25-34. Survey completion was incentivized by being part of a number of voluntary tasks to become eligible for a grade bonus on the final exam. The validation dataset also allowed for a useful complementary comparison with the sample of AI-competent individuals due to their shared similarities in demographic features (gender balance, age and country of origin).

Our home institution does not require an ethics approval for questionnaire-based online studies. All participants in the dataset were informed about the procedure, the length and the basic premise of the study, and gave consent to the use of the data for research purposes. Participants could drop out at any point in the survey, or could exit the survey if they did not agree with the use of their data for research purposes. All analysis data was fully de-identified and the privacy of all subjects was preserved at all times during the study. The service used to collect the data guaranteed compliance with the European Union's General Data Protection Regulation (GDPR). The compensation offered in the two paid studies was above minimum wage.

Vignette study

Experimental vignette studies are a common instrument to study people's perceptions and judgments in a variety of hypothetical scenarios [106, 107, 108, 33, 295, 111, 296, 297, 298]. The design of our factorial vignette study is based on our prior work [445]. It consists of two hypothetical decision scenarios: participants were either drawn into a low-stake advertisement (AD) or a high-stake hiring (HR) scenario. In both scenarios an AI system scans a portrait picture and makes a variety of inferences about an individual. Based on these and other inferences, in the AD context, a social media user will be shown a particular advertisement. In the HR context, an applicant will either be selected or rejected for a job position (see Figure 1 in Appendix A.2). Participants then rated on a 7-point Likert scale their level of agreement or disagreement (1 = "strongly agree", 7 = "strongly disagree") with eight distinct AI-made inferences from a portrait picture, drawn for the above described purpose of the application context: *gender*, *emotion expression*, *wearing glasses* and *skin color*, *intelligent*, *trustworthy*, *assertive*, and *likable*. These ratings are hereafter called *inference ratings*. After each inference rating and before proceeding to the next inference, participants were asked to justify their rating in one to two sentences.

5.3.4 Measuring AI competence

We developed an AI knowledge test with a total of nine questions. Four of them were directed at computer vision, out of which three were based on the computer vision textbook by [459]. The other six questions were based on an instrument designed to assess student's AI and machine learning knowledge by [460]. Here, we adjusted questions for the purpose of this study and removed some items (see Appendix A.3). The AI knowledge test was first discussed with three researchers and the resulting feedback was implemented. The scale was evaluated via a pre-study with three participants, who had varying AI knowledge levels. The pre-study additionally included one question on the difficulty of each item. The pre-study illustrated that the AI knowledge test has easy, moderate and difficult questions, and was able to map out a variety of AI knowledge levels.

Mixed method analysis strategy

All analyses were performed in R and Python.

Content-structuring qualitative content analysis

The design of our research study followed an embedded design, which we analyzed using mixed methods by integrating qualitative and quantitative data [461, 462]. To analyze the application of justification themes, we applied content-structuring qualitative content analysis and developed a detailed category scheme to map justification patterns within the responses by participants [463, 74, 461, 462, 464]. First, one researcher labeled 15% of the two main datasets and formulated 57 detailed categories, which were discussed with a second researcher and grouped into 21 super-ordinate categories. Second, both researchers independently applied this category scheme to 10% [461] of both datasets using the instructions documented in the code book in Appendix C. The inter-coder reliability was above Krippendorff's $\alpha \geq 0.8$ for each of the inferences [465]. Differences were discussed with a third researcher. No further categories were included. Finally, one researcher labeled the entire dataset using the final category scheme. The coding occurred at the word level. This meant that as little as one word up to the entire answer could be assigned a code. Three researchers labeled the validation dataset applying the previously developed category scheme. They achieved Krippendorff's $\alpha \geq 0.7$ for each of the inferences. Differences were discussed and resolved among the three researchers.

Frequency and co-occurrence analysis of justification themes

We analyzed the justification themes using co-occurrence and frequency analysis. We compared the results for subgroups of the sample, e.g., AI-competent vs. laypeople, AD vs. HR context. First, the frequencies of the individual themes were analyzed independently of the co-occurrence with other themes. Second, the frequencies of all unique theme pairs, e.g., the likelihood of two themes being mentioned in combination with each other, were explored.

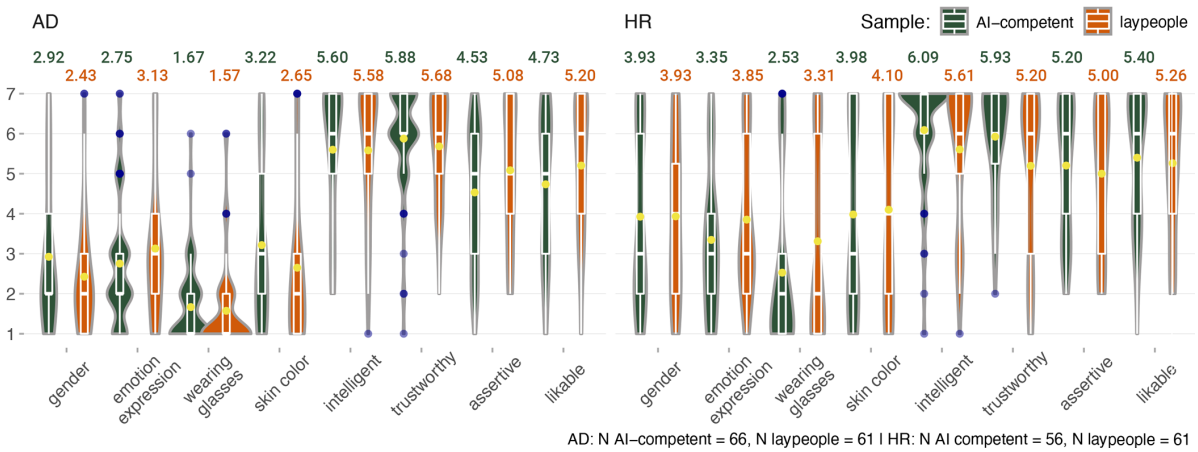


Figure 5.13: Mean inference ratings in AD vs. HR context by sample. Means of inference ratings for each inference by context and sample show that the AI-competent and laypeople (MT) largely agree in their ratings of facial AI inferences. Rating score 1: "strongly agree", rating score 7: "strongly disagree".

Factor analysis, Welch two-sample t-test and analysis of variances

To analyze subjects' ratings, we performed an exploratory factor analysis with orthogonal rotation (varimax), minres factor extraction and regression factor estimation for all three samples. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis [420, 466] and Barlett's test of sphericity indicated that the correlations between items were sufficiently large. For all samples, parallel analysis, BIC, the Velicer MAP and the Kaiser criterion, amongst other tests, suggested retaining two factors (see Appendix B.2 for details). Furthermore, Welch two-sample t-tests and analysis of variances (ANOVA) were computed to directly compare the inference ratings.

5.3.5 Results

Inference ratings show no significant differences between AI-competent and laypeople.

Welch two-sample t-test results

Comparing the inference ratings of the two main samples, none of the Bonferroni-corrected Welch two-sample t-tests shows significant group differences (see Figure 5.13 and Appendix B.1). A robustness check of the results using Yuen's test for trimmed means confirms that there are no significant group differences. The validation laypeople dataset validates the absence of group differences for all inference ratings except for the inference *wearing glasses* ($p_{\text{Bonf.}}$=.04) in the AD context.

Exploratory factor analyses suggest all samples perceive the same two constructs underlying the eight inferences.

Exploratory factor analyses produced the same structure of factor loadings, i.e. two factors, for all three samples. The first factor included the inferences *intelligent*, *trustworthy*, *assertive* and *likable*, which will be referred to as *character and personality traits* in the following. The second factor included the inferences *gender*, *emotion expression*, *wearing glasses* and *skin color*, which will be referred to as *social constructs and features*. Although prior tests (see Appendix B.2) proved the data to be appropriate, some factor loadings did not exceed 0.6 [467], and some of the items (e.g., *gender*) loaded on two factors [468]. We assume that this is due to our rather small sample sizes [467]. Next, we performed robustness checks by repeating the analysis on random sub-samples of 85% of the datasets. The robustness checks validated the findings. These results replicated findings with a large sample in [445]. The observations also confirmed the results from the Welch two-sample t-test: participants in both samples gave similar agreement-disagreement ratings to each of the inferences.

AI-competent and laypeople apply similar levels of complexity to their justifications.

To understand how AI-competent and laypeople justified their inference ratings, we first performed a complexity analysis of the open-text justifications. The analyzed justifications consisted of as little as one word up to a few sentences. Depending on the number of arguments embedded in the justification, we assigned a varying amount of themes during the labeling process. For instance, one participant gave the inference *likable* the rating "strongly disagree" and explained that one "absolutely can't tell if someone is likable because of the way they look. It's actually insulting and misleading and unfair to do that." This justification was labeled with the two themes "not sufficient/ good evidence (data) for task", and "bias/ stereotypes/ discrimination". We refer to justifications of this type as two-theme justifications. The use of fewer arguments could indicate that participants have a clear opinion regarding an inference. The use of more themes could indicate a more diverse and complex spectrum of viewpoints regarding an inference.

The analysis (Table 5.10) shows slight differences in the complexity of justifications by context and inference type. Subjects in the HR context and additionally laypeople in the AD context, provided somewhat more one-theme and less two-theme justifications when justifying their ratings on *character and personality trait* inferences than when justifying their ratings on *construct and feature* inferences. This suggests that evaluations were somewhat clearer for inferences on *character or personality traits*. In contrast, participants discussed inferences on *constructs and features* more diversely.

Context matters: People agree more with AI inferences in the AD than in the HR context.

We then turned our attention to the experimental variable *context* (AD context vs. HR context) to understand whether and how it influences ratings and justifications of participants.

Table 5.10: Complexity of subject's justifications (in %)

	AI-competent		laypeople		validation^{+}	
Type	AD	HR	AD	HR	AD	HR
Inferences on constructs and features						
One theme	66.7	64.3	70.9	74.6	62.3	56.4
Two themes	29.2	31.2	27	23.8	30.9	29.4
Three themes	3.8	4.5	2	1.6	6.9	14.2
Four themes	0.4	-	-	-	-	-
# open text answers	*264	224	244	244	204	204
Inferences on character and personality traits						
One theme	66.3	76.8	79.5	80.7	58.8	64.7
Two themes	28.8	19.2	19.3	18.4	32.4	25
Three themes	4.9	2.7	0.8	0.8	8.8	10.3
Four themes	-	-	-	-	-	-
# open text answers	*264	224	244	244	204	204

* After cleaning of the data, more participants from the AI competent sample happened to be in the AD than HR context.
$^{+}$ More multi-theme justifications by the validation sample may be explained by the longer survey duration.

People agree more with AI inference-making in the low-stake AD context and less in the high-stake HR context.

In all three samples, subjects in the HR context showed significantly less agreement with AI facial inferences than subjects in the AD context (AI-competent ($mean_{\text{AD}}$ =3.90, $mean_{\text{HR}}$ =4.54): $t_{\text{Welch}}(99.08)$ =-3.35, p<.01, $\hat{g}_{\text{Hedges}}$ =-0.62, $\text{CI}_{95\%}$ [-0.99,-0.25]; laypeople ($mean_{\text{AD}}$ =3.88, $mean_{\text{HR}}$ =4.54): t_{Welch} (118.09) =-3.91, p<.01, $\hat{g}_{\text{Hedges}}$ =-0.71, $\text{CI}_{95\%}$ [-1.07,-0.34]; validation ($mean_{\text{AD}}$ =4.06, $mean_{\text{HR}}$ =4.71): t_{Welch} (98.86) =-3.35, p<.01, $\hat{g}_{\text{Hedges}}$ =-0.66, $\text{CI}_{95\%}$ [-1.06,-0.26]). These results indicate that the application context has an impact on participants' evaluations.

The decision context is the most influential factor in participants' ratings.

We performed one six-way ANOVA for each of the eight inferences to analyze the effect of context on the inference rating while controlling for gender, age, education, country, and sample. The variable sample included the AI-competent and laypeople (MT) sample. Using Pillai's trace, ANOVAs with Bonferroni corrections for the eight tests showed that only the variable *context* had a statistically significant effect on inference ratings of *gender* (p<.001), *emotion expression* (p=.015), *wearing glasses* (p<.001) and *skin color* (p=.001). Bonferroni-corrected ANOVAs including the AI-competent and validation laypeople dataset confirmed these results, except for the inference *emotion expression*. We found no other significant effect for any other variable (see Appendix B.3).

Perceptions on the relevance of 'construct and feature' inferences are mixed; in the HR context, laypeople perceive inferences on 'character and personality traits' as relevant.

The influence of the decision context was particularly evident when participants emphasized the "irrelevance" or "relevance" of *construct and feature* inferences (see Figure 5.14, light and dark orange). Participants evaluated these inferences as more "relevant" in the AD context and more "irrelevant" in the HR context. Similarly, participants used the theme "inference (only) sometimes relevant" more frequently in the HR context. This tendency was observed in all samples.

Both laypeople samples applied themes of "(ir)relevance" more frequently than participants with AI competence. Surprisingly, this was particularly the case for MTurk laypeople in the HR context for inferences on *character and personality traits* ("relevant": 15.7%, see Figure 5.14 light orange). For instance, participants from this sample justified that inferring *intelligence* "would give a hint as to how [...] [applicants] would perform on the job" or that inferring *trustworthiness* "in the workplace can be important and it's not wise to have a dishonest person around". For inferences on *constructs and features*, laypeople underlined the "irrelevance" of the inferences *wearing glasses* (26.2% of laypeople; 29.4% of validation laypeople) and *skin color* (27.9%; 39.2%) in the HR context and the "relevance" of the inferences *wearing glasses* (26.2%; 33.3%) and *gender* (26.2%; 29.4%) in the AD context. Some AI-competent subjects drawn into the AD context agreed that the inferences *wearing glasses* (21.2%) and *gender* (18.2%) are relevant to be inferred (see Appendix D.1).

Participants justify ratings on *construct and feature* inferences with a wide variety of themes; ratings on *character and personality* inferences with "insufficient data" themes.

Next, we analyzed whether specific themes were of special importance when justifying inference ratings on *constructs and features* or *character and personality traits*.

Ratings on 'construct and feature' inferences are explained by a variety of justification themes.

As depicted in Figure 5.14, all subjects frequently applied themes highlighting "AI ability", "sufficiency" of the data, and – depending on the AD or HR context – the "relevance" or "irrelevance" of an inference. AI-competent participants raised somewhat more "ethical and discriminatory concerns". Overall, justifications included a substantial variety of justification themes.

Ratings on 'character and personality trait' inferences are predominately explained by the "insufficiency" of a profile picture as evidence.

The use of the "insufficiency" theme was particularly prevalent for laypeople in the HR context (AI-competent: 37.5%, laypeople: 56.7%; validation: 39.3%). Again, individuals with AI competence raised "ethical and discriminatory concerns" more often than participants in

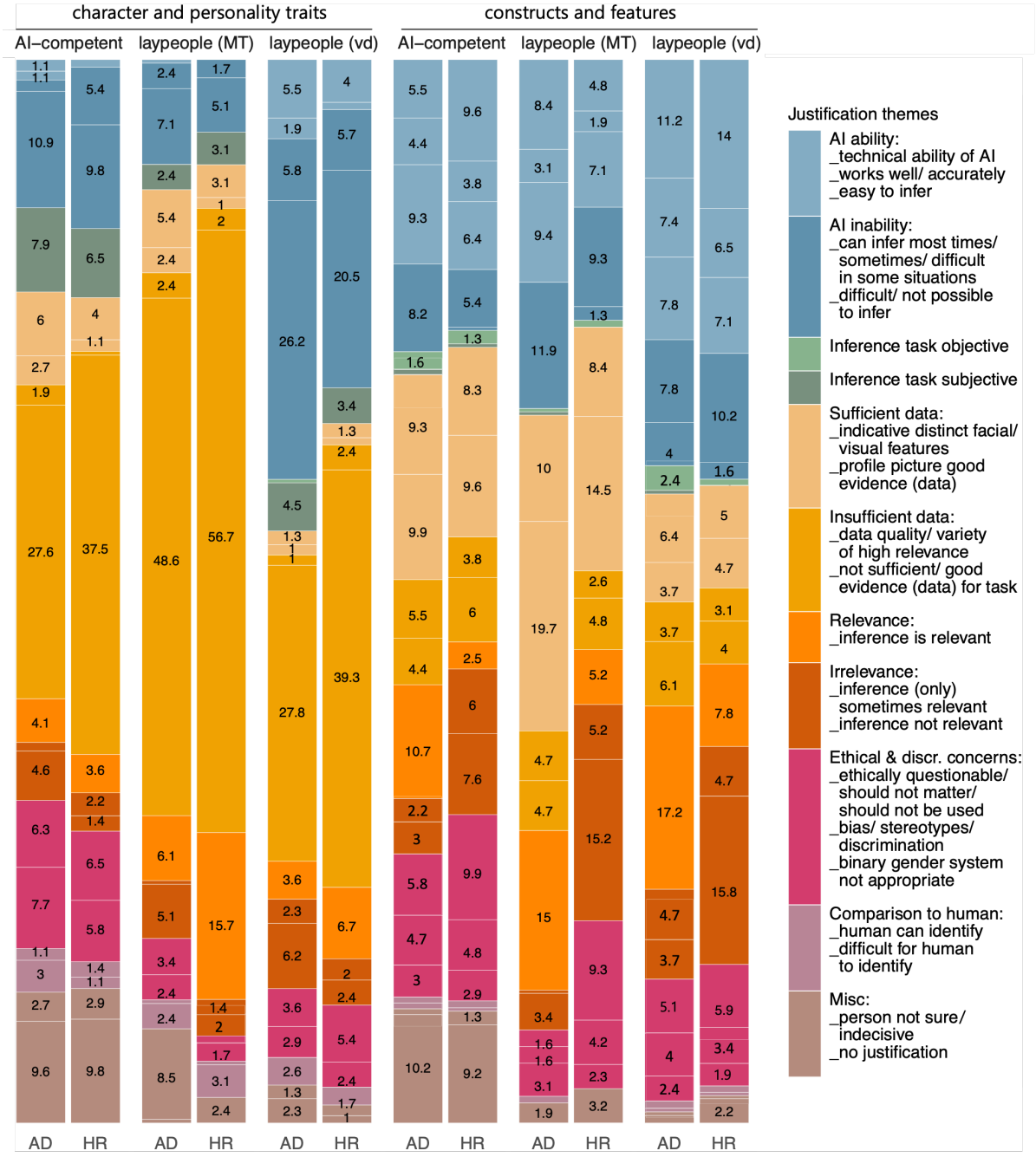


Figure 5.14: Percentages of individual themes grouped by super-ordinate topic, by context, and by sample. Stacked bars add up to 100% and represent the total of individual themes used by the specific sample. Only percentages > 1% are labeled on the graph.

Percentage of individual themes used.

both laypeople samples. Furthermore, participants made references to the "subjectivity" of the inference task.

Participants believe "AI can infer" whether a person is wearing glasses on a portrait picture; they are skeptical about AI's ability to infer emotional expression.

All three samples used the themes "technical ability of AI", "accurate and well working" models, and "easy to infer" most frequently to justify ratings on the inference *wearing glasses*. They applied the theme "can infer sometimes/ difficult in some situation" most often to justify ratings on *emotion expression* and *gender*. For instance, one participant explained that while "the majority of people can have a gender revealed through just a picture, not everyone fits that mold."

Some participants from both main samples believed that a "profile picture is good evidence" for the inferences *wearing glasses* and *emotion expression*. At the same time, there were critical voices stating that a profile picture is not sufficient evidence to infer *emotion expression*, e.g., "Emotion changes by the hour or minute. Can't make an inference based on that." The validation dataset supported these latter results.

Co-occurrence analysis: "AI (in)ability" and data-related themes co-occur most often with other themes.

We then analyzed the co-occurrence of themes with each other to identify patterns in the use of multiple justification themes (see Appendix D.2). We found that for inferences on *constructs and features*, the AI-competent raised concerns but acknowledged AI to be able to make certain inferences. Referring to inferences on *constructs and features*, people with AI competence raised "ethical and discriminatory concerns" in combination with almost all other justification themes, however, most frequently in combinations with themes on "AI ability" or the "sufficiency" of the profile picture as evidence (see Figure 5a and 5b-1 in the Appendix). This relationship reversed for justifications of ratings on *character and personality trait* inferences. Here, "ethical and discriminatory concerns" were most frequently brought forward in combination with themes on the "insufficiency" of a profile picture as evidence (see Figure 5a and 5b-3 in the Appendix).

For inferences on *character and personality traits*, laypeople often paired comments on the "(in)sufficiency" or "(in)adequacy" of the data with another theme. For *constructs and features*, a greater variety of theme combinations was observed.

Many inferences are based on questionable norms or resemble social constructs and societal stereotypes.

To understand participants' most critical concerns, we finally focused on themes related to "ethical and discriminatory concerns" and "AI inability" (see Figure 4 in the Appendix).

Individuals with AI competence perceive the inference *likable* as subjective.

More than laypeople, individuals with AI competence and subjects from the validation sample described the inference *likable* as "relative", "based on sympathy", and "subjective", e.g., "Likability is a matter of perspective" and "depends on the observer." Comments also referred to other justification themes such as ethical concerns, e.g., "Likability is a highly subjective measure and inherently biased. In addition, it is highly unethical to have such type of decisions made by systems that are not capable of understanding the impact of this decisions" [sic] or "Likeability itself is an ill defined thing, predicting it from just portraits is wrong". Participants did not consider any other inference as equally subjective as *likable*.

Some subjects state that inferences on 'character and personality traits' cannot be inferred. However, approximately half of subjects highlight that the data is simply insufficient or inadequate.

A considerable amount of subjects from all samples stated that a profile picture is "insufficient" data (26%-79% depending on inference, context, and sample) to infer *character and personality traits*. For instance, subjects commented that "[n]o facial features indicate trust", or that *intelligence* "is not quantifiable through visual data". At the same time, a minority (~15%) of the AI-competent, a small percentage of laypeople, and many participants from the validation dataset argued that AI cannot infer specific *character or personality traits*. An AI-competent participant explained that the "problem here is ill-posed", there "is no general understanding", and "no clear" or "objective definition of intelligence that everyone agrees with!" Given the lack of shared definitions, some asked "how is this measured? How is it implemented during training?", and "What are the parameters for identifying someone as intelligent?" These findings suggest that some participants evaluated inferences such as *intelligent* and *trustworthy* as social conceptualizations that require a common understanding before being used as inference in facial analysis AI.

Participants with AI competence believe that stereotypical judgments enable AI to draw 'character and personality traits'.

Other people with AI competence worried about "stereotypes" embedded in the training data. They elaborated that, e.g., "a categorization of intelligence based on looks seems to correlate features that are not correlated" or that "the training data for trustworthiness depends on societal stereotypes and not actual trustworthyness" [sic]. Conversely, the existence of "stereotypes" was also used to argue in favor of AI being able to make an inference. For instance, a participant explained that the inference *likable* "makes sense because some people's appearance is appealing to more people. But, this inference can only be made on a statistical basis: Person is or is not likable on average." AI-competent participants stated justifications in relation to "bias, stereotypes and discrimination" most frequently when referring to the inferences *trustworthy*, *assertive*, and *likable*, e.g., one participant commented that "it's an unethical idea to give ai systems the ability to inference something so loosely defined and

this will lead to biased choices made in the name of "science"." Laypeople did not show these levels of concern for any of these inferences.

A minority of participants raises concerns regarding the inference skin color.

In the HR context, 23% of subjects from all samples raised "ethical concerns" regarding the inference *skin color*. One subject commented that *skin color* "should not be a criterion for job applications. Furthermore, being of a certain skin color should be a matter of self-description and not be determined by a computer program". Some participants also perceived the inference *skin color* to be based on biased data or to lead to discrimination: "Users will get predictions based on race and race-based stereotypes" or "if the model is biased towards skin color, it may not encourage a fair AI agent." Some subjects highlighted that *skin color* can be inferred but should not be done or used: "Color can be detected easily by computer vision frameworks (though this inference imposes certain ethical questions)" or "While it is possible to determine the skin color of a person from a portrait [...], it is ethically incorrect to base any decisions on skin color" or "Detecting skin colour should be trivial for the software, so it is reasonable to expect that inference. It is NOT reasonable that this information should be used to indicate whether someone is suitable for the job." These comments exemplify the diversity of normative evaluation of the inference *skin color*. Although suggesting that AI can infer *skin color*, this inference – which some specifically relate to "race" or "ethnicity" – was perceived as an impermissible inference by a considerable number of subjects.

A minority of participants highlights that binary gender norms are not appropriate and ethically questionable.

Referring to the inference *gender*, some participants raised "ethical concerns" in the HR context (AI-competent: 16.1%; laypeople: 11.5%). In both contexts, 9% of participants with AI competence believed that inferences on gender are based on biased data: "The AI might learn to assign gender identity based on a heavily biased training data which are influenced by conventional gender identity norms hence making fateful inferences in the real world. Such inferences are unreasonable". Some subjects across all samples specifically highlighted that "gender norms are not appropriate" anymore: "This used to be a more 'objective' decision, however society has changed and persons can decide by themselves their gender, without being guided by their appearance. The most important part is, again, the inability of an AI system to understand the consequences of deciding something like this". Others commented that gender can be inferred but is not appropriate: "this is very apparent and thus somewhat alright, but then again, gender is a fluid concept". Some participants believed gender to be a social construct that is not binary as is often presupposed by facial analysis AI.

5.3.6 Key Observations and Discussion

Overall, our study on the ethical perceptions of facial analysis AI suggests that there are no "common sense" facial analysis inferences. In all samples, there are participants who raise

concerns, in particular, *ethical concerns* that inferences lack epistemic validity, should not matter or should not be used for the purpose of an application. In addition, we find that both AI-competent and laypeople express a variety of normative concerns regarding AI facial inferences. At the same time, only a minority of participants concluded that AI cannot, under any circumstance, make an inference from faces.

Regarding the facial inference *emotion expression*, participants note that a profile picture is only a snapshot and thus, "temporary and short-lived". Recently, emotion researchers have argued that emotion expression is more context-dependent and variable than commonly assumed. The *emotional state* of a person cannot be readily inferred from a person's facial expression [31]. Participants in both samples raised similar concerns. For example, one participant stated that there "are numerous people that tend to hide their emotions through pictures [...]".

Our analysis of justifications clearly shows that participants voice concerns regarding the classification of latent traits by facial analysis. Participants pointed out that the inference of attributes such as *intelligence* from facial information presupposed a highly simplified definition of a multidimensional concept. Similarly, participants mentioned potential problems related to the subjectivity associated with inferring attributes such as *likability* from faces.

We found that participants criticized the ethically problematic application of a binary conceptualization of *gender*. This finding aligns with recent critical data science research on computer vision. Here, authors, too, point to the fact that sensitive categories, such as gender and race, are often treated as "common sense categories" in computer vision datasets [380, 345, 382, 383].

On the other hand, a justification theme among both laypeople and people with AI competence pertains to the *possibility* of an AI inference provided that the "data is correct". This line of reasoning resembles narratives behind facial analysis AI research and commercial tools that try to solve issues with predictive power at the level of data *rather than question their epistemic foundations*. Some of the AI-competent and laypeople used entrenched stereotypical heuristics to evaluate AI facial inferences. While heuristics and stereotypes may initially help humans navigate through complex social interactions, research on the validity of human inferences from faces demonstrates that faces are no "strong and reliable indicator of people's underlying traits" [389, p.569].

Some specific differences between the two main samples could be observed. Both laypeople samples applied more pragmatic justifications referring to the "(ir)relevance" of the visual data for a decision-making procedure. For inferences on *character and personality traits*, more than half of laypeople (MT) described the data as "insufficient" for the inference task. People with AI competence mentioned themes related to "(ir)relevance" and "insufficiency" less frequently than laypeople, but raised "ethical concerns" more frequently than laypeople.

The complexities behind participants' justifications indicate a "struggle" for the power over the creation and attribution of meaning for visual data. Our study asks who can and should participate in this discourse. AI experts currently have free rein over the meaning that their

datasets should be attributed with. However, politicians are aware of the complexities behind the meaning of visual data [e.g., 433] and we highlight again that more and more critics are voicing ethical concerns [e.g., 345, 387, 388, 347, 383, 382]. One of our main concerns is that the inference of perceived traits or features, e.g., "perceived trustworthiness" [e.g., 469] as opposed to "actual trustworthiness" by an AI system ultimately contributes to society remaining trapped in a cycle of stereotypes.

Taken together, we note that participants in all samples showed a tendency *to oppose* facial AI inference-making. Participants' evaluations underline many of the ethical complications of facial analysis AI that have recently been raised by critical data scientists and other scholars. Moreover, we see that people do not apply a consistent and universal justification profile for each of the facial inferences. Facial inferences are not simple constructs but overloaded with epistemic and pragmatic intuitions that are likely influenced by factors including cultural background.

We end by wondering how a justifiable ethical framework for facial AI inference-making could look like. What "standards" would a satisfactory justification fulfill? Given that we deal with *visual* inferences, we believe that they should first achieve reasonable epistemic validity and that this validity should be supported by scientific agreement over the quality of the evidence. The question then is what a reasonable level of scientific agreement should look like. We have pointed out that while a large majority of researchers underline the invalidity of first facial impressions, there is an ongoing stream of research publications that claim to present evidence on the validity of first impressions.

Participants in our samples disagreed with inferences common in human first impression-making (e.g., trustworthiness, likability etc.) by algorithmic systems. Indeed, one of the core findings of this work is that neither individuals with AI competence nor laypeople trust many of the inferences of facial analysis technology. With legislative attempts seeking to ban certain facial processing technologies, with a plethora of scholars pointing to the dangers of an automated version of physiognomy, and the different sample populations expressing their lack of trust toward such AI inference-making, we ask in what context and under what circumstances such facial analysis AI can be justified at all. It appears that, more often than not, there are better *reasons not to develop and deploy AI* that analyzes human faces to draw a variety of inferences that are then used for a particular decision-making context. Weaving together the argumentation threads from our previous results [445], critical remarks of data scientists and policy-makers, we take it that there is a strong case to be made that such AI inference-making is epistemically invalid, pragmatically of little use, and, overall, contributes and perpetuates stereotypes that stand in conflict with a society's welfare.

Limitations and Future Direction

Our samples were composed of comparatively young people with AI competence that are not representative of all AI researchers. This may have introduced a bias in terms of the participants' understanding of and critiques on social constructs such as gender identities. In addition, this study does not include voices from industry. Future research should also

survey corporate AI developers.

This research makes a methodological contribution by providing an AI knowledge instrument as an alternative to self-reported AI knowledge measures. We hope that the results from the application of the AI knowledge test will act as a starting point for the utilization of a more objective and reliable measure of knowledge on AI. It should be noted that given rapid advances in AI, the questions contained in the AI quiz should be regularly updated.

Our sample included participants from the United States (laypeople sample) and Europe (AI-competent and validation laypeople sample). We addressed the limitation of comparability of the two main samples by creating a validation dataset that shows substantial similarity in terms of demographics with the AI-competent sample. Given the international application of AI systems, diverse study participants are vital. Hence, future studies should explore whether cultural differences influence ethical concerns of facial processing technologies such as facial analysis AI. If there are no such cross-cultural differences then this could serve as evidence for the existence of culturally-universal ethical perceptions of facial inferences.

Whereas we evaluated the perception of AI inferences from profile pictures, future research should also evaluate perceptions of AI inferences from videos. Given that videos are used for a variety of inference tasks [470], the perception of somewhat more accurate results can be expected. However, it remains to be seen whether video data will influence whether such traits *should* be inferred.

5.3.7 Conclusion

As the use of AI grows in popularity and as the impact of AI inference-making on societies increases, so does the responsibility of those who develop such AI systems. A special focus must be placed on exploring the perspectives of a diverse group of people both who are potentially driving the implementation of computer vision and AI and those that are subjected to its inference-making.

This work provides insights into perceptions of AI inference-making by the general public compared to perceptions of individuals with high knowledge of AI. It suggests that, by and large, people with AI competence and the general public share many perceptions about AI inference-making and have distinct context- and task-dependent perceptual differences. Being aware of the perceptions and judgments of people with AI competence, on the one side, and users, on the other side, is essential to develop AI systems that are based on democratic discourse, accepted by society, and sustainable.

Concluding this research, we summarize that the application context does have an effect on how people perceive AI inference-making from faces. While differences in AI competence did not have an effect on the inference ratings, specific differences were observable for the ethical justifications. We found that both laypeople and people with AI knowledge showed more agreement with AI inference-making in the low-stake AD context than in the high-stake HR context. In both contexts, people with AI competence – although only a small minority – raised ethical and discriminatory concerns more frequently than laypeople. Laypeople

made more references to themes related to the (ir)relevance of the inference for the context of application.

Having explored the question whether differences in AI knowledge account for changes in the perceptions of AI inference-making across two contexts, this work extends research in the field of perceptions of algorithmic systems and contributes to the nascent literature on AI experts' perceptions on AI inference-making. The results invite a deeper reflection on the similarities and differences in the perceptions of AI among different people within the general population. With this work, we aim to ultimately contribute to the development of sustainable AI systems that are supported, not only by their developers, but also by the general public.

Acknowledgements

We thank the reviewers for their insightful comments that improved the paper. We thank the study participants for taking part in this study. For their valuable feedback, we thank the participants of the 2021 CEPE/International Association of Computing and Philosophy Conference, the participants of the 2021 Ethics and Technology Lecture Series of the Munich Center for Technology in Society, and the participants of the Venice 2019 Metaethics of AI & Self-learning Robots Workshop.

Funding & Support

This research was conducted with the help of a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

5.3.8 Appendix

Research Design and Methods

Overview of Methods Applied in Studies to Measure AI Knowledge

[457] measured Artificial Intelligence (AI) expertise of practitioners in the communications professions using an 8-item quiz, and AI adoption by asking whether participants were using specific AI applications (e.g., Siri) on their phones or AI devices (e.g., Alexa) in their homes or offices. Knowledge on expert models was measured based on the quality of a definition participants were asked to provide in response to an open-end question [455]. Technical knowledge was measured by means of three questions about self-reported knowledge on computer programming, algorithms and AI [454].

Instead of measuring AI knowledge, [453] surveyed researchers who published in leading AI/ML conferences and assumed them to have high AI knowledge. Others surveyed students studying AI [452, 444]. [452] additionally measured their level of knowledge on fairness in algorithmic decision-making or prior training on topics such as algorithm accountability, transparency and fairness through a self-reported 5-point Likert scale.

Again other studies used knowledge in computer programming and numeracy, as measured by [456] using a 11-item numeracy scale by [471] as a proxy. [456] measured familiarity with algorithms by asking participants how certain they were to know what an algorithm is.

Survey Vignette

Figure 5.15 shows the vignette presented to the participants of the AI-competent and validation laypeople samples, which was based on a the vignette presented in [445]. The same wording and order of text passages were used.

AI Knowledge Measure

Construction

In order to better assess respondent's AI knowledge, we complemented a self-rated AI knowledge level instrument (one item 5-point Likert scale) by an AI knowledge measure (see Table 5.11). This measure is based on an instrument used to asses students' AI and Machine Learning (ML) knowledge by [460]. The developed AI knowledge measure contains single-choice questions on ML of varying degrees of difficulty. The measure by [460] was adapted to the purposes of this study as follows: Four questions that originated from [472] were excluded. Those four questions were originally intended to measure the change in knowledge of AI after a workshop-based intervention. Additionally, four questions were removed that did not seem to be fitting for the purpose of this research study. One item in P4 was replaced by an item that is less philosophically disputable. The wording of P5 was changed slightly to make the items shorter. Furthermore, three questions that were not perceived to be fitting for the purposes of this study were removed. Finally, for all four-item

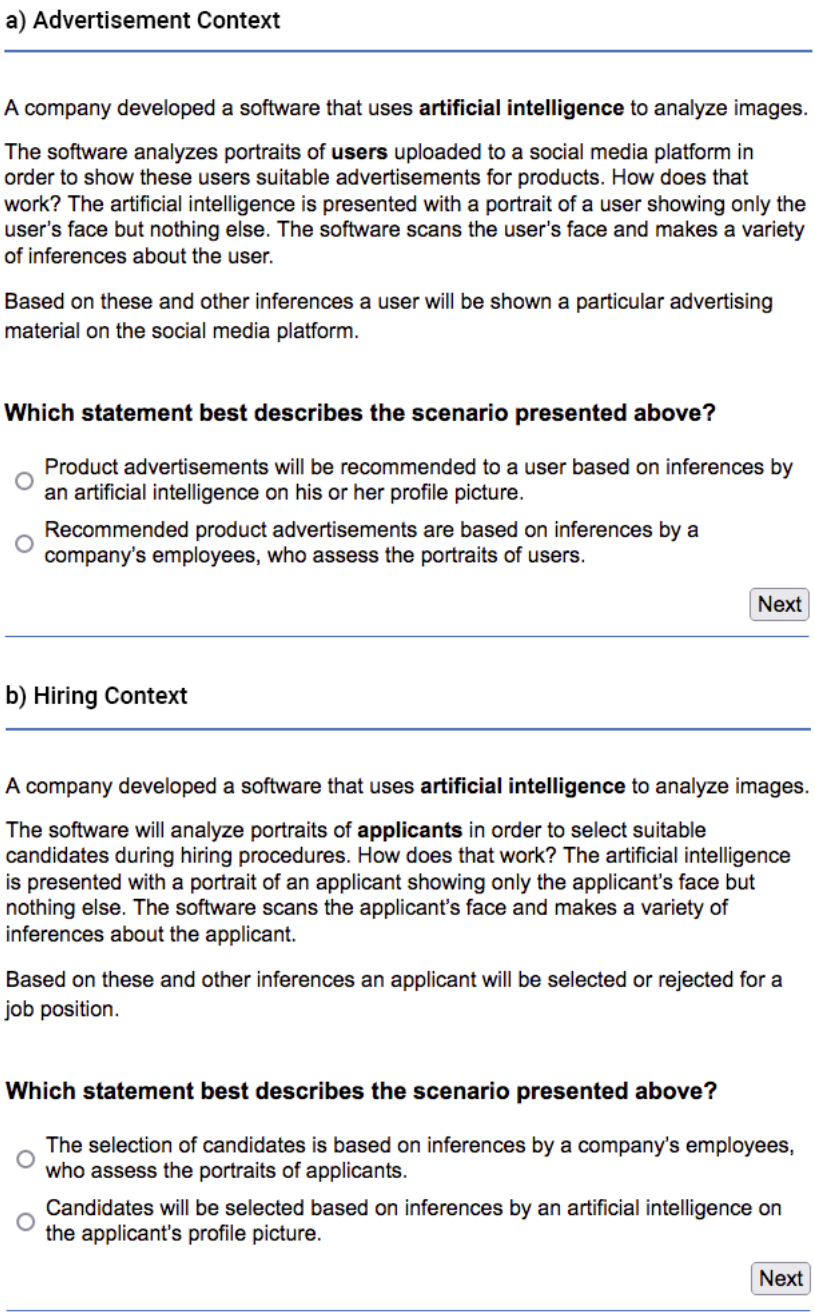


Figure 5.15: Scenario presented to study participants in a) the advertisement context or b) the hiring context.

questions, one wrong item was exchanged with the answer "I don't know". For the question with two items, an "I don't know" option was added. Without the option "I don't know", respondents would have had either to guess or to choose one answer at random, which would have introduced a bias. Given that this research focuses on computer vision, three self-constructed computer vision specific questions (Q7 - Q9) were added, based on [459]. In summary, Q1 through Q5 reflect general questions on ML whereas Q6 through Q9 focus specifically on computer vision and are expected to be answerable by less respondents. In Table 5.11, correct items are marked with an "(X)".

Additional survey questions related to AI knowledge

Besides the questions related to the AI knowledge test, we included a number of additional questions to the survey that allowed us to verify the results from the AI knowledge test. We added two questions on the number of AI courses that the participant took part in (with a technical and with a socio-political or ethical focus). Furthermore, we included three questions to control for the knowledge on the presented AI scenarios, the science of first impression-making, and potential external assistance. To control whether specific AI knowledge might have come from their corporate experiences, we asked participants whether they have an (AI-related) job. We also asked participants how they learned about the survey and what research field best described their research (see Table 5.12).

Validation

Before running our main study, we tested the AI knowledge measure by running a pre-study with three participants. Participants received a survey with the AI knowledge test questions and an additional question designed to indicate the perceived difficulty of each question in the test. Furthermore, the survey asked for an indication of the number of courses with a focus on technical AI, as well as the number of courses with a social-political and/or ethical AI focus. Additionally, participants were asked to indicate their level of AI knowledge on a 5-point Likert-style scale.

Participants were briefed that they were part of a pre-study that helped evaluate the AI knowledge test. Each participant provided feedback on how long it took to complete the survey and whether any questions were misleading. This feedback was gathered and first discussed with the research team. Then, any remaining issues were discussed with an AI expert not part of the research team.

Based on the feedback from the pre-study, the number of mixed examples in P11 for the correct item was increased (from 10 to 1000) to ensure that the strategy described in this item would more clearly result in a the better system. Furthermore, one item was removed from the AI knowledge quiz, because – based on assumptions made by the participant – all of the items might arguably have been correct.

One participant in the pre-study had taken no AI courses and described him-/herself as a novice with regards to AI knowledge. Another person had taken three technical courses

Table 5.11: AI knowledge test: Questions. Changes to original items are indicated.

Name	Orig.	Item and Anchors
Q1	P4	**When an artificial intelligence (AI) system offers results that discriminate in terms, for example, of sex, this is usually due to:** • (X) That the data that was used to train the system was not balanced, that is, that much more data was used for men than women, or vice versa. • That the system is designed to be used by men to a greater extent than by women, or vice versa. • That the system itself tried to be sexist. (new item) • I don't know. (*originally: That the developers of the system had sexist biases.*) • (*deleted item: That the system reflects the sexist reality of human nature.*)
Q2	P9	**In which of the following tasks, to be performed by a computer, would it be appropriate to apply machine learning (ML) techniques?** • (X) Recognize if an email is spam (junk mail). • Count the number of times a key is pressed. • Inform about the hours of a certain business based on the day of the week. • I don't know. (*originally: Add large numbers.*)
Q3	P11	**Both Alicia and Robert want to train a machine learning (ML) system that serves to recognize whether a certain text is "happy / positive" or "sad / negative". Alicia and Robert follow two different training strategies. Who of the two will get the better system?** • (X) Alicia. She has compiled 1000 mixed examples of happy / positive texts and another 1000 mixed examples of sad / negative texts. • Robert. He has collected 1000 examples of happy / positive texts and another 10 examples of sad / negative texts. • I don't know.
Q4	P5	**Imagine we implement machine learning (ML) techniques in a text recognition system. We present the computer with a set of sample texts and the computer, after processing, is able to recognize ...** • only the texts that exactly match those examples. • (X) texts similar to those examples (that is, to recognize new texts that it has not seen before). • any text, image or sound that we present to it. • I don't know. (*originally: any text we present to it.*)
Q5	P6	**Which of the following statements is true about machine learning (ML)?** • (X) Training data is essential for machine learning, without data it is not possible to do machine learning. • The more data we use to train a system that incorporates machine learning, the worse (more inaccurate) are the results offered by that system. • Machine learning does not need data to function, precisely because it is automatic and does not depend on being fed data of any kind. • I don't know. (originally: *With automatic learning, computers learn to think and can recognize any type of data (text, image, sound ...), in the same way that a human being does.*)

Name	Orig.	Item and Anchors
Q6	P7	**Which of the following strategies would be more appropriate to teach a computer to recognize the photo of any apple?** • (X) Train the computer with several photos of different apples, taken in different places and contexts. • Train the computer with several similar photos of the same apple, taken in the same place. • Train the computer with several identical copies of the same photo of an apple. • I don't know. (*originally: Train the computer with photos of dogs.*)
Q7	–	**Which of the following datasets is a classic in the machine-learning community and classifying its content correctly can be considered the "Hello World" of deep learning:** • ImageNet • (X) MNIST • Open Images Dataset • I don't know.
Q8	–	**The best tool for attacking visual-classification problems are ...** • (X) convnets, because they work by learning a hierarchy of modular patterns and concepts to represent the visual world, and the representations they learn are easy to inspect. • densely connected layers, because they learn global patterns in their input feature space, which makes them data efficient when processing images. • basic neural networks, because they learn to associate images and labels, and are energy efficient due to their simplistic computational structure. • I don't know.
Q9	–	**For a multilabel classification, the typical choice for a loss function is ...** • MSE • categorical cross entropy. • (X) binary cross entropy. • I don't know.

Table 5.12: Additional validation questions

Question	Scale
How many courses with a focus on technical AI did you take?	6-point (0 to 5+)
How many courses with a focus on socio-political and/or ethical AI did you take?	6-point (0 to 5+)
In your opinion, how realistic was the scenario?	5-point
How much do you know about the scientific validity of first impressions (based on faces)?	4-point
Did you receive any support for the previous AI quiz? For example, did you consult a search engine (e.g. Google, Bing) or were you helped by nearby friends, colleagues or relatives?	yes/no
Are you currently employed? *exact wording:* yes (IT-related job/company)/ yes (non IT-related job/company)	yes (IT)/ yes (not IT)/ no
How did you learn about this survey? (e.g. which course/ social media/ messaging system)	*open*
Please indicate research field/ study program?	*open*

on AI and two socio-political and/or ethical AI courses and rated his/her AI knowledge as intermediate. Another person had attended five technical courses on AI and three socio-political and/or ethical AI courses and rated his/her AI knowledge as advanced. All respondents had a Master's degree. The reported time needed to complete the quiz was 5, 8 and 10 minutes (order unrelated to presented subjects).

Based on respondents' answers on the perceived difficulty of a question (easy, medium, difficult), a difficulty score was calculated. A question received zero difficulty points when being rated as easy, one difficulty point when being rated as medium and two difficulty points when being rated as difficult. The sum total of the scores collected was then divided by the number of participants. Thus, the difficulty score ranges from 0 to 2. Figure 5.16 displays the questions ordered by their difficulty score.

$$DifficultyScore = \frac{\sum(difficultyPoints)}{N_{\text{respondents}}}$$

People with less knowledge on AI perceived more questions as difficult than people with more knowledge on AI. More specifically, a question that has been perceived as difficult by a respondent with little AI knowledge, was considered as medium by the other two respondents with more AI knowledge. Two questions were rated as easy by all participants: statement about machine learning (training data is essential) and strategy to train an image recognition system (several photos of different apples taken in different places and contexts).

Furthermore, the results from the pre-study hint at a difference in answering behavior, i.e.,

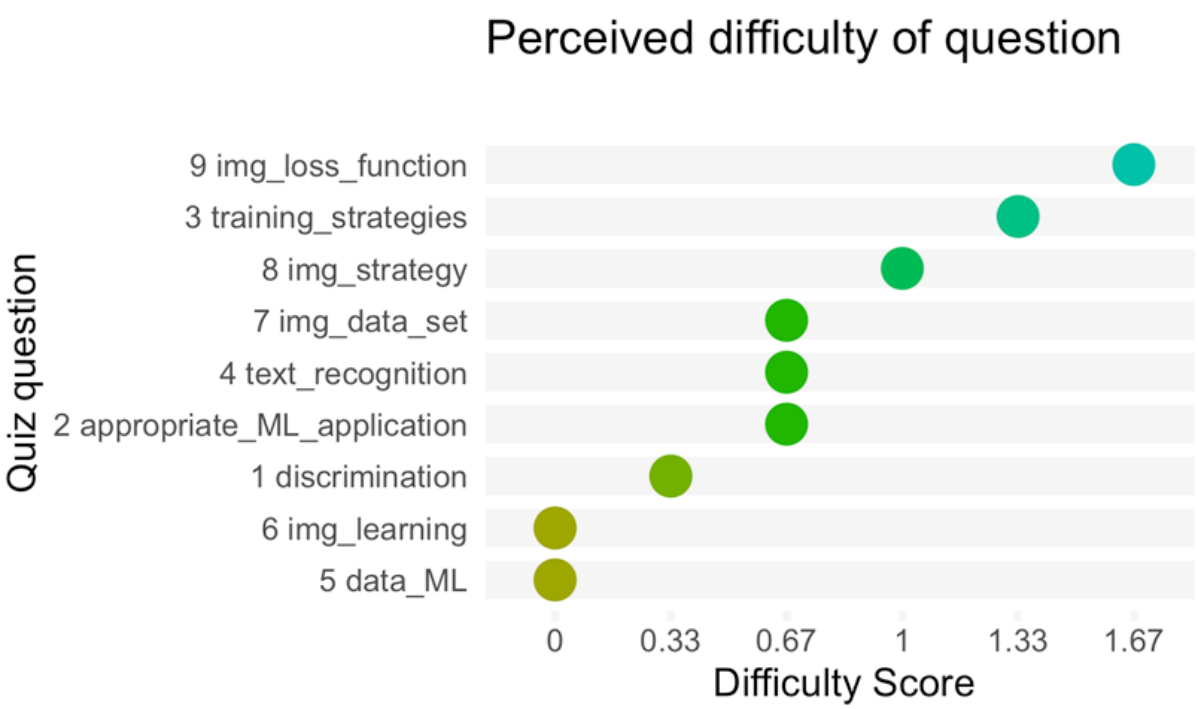


Figure 5.16: Perceived difficulty of AI knowledge test question by participants of the pre-study.

respondents with a higher self-identified AI knowledge tended to avoid choosing the option "I don't know", and rather risked to select a wrong answer. Instead, the respondent with little AI knowledge tended to select the option "I don't know" more frequently, and in contrast to the other two participants, did not select any incorrect answer.

We observed a positive association between the self-rated AI knowledge and the AI knowledge test. This association is also in line with the number of courses taken, i.e., respondents who took fewer AI courses had fewer correct answers than respondents who took more AI courses.

Overall, the test seems to reflect knowledge on AI. Compared to the self-rated AI knowledge, the AI knowledge test seems to be more objective and less influenced by personal reflections on knowledge or personal characteristics such as diffidence (e.g., one subject had 90% correct answers but indicated to only have intermediate AI knowledge).

AI-competent Dataset

The AI knowledge test was included in the questionnaire when surveying the AI-competent sample. Figure 5.17 presents the relationships between self-rated AI knowledge, the number of questions in the AI knowledge test answered correctly, and the number of technical courses on AI taken. Figure 5.17 illustrates that the number of courses taken also influenced self-perception. Participants who attended many courses rated their level of knowledge on

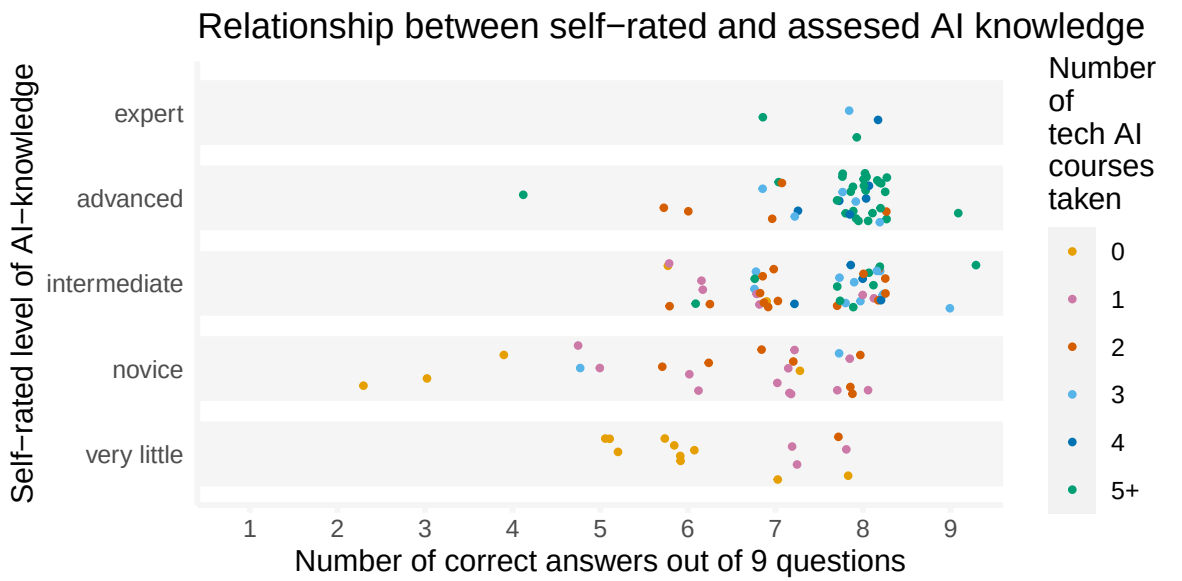


Figure 5.17: Knowledge representation based on different measures. The 'number of correct answers' is based on the AI knowledge quiz included in the survey. Participants who did not answer the manipulation check correctly and who consulted external help are not included in the plot. N = 122.

Relationship between self-rated, identified AI knowledge and number of technical AI courses taken.

average higher than participants who attended fewer courses focusing on technical AI.

Correlations found supported these observations: In order to assess the relationship between the above described AI Knowledge variables, we computed Spearman's rank correlation[23] (not all of the variables were normally distributed). There was a weak positive correlation between the number of correct answers in the AI knowledge test and the self-rated AI knowledge level, r_S=.37, p<.001. There was a moderate positive correlation between the number of correct answers and the number of courses taken on technical AI, r_S=.57, p<.001. There was a strong positive correlation between the self-rated AI knowledge level and the number of courses taken, r_S=.72, p<.001. For this subject pool, we defined participants to be AI-competent when they had correctly answered at least six out of nine questions.

Data Cleaning

The AI-competent data sample was cleaned based on the criteria listed in Table 5.13. Participants who had indicated to have consulted external help for the AI knowledge test were removed from the dataset.

[23]Spearman's rank correlation rho (absolute correlation values): 0-.19: very weak, 20-.39: weak, .40-.59: moderate, .60-.79: strong, .80-1.0: very strong

Table 5.13: Data Cleaning Criteria

	removed cases	N
Original N		160
< 18 years	0	160
Attention check AD	7	153
Attention check HR	14	139
Duration < 120 seconds	0	139
External help	7	132
Low knowledge quiz score	10	122
Final N		**122**

Analysis of Inference Ratings

Welch Two Sample t-test

The Welch two-sample t-tests produced the following results for the AD context. **Gender** ($mean_{\text{AI-competent}}$ =2.92, $mean_{\text{laypeople}}$ =2.43): t_{Welch} (122.85) =1.72, p>.05, $p_{\text{Bonf.}}$ =0.70, $\hat{g}_{\text{Hedges}}$ =0.30, $\text{CI}_{95\%}$ [-0.05,0.65]; **Emotion expression** ($mean_{\text{AI-competent}}$ =2.75, $mean_{\text{laypeople}}$ =3.13): t_{Welch} (120.99) =-1.28, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.23, $\text{CI}_{95\%}$ [-0.58,0.12]; **Wearing glasses** ($mean_{\text{AI-competent}}$ =1.67, $mean_{\text{laypeople}}$ =1.57): t_{Welch} (119.85) =0.50, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.09, $\text{CI}_{95\%}$ [-0.26,0.44]; **Skin color** ($mean_{\text{AI-competent}}$ =3.22, $mean_{\text{laypeople}}$ =2.65): t_{Welch} (122.67) =1.64, p>.05, $p_{\text{Bonf.}}$ =0.83, $\hat{g}_{\text{Hedges}}$ =0.29, $\text{CI}_{95\%}$ [-0.06,0.64]; **Intelligent** ($mean_{\text{AI-competent}}$ =5.60, $mean_{\text{laypeople}}$ =5.58): t_{Welch} (122.38) =0.06, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.01, $\text{CI}_{95\%}$ [-0.34,0.36]; **Trustworthy** ($mean_{\text{AI-competent}}$ =5.88, $mean_{\text{laypeople}}$ =5.68): t_{Welch} (121.95) =0.82, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.15, $\text{CI}_{95\%}$ [-0.21,0.50]; **Assertive** ($mean_{\text{AI-competent}}$ =4.53, $mean_{\text{laypeople}}$ =5.08): t_{Welch} (120.23) =-1.79, p>.05, $p_{\text{Bonf.}}$ =0.61, $\hat{g}_{\text{Hedges}}$ =-0.32, $\text{CI}_{95\%}$ [-0.68,0.04]; **Likable** ($mean_{\text{AI-competent}}$ =4.73, $mean_{\text{laypeople}}$ =5.20): t_{Welch} (120.99) =-1.45, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.26, $\text{CI}_{95\%}$ [-0.62,0.10].

The Welch two-sample t-tests produced the following results for the HR context. **Gender** ($mean_{\text{AI-competent}}$ =3.93, $mean_{\text{laypeople}}$ =3.93): t_{Welch} (107.18) =-0.02, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.00, $\text{CI}_{95\%}$ [-0.37,0.37]; **Emotion expression** ($mean_{\text{AI-competent}}$ =3.35, $mean_{\text{laypeople}}$ =3.85): t_{Welch} (113.74) =-1.47, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.27, $\text{CI}_{95\%}$ [-0.64,0.01]; **Wearing glasses** ($mean_{\text{AI-competent}}$ =2.53, $mean_{\text{laypeople}}$ =3.31): t_{Welch} (112.77) =-1.84, p>.05, $p_{\text{Bonf.}}$ =0.55, $\hat{g}_{\text{Hedges}}$ =-0.34, $\text{CI}_{95\%}$ [-0.70,0.03]; **Skin color** ($mean_{\text{AI-competent}}$ =3.98, $mean_{\text{laypeople}}$ =4.10): t_{Welch} (108.79) =-0.27, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.05, $\text{CI}_{95\%}$ [-0.42,0.32]; **Intelligent** ($mean_{\text{AI-competent}}$ =6.09, $mean_{\text{laypeople}}$ =5.61): t_{Welch} (111.91) =1.63, p>.05, $p_{\text{Bonf.}}$ =0.85, $\hat{g}_{\text{Hedges}}$ =0.30, $\text{CI}_{95\%}$ [-0.07,0.66]; **Trustworthy** ($mean_{\text{AI-competent}}$ =5.93, $mean_{\text{laypeople}}$ =5.20): t_{Welch} (103.70) =2.30, p=.02, $p_{\text{Bonf.}}$ =0.18, $\hat{g}_{\text{Hedges}}$ =0.42, $\text{CI}_{95\%}$ [0.05,0.79]; **Assertive** ($mean_{\text{AI-competent}}$ =5.20, $mean_{\text{laypeople}}$ =5.00): t_{Welch} (112.37) =0.63, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.12, $\text{CI}_{95\%}$ [-0.25,0.48]; **Likable** ($mean_{\text{AI-competent}}$ =5.40, $mean_{\text{laypeople}}$ =5.26): t_{Welch} (112.72) =0.44, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.08, $\text{CI}_{95\%}$ [-0.29,0.45].

Exploratory Factor Analysis

We conducted an exploratory factor analysis on the eight inferences (items) with orthogonal rotation (varimax) for each of the three samples. For the analysis, cases with missing values, i.e., "Can't Answer" responses, were removed from all three samples, which reduced the sample size for the laypeople sample N=118, for the AI-competent sample to N=112, and for laypeople validation sample to N=91. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis for the laypeople (MT) sample KMO = 0.76, for the AI-competent sample KMO = 0.75, and for the laypeople validation sample KMO = 0.68. All KMO values for individual inferences were ≥ 0.70 for laypeople (MT) sample, ≥ 0.68 for the AI-competent sample, and ≥ 0.64 for the laypeople validation sample. Hence, all values were above the acceptable limit of 0.5 [69, 466, 420]. Bartlett's test of sphericity indicated that correlations between inferences were sufficiently large for the laypeople (MT) sample $\chi^2(28) = 283.9352$, $p < .001$, the AI-competent sample $\chi^2(28) = 227.8268$, $p < .001$, and the laypeople validation sample $\chi^2(28) = 192.1025$, $p < .001$ [69].

Multiple criteria for the identification of the number of factors to extract suggested two factors. For examples, for all three samples two factors had eigenvalues over Kaiser's criterion of 1. The scree plot, very simple structure of complexity 1, as well as the Velicer MAP all suggested two factors for all of the three samples. Given these analysis, we extracted two factors in the final analysis. For all three samples, oblique rotation resulted in factors with correlations <.32 [423], yet the same pattern structure. Hence, orthogonal rotation was chosen. Table 20 shows the factor loadings after rotation for all of the three samples separately. It should be noted that some factor loadings do not exceed .6 and our sample size is rather small [467].

We performed robustness checks with sub-samples of 85% of the data. The results from the robustness checks validate the findings from the main analysis. However, the solutions were not always stable. Some items loaded on two factors, and hence, did not achieve simple structure. This is because there are variables with loadings >.3 on more than one factor [468], e.g., *gender* or *wearing glasses*.

While small factor loadings and unstable factor solutions during the robustness check suggest that the interpretation of the factor analyses should be considered with caution, the structure of the factor loadings replicates findings from [445]. We assume that both, small factor loadings and the lack of simple structure, emerge from the small sample size.

ANOVAs for each of the inferences

AI-competent vs. MTurk Laypeople Sample

Table 5.15 to Table 5.22 present the results from the Bonferroni corrected ANOVAs for each of the eight inferences.[24]

[24]Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5.14: Exploratory factor analysis for all three samples: Varimax rotated factor loadings

	Laypeople (MT)		**AI-competent**		**Laypeople (Validation)**	
	Character and personality	Social constructs and features	Character and personality	Social constructs and features	Character and personality	Social constructs and features
gender	-0.01	0.53	0.36	0.67	0.07	0.53
emotion expression	0.15	0.53	0.18	0.49	0.24	0.42
wearing glasses	-0.29	0.75	-0.09	0.64	0	0.76
skin color	-0.13	0.8	0.02	0.68	-0.05	0.83
intelligent	0.69	-0.07	0.6	0.05	0.7	-0.1
trustworthy	0.74	-0.15	0.8	-0.14	0.8	-0.02
assertive	0.72	0.08	0.7	0.2	0.64	0.16
likable	0.69	-0.05	0.56	0.22	0.53	0.22
Eigenvalues	2.17	1.81	1.98	1.67	1.88	1.82
% of variance	0.27	0.23	0.25	0.21	0.23	0.23
α	0.81	0.75	0.75	0.71	0.75	0.73

Table 5.15: ANOVA for inference: gender

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	82.63	1	23.60	0.000	0.000	***
gender	6.57	2	0.94	0.393	1.000	
age	16.28	5	0.93	0.463	1.000	
education	15.27	7	0.62	0.736	1.000	
country	105.05	27	1.11	0.330	1.000	
sample	0.43	1	0.12	0.725	1.000	
Residuals	689.67	197				

Using Pillai's trace, there were significant main effects at an α-level of 0.05 for *context* on the inference ratings for gender, emotion expression, wearing glasses and skin color. There were no other significant effects.

AI-competent vs. Validation Laypeople Sample

Table 5.23 to Table 5.30 present the results from the Bonferroni corrected validation ANOVAs for each of the eight inferences. For this comparison, we were able to include participant's information on whether they have a job (no; yes, IT-related; yes, not IT related).

Using Pillai's trace, there were significant main effects at an α-level of 0.05 for *context* on the inference ratings for gender, wearing glasses and skin color, but not for emotion expression. There were no other significant effects.

Table 5.16: ANOVA for inference: emotion expression

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	28.22	1	9.34	0.003	0.015	*
gender	2.25	2	0.37	0.689	1.000	
age	15.40	5	1.02	0.407	1.000	
education	47.21	7	2.23	0.033	0.199	
country	70.14	27	0.86	0.669	1.000	
sample	1.24	1	0.41	0.523	1.000	
Residuals	598.40	198				

Table 5.17: ANOVA for inference: wearing glasses

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	85.37	1	26.31	0.000	0.000	***
gender	1.43	2	0.22	0.803	1.000	
age	5.65	5	0.35	0.883	1.000	
education	5.64	7	0.25	0.972	1.000	
country	99.93	27	1.14	0.297	1.000	
sample	0.74	1	0.23	0.633	1.000	
Residuals	645.73	199				

Table 5.18: ANOVA for inference: skin color

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	68.18	1	14.17	0.000	0.001	**
gender	4.36	2	0.45	0.637	1.000	
age	16.34	5	0.68	0.640	1.000	
education	17.33	7	0.51	0.823	1.000	
country	112.15	27	0.86	0.664	1.000	
sample	0.00	1	0.00	0.984	1.000	
Residuals	933.58	194				

Table 5.19: ANOVA for inference: intelligent

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	5.41	1	2.11	0.148	0.889
gender	0.01	2	0.00	0.999	1.000
age	28.72	5	2.24	0.052	0.311
education	9.67	7	0.54	0.805	1.000
country	94.04	26	1.41	0.099	0.593
sample	4.11	1	1.60	0.207	1.000
Residuals	507.89	198			

Table 5.20: ANOVA for inference: trustworthy

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	2.07	1	0.86	0.356	1.000
gender	4.23	2	0.87	0.419	1.000
age	15.44	5	1.28	0.275	1.000
education	5.93	7	0.35	0.929	1.000
country	56.01	25	0.93	0.568	1.000
sample	0.61	1	0.25	0.616	1.000
Residuals	478.66	198			

Table 5.21: ANOVA for inference: assertive

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	3.76	1	1.24	0.266	1.000
gender	6.14	2	1.01	0.364	1.000
age	9.43	5	0.62	0.682	1.000
education	14.97	7	0.71	0.666	1.000
country	77.99	25	1.03	0.429	1.000
sample	14.54	1	4.80	0.030	0.177
Residuals	593.24	196			

Table 5.22: ANOVA for inference: likable

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	7.75	1	2.42	0.121	0.728
gender	3.75	2	0.59	0.558	1.000
age	17.35	5	1.08	0.371	1.000
education	20.61	7	0.92	0.492	1.000
country	48.06	26	0.58	0.951	1.000
sample	4.33	1	1.35	0.246	1.000
Residuals	627.60	196			

Table 5.23: Validation ANOVA for inference: gender

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	70.24	1	19.43	0.000	0.000	***
gender	1.61	2	0.22	0.800	1.000	
age	11.78	4	0.81	0.518	1.000	
education	11.27	6	0.52	0.793	1.000	
country	158.26	35	1.25	0.177	1.000	
student job	7.36	2	1.02	0.364	1.000	
sample	0.25	1	0.07	0.794	1.000	
Residuals	611.07	169				

Table 5.24: Validation ANOVA for inference: emotion expression

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	3.70	1	1.37	0.244	1.000
gender	0.87	2	0.16	0.851	1.000
age	20.24	4	1.87	0.118	0.824
education	34.05	6	2.10	0.056	0.390
country	127.58	35	1.35	0.110	0.767
student job	1.63	2	0.30	0.740	1.000
sample	0.01	1	0.00	0.959	1.000
Residuals	454.11	168			

Table 5.25: Validation ANOVA for inference: wearing glasses

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	30.60	1	8.59	0.004	0.027	*
gender	2.16	2	0.30	0.738	1.000	
age	14.38	4	1.01	0.404	1.000	
education	3.84	6	0.18	0.982	1.000	
country	93.19	35	0.75	0.844	1.000	
student job	0.89	2	0.13	0.882	1.000	
sample	4.58	1	1.29	0.258	1.000	
Residuals	601.86	169				

Table 5.26: Validation ANOVA for inference: skin color

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	38.58	1	7.80	0.006	0.041	*
gender	13.63	2	1.38	0.255	1.000	
age	11.97	4	0.61	0.659	1.000	
education	13.69	6	0.46	0.836	1.000	
country	178.70	34	1.06	0.386	1.000	
student job	1.61	2	0.16	0.850	1.000	
sample	0.01	1	0.00	0.971	1.000	
Residuals	825.81	167				

Table 5.27: Validation ANOVA for inference: intelligent

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	14.11	1	5.86	0.017	0.116
gender	0.36	2	0.07	0.928	1.000
age	9.09	4	0.94	0.441	1.000
education	3.78	6	0.26	0.954	1.000
country	90.77	34	1.11	0.327	1.000
student job	6.16	2	1.28	0.281	1.000
sample	5.18	1	2.15	0.144	1.000
Residuals	409.63	170			

Table 5.28: Validation ANOVA for inference: trustworthy

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	0.00	1	0.00	0.962	1.000
gender	4.00	2	1.02	0.362	1.000
age	2.36	4	0.30	0.877	1.000
education	18.30	6	1.56	0.162	1.000
country	60.96	33	0.94	0.560	1.000
student job	9.12	2	2.33	0.100	0.703
sample	0.85	1	0.43	0.511	1.000
Residuals	330.78	169			

Table 5.29: Validation ANOVA for inference: assertive

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	9.05	1	3.31	0.071	0.496
gender	5.93	2	1.08	0.341	1.000
age	2.87	4	0.26	0.902	1.000
education	17.37	6	1.06	0.390	1.000
country	104.36	33	1.16	0.273	1.000
student job	17.97	2	3.28	0.040	0.280
sample	8.13	1	2.97	0.087	0.607
Residuals	459.79	168			

Table 5.30: Validation ANOVA for inference: likable

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	11.75	1	3.50	0.063	0.443
gender	2.54	2	0.38	0.686	1.000
age	5.22	4	0.39	0.817	1.000
education	43.68	6	2.17	0.049	0.341
country	63.58	34	0.56	0.977	1.000
student job	5.70	2	0.85	0.430	1.000
sample	4.93	1	1.46	0.228	1.000
Residuals	571.63	170			

Code Book

This code book was provided to all researchers who were involved in the process of labeling the datasets. It provides information on the context of the data and guidelines on how to label the data. Figure 5.15 was added for better understanding of the survey scenarios.

General Notes and Background Information on the Study

Study Description

The comments to be categorized originate from surveys on the perception of AI inference-making in the context of advertisement and in the context of hiring. After rating how much a participant agrees or disagrees with a certain inference made by a software application using AI, the participant was asked to justify his/her rating in one to two sentences. In total, the participant was asked to repeat this process for eight different inferences: *gender*, *skin color*, *wearing glasses*, *emotion expression*, *intelligent*, *trustworthy*, *assertive*, *likable*.

Experimental Set-up

One participant was either drawn into the context of advertisement (AD) or into the context of hiring (HR). Figure 5.15 contains examples of two scenarios shown to two different participants (one drawn into the AD context and the other drawn into the HR context).

Coding Instructions

Case-wise analysis

The answers to the open questions are analyzed case-wise, i.e., one respondent at a time. Given partially very little text per answer and occasional references to previous answers, a case-wise coding of all answers per participant ensures the preservation of participant-based contextual information.

Scope of material

The **unit of evaluation** corresponds to all justification texts by respondents in the samples. The justification texts are answers to eight brief open questions in the survey.

The **unit of context** determines the material that can be consulted for coding. In this study a participant may reference a previous answer; hence, the context unit equals all responses from one participant.

The **unit of coding** resembles the minimal textual element that can be assigned to one category; here, parts of one sentence of a response from one study participant.

Repeated information and multiple codes per justification

Multiple codes can be assigned to one justification of a participant. This approach allows accounting for complex justification patterns, where participants discuss different topics

within one comment. The rating is to be considered when assigning a code, because it usually helps understanding the justification better.

Missing responses

Some respondents did not justify their rating ("NA") or wrote, e.g., "None". In these cases, the theme "no justification" is assigned.

Categories

The following Table 5.31 summarizes all categories, gives definitions as well as application descriptions and differentiates the categories from related ones. Examples are provided.

Table 5.31: Code Book: Definition of categories and examples.

Category	Description	Example
AI (in)ability		
technical ability of AI	**Definition:** The AI has the technical ability to draw an inference. **Application:** This code is applied when a participant gives a specific explanation to why s/he believes the AI to be able to draw the inference.	*Obviously an AI can identify the shade of skin* | *Machine learning can be used to determine whether or not the person is expressing an emotion.*
works well/ accurately	**Definition:** The AI accurately draws a certain inference. This task is known to work well. **Application:** This code is used when a participant highlights high accuracy scores for a specific task or mentions that a specific task is regularly and successfully solved by the AI system.	*Yes, would work, as it's already being done* | *emotion regonition based on facial expression is a very popular AI Task [sic]* | *solid results can be achieved by an AI*
easy to infer	**Definition:** The inference task is easy to solve for an AI system. **Application:** This code is used when a participant highlights that a certain inference can easily be drawn by an AI system.	*This is something that should be easy for an AI to determine.* | *the gender of a person based on a picture is in itself a relative easy classification task*
can infer most/sometimes/ difficult in some situations	**Definition:** The AI systems can most times or sometimes draw the inference. However, for specific cases, such as the gender "other", a correct inference is difficult or accompanied by (many) mistakes. **Application:** This code is assigned when a participants highlights that the system will make mistakes for some cases, e.g, for "other".	*except for very small amount of situations like trans, gender can be analysed easily by ai* | *The vast majority of men and women have features that make their gender clearly identifiable. However, gender-neutral persons or people who simply don't look like or conform to a gender would be difficult* | *it won't be perfect as people express emotions differently*
difficult/ not possible to infer	**Definition:** The inference task is difficult or impossible to be solved for an AI system. **Application:** This code is used when a participant highlights that a certain inference is impossible or difficult for an AI system.	*Impossible to infer* | *Too complex a notion to quantify, even for humans* | *Hard for AI to decide.*
Inference task		
inference is objective	**Definition:** The inference task is objective. **Application:** This code is used when a participant highlights that the evaluation of the inference is not dependent on the observer or specifically uses the word "objective".	*This can also be easily and objectively answered by an ML algorithm.* | *can be determined objectively*
inference is subjective	**Definition:** The inference task is subjective. **Application:** This code is used when a participant highlights that the evaluation of the inference is dependent on the observer or specifically uses the word "subjective".	*trustworthy is a subjective trait* | *An image can not show whether a person is likeable or not. Likeability is largely subjective and can not be judged in objective terms.*

Category	Description	Example
Reference to data		
indicative distinct facial/visual features	**Definition:** Distinct facial tendencies or features indicate a certain inference. **Application:** The code is used when a participant highlights certain facial properties as key for drawing a correct inference.	*Learning to recognize the traits that show specific emotions is possible* | *I supposed a computer can be programmed to detect certain facial tendencies that express emotions*
profile picture good evidence (data)	**Definition:** The profile picture provides good evidence for an AI system to draw a certain inference. **Application:** The code is used when a participant highlights that the inference can be drawn based on the provided data, here, the profile picture, or comments that a certain inference can be "seen".	*for the most part it's fairly obvious to see if someone is expressing anger, hate, love, etc.* | *visually can be determined by the AI* | *It seems reasonable that an AI could figure out whether someone is wearing glasses by their picture.*
data quality/ variety of high relevance	**Definition:** Data quality, including a varied dataset, is of high importance to train an AI system to draw specific inferences. **Application:** The code is used when a participant highlights that the success of the AI system is based on the quality of the data. This code is also used when a participant mentions that the data can be manipulated in such a way that it is difficult to draw correct inferences.	*diversity on the dataset would be required to make sure no skin tone under different lighting is left out* | *If there's valid, reliable data to support it, it's reasonable* | *If the data is valid and reliable, it's reasonable.*
not sufficient/ good evidence (data) for task	**Definition:** The profile picture does not provide sufficient or good evidence for an AI system to draw a certain inference. **Application:** The code is used when a participant highlights that further data or different data would be required to properly draw the inference, e.g., because the image only captures a single moment. This code is also used when it is mentioned that facial expressions do not resemble how a person actually feels or what they identify with.	*A personality cannot be inferred from facial traits. It is inferred by actions, which cannot be shown in a profile pic* | *There are numerous of people that tend to hid their emotions through pictures and everyday lifestyle but end up taking their own lives. Nothing looks like it seems.* [sic] | *[it's no] real indicator whether they are nice.*
Reference to (ir)relevance of inference for purpose of AI system		
inference relevant	**Definition:** The inference is relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Application:** The code is used when a participant highlights that drawing an inference is useful/helpful/relevant to the purpose of the AI system.	*I agree that different genders need different products and services, so this would be reasonable* | *There are products that target just men or just women so i can see this being helpful*
inference sometimes relevant	**Definition:** The inference is (only) sometimes relevant to the decision of the AI, e.g. advertisement choice or applicant selection. **Application:** The code is used when a participant highlights that drawing an inference is not always, but only sometimes, useful/helpful/relevant to the purpose of the AI system.	*There are cases where a certain gender could be preferable to another (babysitters, private tutors), but there are also cases where this distinction does not matter (corporate jobs, waiters, sellers).*
inference not relevant	**Definition:** The inference is not relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Application:** The code is used when a subject highlights that drawing the inference is not useful/helpful/relevant or does not have anything to do with the purpose of the AI system.	*does not seem like a valueable information.* | *skin color doesn't influence a persons consumer behavior* [sic]

Category	Description	Example
Ethics and Norms		
ethically questionable/ should not matter/ should not be used	**Definition:** Drawing the inference is ethically questionable or should not matter. An inference should not be used to make subsequent decisions. **Application:** The code is used when a participant highlights or critiques drawing a specific inference. Some participants stress that an inference should not matter for a decision made by an AI or that an inference, even when drawn, should not further be used in subsequent AI decision-making.	*It should not matter for the job* | *I don't think that the AI should consider race or skin color when deciding what advertisements to show people* | *I think this is a bit too touchy of a subject due to being politically correct is very important at this time on history*
bias/ stereotypes/ discrimination	**Definition:** Drawing the specific inference leads to bias or discrimination. Making decisions based on the inference is based on stereotypes. **Application:** The code is used when a participant highlights or critiques drawing specific inferences because the resulting AI decision-making would be biased, be based on stereotypes or discriminate.	*I think this can be racist.* | *This will end badly, if white people are more likable than black people. Won't to that.* [sic]
binary gender system not appropriate	**Definition:** A binary concept of the inference gender is not appropriate and does not reflect today's society. **Application:** The code is used when a participant highlights or critiques drawing the inference gender based only on two categories, females and males.	*Gender norms are a thing of past!* | *gender is a fluid concept* | *many people do not identify with their sex and birth or with a binary gender system which may lead to incorrect classification*
Comparison to human		
easy for human to identify	**Definition:** The inference task is easy to solve for a human. **Application:** This code is used when a participant makes a comparison to a human being and highlights that a certain inference can easily be drawn by a human.	*This is an easy task even for a human*
difficult for human to identify	**Definition:** The inference task is difficult to solve for a human. **Application:** This code is used when a participant makes a comparison to a human being and highlights that a certain inference is difficult to be drawn by a human.	*Disagree, because its also hard for humans to guess that from experience* | *there are some difficulties (androgynous), which even humans have problems with*
Miscellaneous		
person not sure/ indecisive	**Definition:** A respondent is indecisive whether to agree or disagree and/or does not have any opinion. **Application:** This category is assigned if a respondent is unsure. In such cases the respondent gave a rating "4", i.e. neither agree nor disagree. Occasionally, the text-field is left empty.	*What is skin color? Do you mean race? What are the categories?* | *Not sure if a single picture can determine the assertiveness of a person. People generally put a brave face on social media.*
no justification	**Definition:** A respondent did not provide an open-text response.	*NA* | *none*

Figure 5.18: Percentages of justification themes used by participants for each inference (number of participants per context and inference as baseline). Each column adds up to more than 100%, because participants used up to four themes in one justification.

Analysis of Justification Themes

Comparison of Usage of Justification Themes

Co-occurrence Analysis: AI (in)ability and data-related themes are the themes most frequently used in combination with other themes.

We analyzed the co-occurrence of themes with each other to identify patterns in the use of multiple justification themes. Figure 5.19a) depicts the frequencies of two themes used in combination. Please note that this analysis includes only the AI-competent and laypeople (MT) sample, and refers to all justifications containing two or more themes, i.e. 20% to 35% of justifications (see Table 1 in main article). Figure 5.19b) illustrates networks of the co-occurrences of themes by sample and by inference type.

For inferences on 'constructs and features', the AI-competent raise concerns but recognize AI to be able to make certain inferences.

Referring to inferences on *constructs and features*, people with AI-competence raise "ethical and discriminatory concerns" in combination with almost all other justification themes, however, most frequently in combinations with themes on "AI ability" (11.5%) or the "sufficiency" of the profile picture as evidence (10%; Figure 5.19a) and 5.19b-1)). The "sufficiency" of the profile picture as evidence is also often mentioned in combinations with themes on AI ability (10%).

This relationship changes for justifications on the inferences on *character and personality traits*. "Ethical and discriminatory concerns" are most frequently brought forward in combination with themes on the "insufficiency" of a profile picture as evidence (11.4%; Figure 5.19a) and 5.19b-3)).

For inferences on 'character and personality traits', laypeople often pair comments on the (in)sufficiency or (in)adequacy of the data with another theme.

Referring to inferences on *constructs and features*, laypeople highlight the "sufficiency" of the data in combination with comments on the "ability" (17.2%) and "inability" of AI (15.2%) (Figure 5.19a) and 5.19b-2)). With reference to the inferences on *character and personality traits*, laypeople frequently mention themes related to "insufficiency" of the data in combination with "inability" of AI to make such an inference (12.5%), "ethical and discriminatory concerns" (15.5%), and "comparison(s) to human" abilities (12.5%). For instance, a comment on the inference *assertive* states: "I can't see how even a person could determine this from a picture." However, the "insufficiency" of the data is also frequently mentioned in combination with the "relevance" of the inference (10.6%), e.g., a comment on the inference *intelligent* states: "You want to hire smart people but i dont think that can be analyzed from a photo" [sic] (Figure 5.19a) and b-4)).

More theme combinations are used to explain ratings on inferences referring to 'constructs and features' and by individuals with AI-competence.

Generally, a greater variety of combinations are used to justify ratings on inference ratings referring to *constructs and features*. This applies to both main samples (see Figure 5.19a), 5.19b-

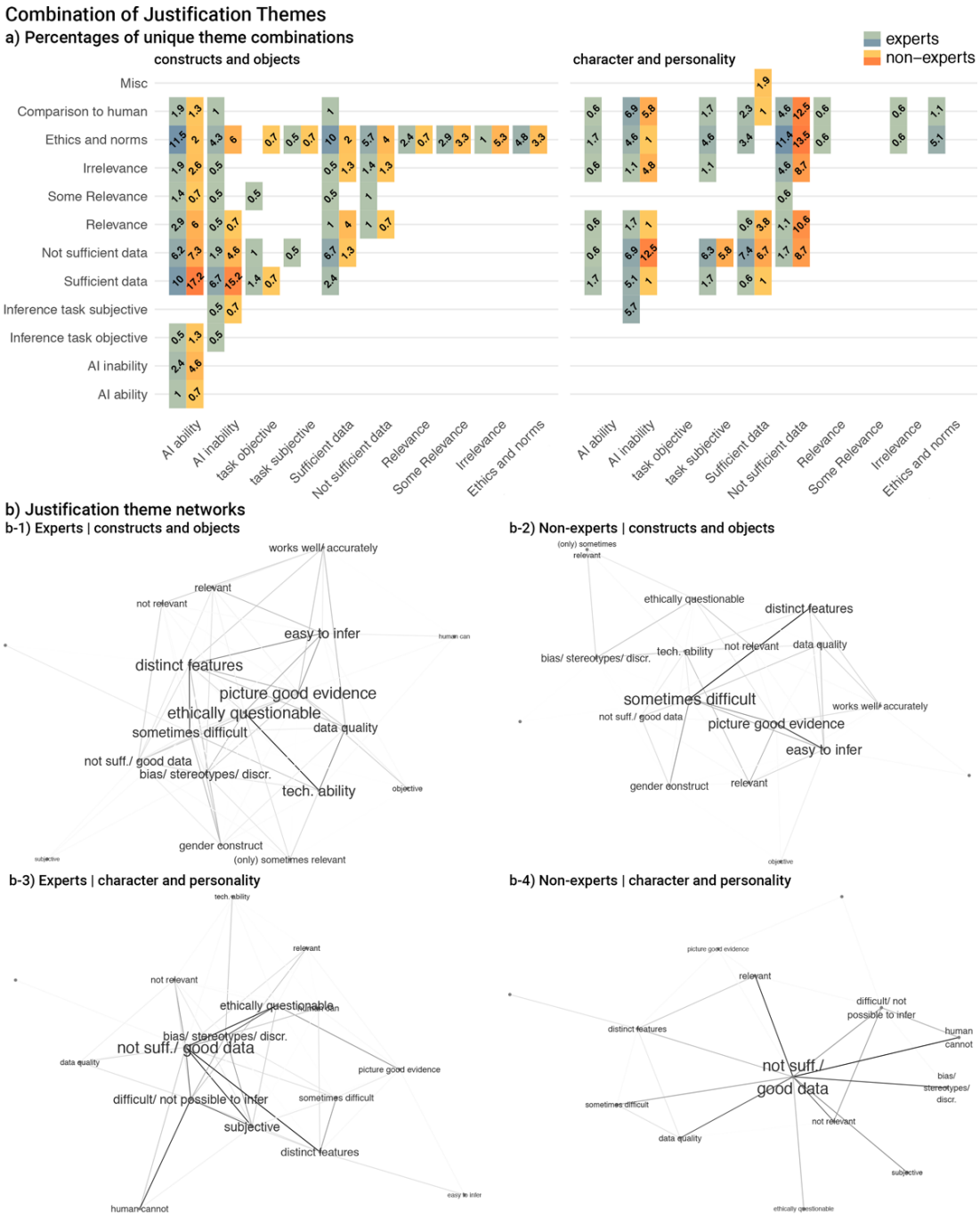


Figure 5.19: Combination of justification themes by inference type and sample. a) Unique combinations of two themes (i.e. super-ordinate theme topic) by inference type and sample. Analysis refers to justifications containing more than one theme. b) Network analysis of co-occurrences of themes. We calculated undirected weighted one-mode networks.

1) and b-2)). Figure 5.19a), 5.19b-3) and b-4) show a smaller variety of theme combinations for justifications on inference ratings referring to *character and personality traits*. This implies that opinions on inferences referring to *character and personality traits* are clearer. In contrast, opinions on inferences referring to *constructs and features* are more varied. People with AI competence use a greater variety of theme combinations than laypeople (MT) for both types of combinations (Figure 5.19a), 5.19b-1) and 5.19b-3)).

6 Discussion & Final Remarks

6.1 Discussion

Our research on the Chinese SCS, on the procedural dimensions of social media classification, as well as on facial analysis AI demonstrates the multidimensional nature of digital classification systems. In the case of the Chinese SCS, a techno-autocratic government motivates and realizes a nationwide digital classification system with the promise to solve society's moral, social, and legal challenges. Social media classifications promise marketers a classification market that facilitates the delivery of advertisement to fine-grained audiences consisting of billions of users. Facial analysis AI has been translated into numerous commercial applications. However, it is largely driven by reference to a body of research that assumes facial impressions to be epistemically valid despite overwhelming evidence that has falsified such a view.

The purpose of this research has been to advance our understanding on the legitimacy of different digital socio-technical classification systems. I have approached this sizable territory from three perspectives. The prime motivation to conduct research on the Chinese SCS was the fact that very little was known about its actual implementation. In light of contradictory media reports and a body of theoretical accounts on the Chinese SCS, we observed that the technological realization of the listing infrastructure remained poorly understood. Decisive for our research on social media classifications was the fact that most studies had paid attention to the consequences of social media classifications largely neglecting the procedural legitimacy of such classifications. Finally, in an increasingly *visual* data culture, we hoped to contribute to discussions around ethical computer vision AI by conducting participatory studies with non-experts and people with AI-competence.

In the following section, I will summarize what I believe are the most important takeaways from these three research perspectives and offer final remarks.

6.2 Understanding the legitimacy of the Chinese SCS

Summary of key takeaways

China's unique policy-making procedures: Central planning, de-centralized implementation. It is important to understand how techno-autocracies, in comparison to techno-democracies, use digital socio-technical classification systems to accomplish long-term policy goals. This is particularly important for the Chinese SCS given China's geopolitical power and the fact that the SCS shapes the behavior of about 1.4 billion Chinese citizens and all

companies doing business in China. There are multiple different SCSs in China. Apart from commercial SCSs and SCS experiments in cities, the government-led SCS reputational blacklist and redlist infrastructure operates across all provinces, municipalities, and administrative divisions.

The SCS is planned by the Chinese Communist Party's (CCP) top officials while provincial and municipal governments are in charge of implementation. This necessarily results in a diverse realization of the central government's policy-plans that are kept deliberately vague. Each SCS platform operates a different web server with its own front-end, back-end and database design. Moreover, the de-centralized implementation of the SCS leads regional governments to focus on those sanctions and norms that are most relevant to their region. Understanding the impact that the Chinese SCS exerts on Chinese society, now and in the future, means taking into account its unique policy-making cycles as well as a comprehensive analysis of the different SCS implementations across the entire country.

Current focus: economic sanctions and rewards. But new list types can be set up quickly. The blacklist and redlist infrastructure represents the government's efforts to govern society through mechanisms that go beyond common legal and regulatory practices. Being listed sends a strong reputational signal, a verification of "goodness" or "badness" by the government itself in a one-party, authoritarian state. Moreover, redlisted and blacklisted entities receive a variety of material rewards or punishments realized through the "joint punishment and reward mechanism". Citizens on blacklists are excluded from a variety of services and goods, for example, they cannot purchase train or plane tickets. Redlisted entities receive a variety of rewards such as fast tracks for visa applications or free entrance to museums.

The current SCS implementation focuses on entities that fail to fulfil a court order issued due to a variety of economic transgressions. In our exploratory study, all administrative divisions had implemented a so-called "List of Dishonest Persons subject to Enforcement" also called the "Lao Lai" blacklist. However, we also identified lists that rewarded moral behaviors whereby morally virtuous behaviors always meant working towards the Party's social and political goals. When we conducted the Coronavirus case study in February 2020, we found that the listing infrastructure is dynamic and flexible, which makes the Chinese SCS a powerful regulator of novel transgressions: lists can be set up whenever central or regional governments see the need to sanction or reward novel behaviors.

Virtue ethics principles support the legitimization of the Chinese SCS. We found that the Chinese government publishes narratives on "blameworthy" and "praiseworthy" role model citizens on its central SCS platform creditchina.gov.cn. In Chinese ethics, the narrative format plays an important role in propagating ethical principles *represented* by the thoughts, emotions, and behaviors of individuals in ethically-charged scenarios. Revealing the "goodness" or "badness" of characters through narratives is a virtue ethics approach particularly prevalent in China. Besides deontology and consequentialism, virtue ethics is the third major branch of normative ethics [473]. It is concerned with the development of a moral character that *acts*

virtuous emphasizing the importance of practical morality in virtue ethics. "Blameworthy" and "praiseworthy" SCS role model narratives showcase how individuals act morally wrong or morally right with respect to the overall *economic* goals of the SCS. Given China's one-party authoritarian government and unique economic system means that the SCS cannot be legitimized by the promises of economic growth and development attributable to a liberal market economy. Instead, its legitimization is supported by the mapping of economically desirable behaviors to the virtue ethical principles of honesty and harmony that play an important role in Confucianism.

Discussing the legitimacy of the Chinese SCS

We know that regional governments implement the SCS listing infrastructure according to the types of transgressions that are most prevalent in their respective areas of governance. The goal of the regional implementation, besides the fact that it follows Chinese policy-making more generally, is to strengthen the administrative and legal enforcement of economic sanctions. But, at the same time, it increases the oversight and control of the CCP's Central Committee over regional governments. First, the development of the listing infrastructure itself indicates how quickly and efficiently regional governments manage to establish the SCS in their province or municipality. Regional governments stand in competition to implement the CCP's policy plans. This requires not just the technical means to display credit records publicly but a technological infrastructure of data sharing between different administrative and governmental offices. Second, the amount of blacklist *records* necessarily indicates the prevalence of transgressions within an administrative region. A region with a lot of blacklist records on display signals efficient enforcement. On the other hand, there may be a point at which a regional government showcases "too many records" raising questions about the level of transgressions in that region. Thus, the listing infrastructure also serves as a reputation signal of regional governments' abilities to enforce social order. However, this positive reputational signal can flip around when regional governments list significantly more transgressions than SCS listings in other administrative regions.

I believe that it is important to reflect on the Chinese government's use of virtue ethics principles in the narratives on role model citizens. The Chinese SCS represents a technological surveillance infrastructure that enables the government to enforce social control more effectively. One the one hand, we see a technological project of digital reputation lists and, on the other hand, a folklore-like use of the narrative format to communicate what constitutes "good" and "bad" citizenship. Narratives describe the ideal of a harmonious society where citizens work for the collective good while economic dishonesty is rigorously punished thanks to the technological surveillance apparatus of the SCS. There's an SCS label for "bad" citizen "Lao Lai" that corresponds to the morally reprehensible "Xiaoren" in Confucianism (literally translated as "small man" [228]). My assertions is that one core weakness of virtue ethical principles is that they are empty concepts until an entity defines what it means to be virtuous and in relation to what activity or object. Virtue ethical principles are ethically weak when they are not embedded in a society that can engage with the semantic plurality of virtues

and vices. This weakness of virtue ethics supports the CCP's legitimization of a nationwide reputation system as there is no entity that can contest the CCP's definition of what should essentially reflect a virtuous trait or person.

6.3 Understanding the procedural legitimacy of social media classifications

Summary of key takeaways

Defining challenges of procedural legitimacy in social media classification. Our work on social media classification addresses procedural challenges inherent to user profiling. Platforms determine in a largely arbitrary manner the quantity and quality of evidence that enables the classification of user attributes for billions of social media users worldwide. Analyzing a theoretical framework of personal identity in philosophy, we argued that social media user profiling generates *formalistic self-concepts* when it assigns meaning to views, clicks, posts, relationships, or location data of social media users. We noted that philosophers generally agree that individuals have the capacity to *justify* and *control* essential elements of their self-concept. For example, to the philosopher Harry Frankfurt, a person is able to justify and control what motives they wish to be moved by. The philosopher Marya Schechtman believes that a person has the capacity to psychologically compare, organize, and relate experiences into a culturally-accepted narrative. The ability to create formalistic self-concepts makes social media platforms powerful classification systems, in part, because they do not offer any means for justification and control over formalistic self-concepts as philosophical theories of personal identity suggest.

Defining normative trade-offs in social media classification. As a consequence of our theoretical engagement, we formulated two normative trade-offs inherent to social media classification. First, should a person's formalistic self-concept accurately represent their actual self-concept at the expense of significant data collection, data processing, and data analysis of users? Second, should a person's formalistic self-concept be made transparent to users if such identity claims then influence a person's self-concept? In other words, the transparency of their formalistic self-concept could exert influence on individuals' actual self-concept thereby undermining individuals' autonomy to self-determine when they internalize "how the machine classifies" them.

Understanding how social media users evaluate normative trade-offs in social media classification. We then asked social media users how they evaluate these normative trade-offs in an empirical vignette study. We found that social media users place more value on their privacy, conceptualized as minimizing data collection and analysis, than an accurate formalistic self-concept. Moreover, social media users strongly desire to view their formalistic self-concept. Participants in our study claimed that viewing their formalistic self-concept

would not influence their actual self-concept. They stated that they would be "immune" against social media identity declarations.

Moreover, our vignette study explored how social media users relate to social media classifications more generally. We found that users believed that social media platforms can accurately infer their interests, values, and even whether they have changed as a person since using the platform. They also stated that their formalistic self-concept is unique and that others have a different formalistic self-concept. However, they claimed that their formalistic self-concept accurately represents only part of their actual self-concept. Users expressed the desire to compare their own self-concept with the social media identity declarations of their formalistic self-concept. Finally, if made transparent, subjects would correct wrong identity declarations but our data show no clear motivation whether they would manage them over time.

Discussing the procedural legitimacy of social media classifications

Overall, our study presents a novel normative framework that gives ethical shape to the normative tensions of social media user profiling. Theories of personal identity in philosophy argue that the justification of and control over one's own hermeneutic self-concept is an essential process for the development of an autonomous self-concept. To philosophers, this process *constitutes* the freedom to self-determine, on the one hand, and with it the obligation to take responsibility over our self-concept, on the other hand. In philosophical scholarship, justifying and controlling essential aspects of our self-concept is intimately tied to intrinsic values such as autonomy, freedom, and responsibility.

Our empirical work indicates that people attribute some narrative capacity to social media classifications. Our study suggests that users believe their formalistic self-concept to be unique, other users have a different formalistic self-concept. I wonder whether perceiving one's formalistic self-concept to be "unique" results from a psychological disposition to believe in the uniqueness of one's self-concept *in general* or whether it rests on the belief in the *technological capacity* to produce a "unique" formalistic self-concept. Future studies could look into this question in more detail.

In our study, participants claimed that their self-concept would not be influenced by social media classifications if they were made transparent to them. Here, too, future studies could further investigate whether the claim of "self-concept immunity" truly holds or whether people overestimate the robustness of their self-concept. To me, it seems hard to believe that being confronted with the identity declarations of one's own formalistic self-concept would not exert any influence on one's actual self-concept, particularly, if specific classifications appear repeatedly in the formalistic self-concept. Our study did not delve deeper into the mental models that people create of their own formalistic self-concept or individual social media classifications. Our account provides an initial understanding on how individuals perceive social media identity declarations leaving ample opportunity for follow-up studies.

What I see as a key takeaway from the engagement with the procedural normativity

of digital classification as well as the normativity of specific classifications (i.e., our work on facial analysis AI) is the importance of understanding the epistemic validity of digital classifications in general. Concisely defined, I mean the legitimacy of classifications that is constituted by the epistemic proportionality governing the relationship between data and inference. I will reflect on this point in more detail when discussing our findings on facial analysis AI classifications as well as in the final conclusion of this thesis.

6.4 Understanding the legitimacy of facial analysis AI

Summary of key takeaways

Visual classifications: tensions between epistemic and pragmatic legitimization. Given the semantic ambiguity of visual data, fixing the large space of interpretive possibilities to a selection of target variables is an act of classification. And this act of classification necessarily demands an ethical justification. We believed that this discussion must take note of non-experts' ethical evaluations and that these evaluations must at least complement conceptual critical analyses. In our conceptual work, we argued that classifications such as sexual orientation, mental inferences such as trustworthiness or intelligence as well as aesthetic inferences are unreasonable due to their non-falsifiability by more evidence of the same type. Confirming our conceptual proposition, in our first empirical study with non-experts only, we found that subjects believe AI should not draw inferences such as intelligence or assertiveness because of their epistemic invalidity. However, some subjects used pragmatic considerations to legitimize AI facial inferences when they believed that an AI inference is relevant for a decision's outcome. We argued that the legitimization of an inference by considering *positive outcomes* does not rationalize the underlying epistemic belief of the inference. This type of false legitimization may be applied in other computer vision systems from social robots interacting with people to mood detection systems in cars or visually-based hiring procedures.

Both non-experts and people with AI-competence tend to oppose facial analysis AI. Our research with both sample groups highlights the normative complexity behind facial AI inferences. Indeed, based on our findings, there are no "common sense" facial analysis classifications. This is particularly true for classifying gender or skin color from facial features. Here, participants in both samples voiced concerns about the ethically problematic application of a binary conceptualization of gender as well as the conflation of skin color with race. Participants also observed the problematic confusion between a person *appearing* to have a certain trait (e.g., trustworthiness, assertiveness etc.) and whether they actual have that trait. It seems to me to be a slippery slope from legitimizing the classification of "apparent" traits from visual data to using such inferences as "actual" traits.

Overall, the majority of participants with and without AI-competence tended to reject AI classifications, in particular, those common in human first impression-making. Still, when agreeing with facial analysis AI, participants in both samples overestimated the technological

capabilities of AI. In this case, the legitimacy of a classification was dependent on the "correctness of the data" shifting the problem to the level of data collection and processing thereby neglecting the classifications' basic epistemic invalidity. Finally, the complexities behind participants' justifications raise important questions regarding who has the power over the imposition of meaning in a visual data culture.

Discussing the legitimacy of facial analysis AI

There are three main points that I believe are most important for understanding the legitimization of facial analysis AI: first, taking human inference-making as a normative benchmark for AI classification. Second, trying to legitimize facial analysis AI by correlational evidence of statistical analyses and, third, the justification of visual inferences by pragmatic considerations.

Analyzing the literature on facial analysis AI, we observe that facial analysis classifications are legitimized – at their core – by taking *human* inference-making as a normative benchmark for AI classification. For example, emotion recognition technology presupposes that reading off emotional states from people's faces is possible due to the facial action units that humans use to decipher emotional states from facial expressions. The fact that humans supposedly read off emotions from people's faces makes it normative for AI classification to do the same irrespective of whether such inference are valid. In legitimizing digital classifications, it is critical to ask in what cases human inference-making is in fact normative and in what cases it is not.

Second, and adding to the first point, in the case of facial analysis AI, the scientific literature on the validity of human facial inference-making is divided, but this division is far from balanced. Only a small share of research papers provide evidence for the validity of human facial inference-making while the overwhelming share of research articles demonstrate their epistemic invalidity. Commonly, this much smaller share of research papers serves as an "epistemic reference of validity" to justify the development of facial analysis AI. In many studies that we have referred to in our research articles, researchers engage in extensive statistical testing for significant *correlations* between facial looks and a variety of aesthetic, mental, and character traits. We found that many of these papers report significant correlations without accounting for multiple comparison problems, for example, by Bonferroni correction. Taken together, studies pick the most fitting evidence as a motivation to pursue research on facial analysis AI and report the veracity of such classifications by reference to weak correlational evidence only.

Third, results from our empirical studies show that some people rationalize visual data inferences by pragmatic considerations. We argued that this is a legitimization pitfall because the normativity of a vision-based inference should not be evaluated by criteria other than epistemic evidence. For example, even if people say that inferences such as intelligence cannot be told from faces, they may still believe that they should be drawn because of their relevance for the decision outcome. In other words, those that want to draw an epistemically invalid visual inference could claim that it nonetheless should be drawn because of its beneficial

consequences. Taken together, I see a problematic negligence of the epistemic quality of digital classifications. Technical fixes to solve bias in classification procedures and outcomes typically do not address such epistemic challenges. The de-biasing of datasets does not include a negotiation of the epistemic quality of the inferences in the dataset. I believe that a much more serious engagement with the epistemic assumptions that underlie specific classifications is necessary to ensure the legitimacy of digital classification systems.

6.5 Final Remarks

I wish to end with two remarks. First, analogical to the historic setbacks in AI development referred to as "AI winters" [474], working on AI ethics in the past years has led me to believe that we may have entered an "AI classification winter". We need a more profound and sincere commitment to determine what kind of epistemic expectations we should justifiably place on digital classification. How do we define standards of epistemic validity for commercial classification markets such as social media platforms? This critically requires a conceptual advancement that, I believe, has not been produced by philosophical scholarship in the past years. Novel digital classification systems challenge established philosophical notions. For example, scholarship on the ethics of belief [475] relies too much on the notion that an epistemically valid belief is necessarily one that has considered all available evidence. Such notions are not desirable for the normativity of classifications since they lead to significant privacy problems. Here, it is necessary to understand the advantages and disadvantages of applying other conceptualizations. For example, can we legitimize the classifications of a system by adhering to a more moderate, less strict form of evidentialism? Would a justification to apply such a less strict form of evidentialism depend only on the gravity of the consequences of the classification context or are there also procedural norms that we should consider? Digital classification systems can continue to contribute to the welfare of society when they are legitimized by both epistemic and pragmatic considerations. Critically, advancing such considerations requires a commitment to understand the underlying conceptual plurality that digital classification systems necessarily reduce to the operationalization and application of a single conception.

Thus, and this is my final remark, novel digital classification systems create opportunities to advance philosophical or conceptual scholarship. Digital classification systems represent the applied version of an otherwise essentially contested concept. This is true for ethical conceptions implemented in digital classification systems. Engaging with digital classification systems means engaging with ethical concepts *in their applied form*. This allows us to better pinpoint, understand, and explain the weakness of ethical concepts when we "see" them in their applied form in digital classification systems. For example, the strength and value of virtue ethics to justifiably legitimize a digital classification system critically depends on accounting for the conceptual plurality inherent in the concept of virtues and vices. Taking a meta-level perspective, ethical concepts lose their legitimacy to justify the legitimacy of digital classification systems when there is no engagement with their own conceptual contestedness.

Finally, all of this, I think, requires a serious commitment to understand both the technological affordances of digital classification and the underlying concepts that they aim to operationalize.

7 Published Versions of Papers

In order of appearance:

- **Engelmann*, S.**, Chen*, M., Fischer, F., Kao, C. & Grossklags, J. (2019). Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

- **Engelmann, S.**, Chen, M., Dang, L., & Grossklags, J. (2021). Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness. *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- Chen, M., **Engelmann, S.**, & Grossklags, J. (2022). Ordinary people as moral heroes and foes: Digital role model narratives propagate social norms in China's Social Credit System. *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- **Engelmann, S.**, Scheibe, V., Battaglia, F., & Grossklags, J. (2022). Social media profiling continues to partake in the development of formalistic self-concepts. Social media users think so, too. *Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM AIES)*.

- **Engelmann, S.**, & Grossklags, J. (2019). Setting the Stage: Towards Principles for Reasonable Image Inferences. *Workshop on Fairness in User Modeling, Adaptation and Personalization (FairUMAP), 27th Conference on User Modeling, Adaptation and Personalization (ACM UMAP)*.

- **Engelmann*, S.**, Ullstein*, C., Papakyriakopoulos, O., & Grossklags, J. (2022). What People Think AI Should Infer from Faces. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

- Ullstein*, C., **Engelmann*, S.**, Papakyriakopoulos, O., Hohendanner, M., & Grossklags, J. (2022). AI-competent individuals and laypeople tend to oppose facial analysis AI. *Proceedings of the Second ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO).

*Denotes equal contribution.

Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior

Severin Engelmann*
engelmas@in.tum.de
Chair of Cyber Trust
Faculty of Informatics
Technical University of Munich

Mo Chen*
mo.chen@tum.de
Chair of Cyber Trust
Faculty of Informatics
Technical University of Munich

Felix Fischer+
felix.fischer@tum.de
Chair of Cyber Trust
Faculty of Informatics
Technical University of Munich

Ching-yu Kao
ching-yu.kao@aisec.fraunhofer.de
Fraunhofer Institute for Applied and Integrated Security

Jens Grossklags
jens.grossklags@in.tum.de
Chair of Cyber Trust
Faculty of Informatics
Technical University of Munich

ABSTRACT

China's Social Credit System (SCS, 社会信用体系 or shehui xinyong tixi) is expected to become the first digitally-implemented nationwide scoring system with the purpose to rate the behavior of citizens, companies, and other entities. Thereby, in the SCS, "good" behavior can result in material rewards and reputational gain while "bad" behavior can lead to exclusion from material resources and reputational loss. Crucially, for the implementation of the SCS, society must be able to distinguish between behaviors that result in reward and those that lead to sanction. In this paper, we conduct the first transparency analysis of two central administrative information platforms of the SCS to understand how the SCS currently defines "good" and "bad" behavior. We analyze 194,829 behavioral records and 942 reports on citizens' behaviors published on the official Beijing SCS website and the national SCS platform "Credit China", respectively. By applying a mixed-method approach, we demonstrate that there is a considerable asymmetry between information provided by the so-called Redlist (information on "good" behavior) and the Blacklist (information on "bad" behavior). At the current stage of the SCS implementation, the majority of explanations on blacklisted behaviors includes a detailed description of the causal relation between inadequate behavior and its sanction. On the other hand, explanations on redlisted behavior, which comprise positive norms fostering value internalization and integration, are less transparent. Finally, this first SCS transparency analysis suggests that socio-technical systems applying a scoring mechanism might use different degrees of transparency to achieve particular behavioral engineering goals.

CCS CONCEPTS

• **Social and professional topics** → *Government technology policy*; • **Information systems** → *Decision support systems*; • **Security and privacy** → *Social aspects of security and privacy*; • **Applied computing** → *Anthropology*;

KEYWORDS

Social Credit System, Socio-Technical Systems, Transparency, Behavioral Engineering.

ACM Reference Format:
Severin Engelmann*, Mo Chen*, Felix Fischer+, Ching-yu Kao, and Jens Grossklags. 2019. Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287585

1 INTRODUCTION

Moral thinking and action necessarily depend on informational resources. When an individual asks: "What is the right thing to do?", he or she essentially relies on information that renders a conclusion morally justifiable. In philosophy and anthropology, descriptive morality refers to how groups or societies negotiate codes of conduct (or norms) that are morally acceptable or unacceptable [8, 36]. As a consequence, an individual's moral accountability tends to be proportional to his or her knowledge of good and bad moral behavior underlining the epistemic character of morality [7]. In 2014, the Chinese government issued a plan for a nationwide digital scoring system known as the Chinese Social Credit System (SCS) classifying behavior into morally "praise-" and "blameworthy" [29]. Thereby, all legal entities including companies and public institutions (among others) receive an 18-digit ID called the Unified Social Credit Code,[1] which corresponds to the 18-digit ID card number for

We thank the anonymous reviewers for their comments. We appreciate the support from the German Institute for Trust and Safety on the Internet (DIVSI).
* Severin Engelmann and Mo Chen equally contributed to this work.
+ Felix Fischer is also affiliated with Projects by IF, London.

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287585

[1] http://www.gov.cn/zhengce/content/2015-06/17/content_9858.html, last accessed on November 19, 2018.

Chinese citizens. Presumably, based on these IDs, the SCS will collect and evaluate behavioral data and may assign scores that result in material benefits and reputational praise or material exclusion and reputational loss. Or, in the words of the Chinese government, the goal of the SCS is to "allow the trustworthy to roam everywhere under heaven while making it hard for the discredited to take a single step" [21, 29].

But how can citizens, companies, and social institutions know what behaviors are "good" and "bad" in the SCS? Put differently, how transparent is the current SCS in providing information on "good" and "bad" behaviors? Answering this question requires a conceptualization of transparency. Here, we rely on the definition proposed by Turilli and Floridi, which conceptualizes transparency as "the choice of which information is to be made accessible to some agents by an information provider"[30]. First, this definition distinguishes between an information provider, which makes information accessible, in this context the Chinese government, and agents or entities that depend on this information for their decision-making. Secondly, this definition recognizes that information transparency is an "ethically impairing or enabling factor when the information disclosed has an impact on ethical principles"[30]. Both of these components are highly relevant for the SCS since participants are dependent on the information provided to make decisions that can lead to reward or punishment.

Recently, the Chinese government has started issuing behavioral information on several platforms (see Section 2 for more information). In this empirical study, we review a subset of this behavioral information released on two central SCS platforms: the official SCS national website "Credit China" and its equivalent municipal outlet "Credit China (Beijing)".

On the former site, we collect and analyze 156 news reports about "good" behaviors (we refer to as "positive" cases), and 789 equivalent reports about "bad" behaviors ("negative" cases). In these "negative" portraits, individuals are commonly stereotyped as so-called "Laolai (老赖)" – the epitome of a financially dishonest individual in China. Since all stories we collected are news reports about real-life events portraying a morally "good" or "bad" individual, they all include descriptive norms highlighting "desirable" and "undesirable" characteristics of individuals in Chinese society today.

Next, on "Credit China (Beijing)", we retrieve a large number of records of "good" and "bad" behavior from the so-called Redlist and Blacklist. Thus, our approach is as follows: first, we collect and statistically analyze close to 200,000 Blacklist and Redlist records from "beijing.gov.cn/creditbj", the SCS's information platform for China's capital, Beijing. Hence, based on machine learning topic modeling and manual text coding, we identify the common semantic patterns of close to 1000 reports on "good" and "bad" behavior published on the national SCS platform "www.creditchina.gov.cn".

We show several informational asymmetries that characterize the current degree of transparency of the governmental SCS's information platforms. Finally, we discuss how degrees of transparency could correspond to different incentive strategies of socio-technial systems that rate legal entities in society.

Our paper has the following structure. In Section 2, we discuss the development of China's SCS and review related work. In Section 3, we present our data acquisition and data analysis approach. We conduct our analysis in Sections 4 and 5. We discuss our results and offer concluding remarks in Section 6.

2 BACKGROUND

The implementation of the SCS rests on at least three main factors: First, lack of honesty and trust[2] in Chinese society has become a serious issue manifested in the numerous news reports about food poisonings, chemical spills, financial and telecommunications fraud, and academic dishonesty over the past two decades [13, 22]. It is estimated that Chinese enterprises suffer from a loss of 600 billion RMB (around 92 billion USD) per year due to dishonest activities[3]. According to a survey conducted by Ipsos Public Affairs [14], "moral decline" was regarded as the most serious issue in China in 2017. 47% of Chinese respondents ranked it as one of the top 3 greatest concerns, while the same issue was only mentioned by 15% of total respondents worldwide.

Secondly, China's SCS is expected to boost the domestic economy. The Chinese government hopes that the SCS will give millions of Chinese citizens without a financial history access to credit and investment opportunities in the domestic market. China has the largest unbanked population in the world (in absolute numbers), with more than 225 million citizens having no bank account [5]. So far, only 320 million Chinese citizens have a credit record[4]. However, the sustainability of China's economic growth partially depends on an increase in domestic spending. Through the SCS, citizens could apply for loans based on trustworthiness scores without having to prove their financial creditworthiness.

Finally, in Chinese society, the concept of personal identity is largely determined by Confucian principles [6, 32]. Accordingly, personhood is supposed to extend from the private to the public sphere thereby somewhat losing its private and public boundaries. In other words, normative expectations on individuals hardly account for the distinction between a private and a public sphere. The division between a private and a public persona is often conceived as trying to be secretive as privacy is commonly conceived as hiding something shameful [34]. In fact, until recently, privacy was primarily protected under the right of reputation in Chinese civil law [33]. At the same time, the public interest ranks highly in Chinese civil law [3]: "private information protected from disclosure refers to information that is irrelevant to the public interest or to the interests of other persons." However, while the Chinese concept of privacy is evolving, it is expected to remain distinct from other societies [18]. Overall, the introduction of the SCS is hardly perceived as a privacy-violating system in Chinese society, which is perhaps surprising from a Western perspective [16].

2.1 Current state of the SCS

At the current stage, the SCS remains fragmented, being developed at national, provincial, municipal, and ministerial levels with no clear unified structure. In the past years, provinces and cities have

[2]The characters "诚信 (chengxin)" literally mean both honesty and trust in Chinese.
[3]This information is included in the "Report on China's Honesty Building Situation (Zhongguo Chengxin Jianshe Zhuangkuang Baogao)". The full report is not publicly available, but parts of the report (in Chinese) are accessible through: http://society.people.com.cn/n1/2016/0523/c1008-28370202.html, last accessed on November 19, 2018.
[4]See "Inspiration of the US Non-traditional Credit Information Mechanism" available on the platform of "Credit China" at http://www.creditchina.gov.cn/zhengcefagui/tashanzhishi1/201712/t20171207_98701.html, last accessed on November 19, 2018.

developed various prototype models for the SCS [17, 35]. Importantly, the SCS also takes companies, government departments and judicial organizations as its targets [29]. This means that some companies have a special role in the SCS. Since 2015, eight companies were granted permission to run individual credit services with the purpose to implement pilot SCS programs [23]. Individually, none of the eight companies received a licence to continue individual credit services after the two-year trial period ended in 2017. Instead, together with the China Internet Finance Association (run by the People's Bank of China), they recently have become common shareholders of a company called Baihang Credit, which received the first credit scoring licence in February 2018.

2.2 Related Work

We are unaware of *any* research project that conducts a data-driven analysis of the currently observable data practices of key sites of China's SCS. However, we have identified two empirical research studies that help understand how the SCS is being communicated and discussed by Chinese media [23], and how it is being perceived by Chinese citizens [16].

Ohlberg et al. collected official Chinese news articles and public communications, as well as social media postings on Chinese blogs, forums, and bulletin board services about the SCS for a six-month period in 2017 [23]. The large majority of news articles has a positive focus and highlight the SCS as a "cure-all for social and economic problems". Criticism is mostly aimed at the slow implementation progress or directed at commercial initiatives in the SCS. Citizens' social media postings rarely address privacy issues and rather focus on how to game the system to achieve a higher social credit score within commercial SCS applications. Of relevance to the latter point, the implications of gamifying social credit are also being discussed from a non-empirical perspective by other scholars [19, 24].

Kostka [16] conducted an online survey with about 2,200 Chinese citizens that was distributed via different channels including websites and apps. Due to the widespread internet surveillance in China, the validity of such online surveys remains questionable at least to some extent. According to her findings, about 80% of the respondents have a positive perception of governmental and commercial SCS initiatives. Interestingly, older and more educated respondents have a higher approval rating. In contrast, these demographic factors are typically associated with higher privacy concerns in Western societies (see, for example, [1]). Several policy papers address the relationship between the SCS and the danger of mass surveillance (e.g., [20]).

Finally, there is rigorous work on comparing financial credit reporting systems [15], which, however, predates the emergence of the SCS in China and focuses on the financial aspects of credit reporting. Likewise, privacy considerations concerning private entities facilitating credit and background reporting have, for example, been explored by Hoofnagle [12].

2.3 Ethical Issues

Our analysis is built on publicly available data from key sites of China's SCS, which is posted with the intent of public scrutiny. Our paper includes screenshots from the *currently available* implementations. We have blurred any personally identifiable data.

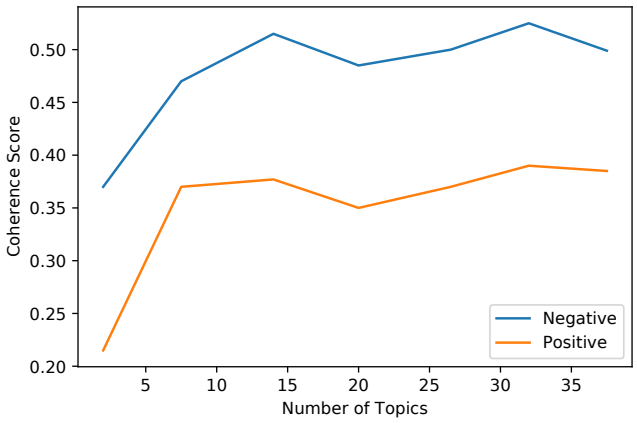


Figure 1: Coherence score C_v for topic models of negative and positive case studies using different topic counts.

3 METHODS

We used computer-assisted content analysis methods to explore the level of transparency of current behavioral information published on the two previously mentioned SCS websites. First, the column-and-row structured records of both the Blacklist and the Redlist on the SCS's Beijing platform[5] were crawled and statistically evaluated. Hence, to understand the semantic and structural patterns of both "positive" and "negative" case studies, we crawled news reports on "bad" behavior labeled as "Typical Cases (典型案例)"[6] and on "good" behavior labeled as "Stories of Integrity (诚信人物/故事)" under the section of "Integrity Culture (诚信文化)"[7] on the national SCS information platform "Credit China"[8]. We then applied statistical topic modeling based on Latent Dirichlet Allocation (LDA) to all available 156 news reports on "good" behavior ("positive" cases) and 789 news reports on "bad" behavior ("negative" cases) on August 12, 2018.

We preprocessed the downloaded documents by applying *jieba*[9] for segmentation and stopword filtering of Chinese text. We used the stopword corpus compiled by the Chinese search engine Baidu[10]. After tokenization of the given text, we applied *tf-idf* to re-weigh term counts. As we had no reasonable expectation for the number of topics k to be detected within the given document corpus, we performed optimal topic number search. Thereby, we created several LDA models for "positive" and "negative" case studies and calculated the topic coherence measures C_v as proposed in [25]. We started with $k = 2$ and increased the number of topics until an upper bound of $k = 40$. As shown in Figure 1, coherence values of models for both document sets increased until $k = 15$ before flattening out. Therefore, we investigated the top-30 most salient terms for each of the fifteen topics produced by these models [4]. Thereby, we set $\delta = 0.6$ within the applied relevance metric [28].

[5] http://www.creditbj.gov.cn/xyData/front/creditService/initial.shtml%20?typeId=4.
[6] https://www.creditchina.gov.cn/home/dianxinganli1/?navPage=6.
[7] https://www.creditchina.gov.cn/chengxinwenhua/chengxingushi/.
[8] https://www.creditchina.gov.cn/.
[9] https://github.com/fxsjy/jieba.
[10] http://www.baiduguide.com/baidu-stopwords.

Figure 2: Three lists publishing records of "negative" behavior: from left to right, the first arrow points to Blacklist, the second arrow to Special Attention List, and the third arrow to Administrative Punishment.

Moreover, we also reviewed the results for $k = 10$, $k = 20$, and $k = 30$ in order to further manually verify the optimal topic number. We found the optimal model with $k = 10$ for both "positive" and "negative" cases. Finally, we further selected 5 main topics for the "positive" cases and 7 topics for the "negative" cases (see Table 3 in the Supplementary Materials for topics selected for the "positive" cases, and Table 4 in the Supplementary Materials for "negative" case topics).

Based on our topic modeling results, we selected the 4 most related cases (highest predicted probability of belonging to the topic) for each of the topics.[11] We then manually analyzed 20 "positive" cases and 26 "negative" cases[12] in detail. One author first reviewed 5 "positive" and 5 "negative" cases, respectively, and drafted a coding guide, which was then reviewed iteratively by another author, refined, and retested to generate consistent definitions. As a result, we developed two coding schemes for "positive" and "negative" cases (see Table 1 for the coding scheme applied to "positive" cases and see Table 2 for the coding scheme used to analyze "negative" cases). After reliability was established, we examined all 46 cases for structural and thematic commonalities. Each coding sheet contained the information from one "positive" or "negative" case. Once the coding sheets were completed, we grouped and analyzed the information contained in them.

4 RESULTS

4.1 Blacklists

On the platform of "Credit China (Beijing)", we found three publicly accessible databases providing information on "bad" behavior, all of which could be queried by search term. Translated from Chinese (see Figure 2), they were termed the following: 1) Blacklist (1,137,546 entries), 2) Special Attention List (9,229,179 entries), and 3) Administrative Punishment (14,885,789 entries).

The Blacklist further contained 16 subcategories for "bad" behavior. For the Blacklist, we crawled two of these subcategories, one containing records of individuals that have been banned from

[11] For "negative" cases, there are only three cases for Topic 6 (measures taken against crime) and Topic 7 (public transport regulation violation), respectively.
[12] There were only 3 cases for 2 out of the 7 topics.

Pattern	Definition	Example
Bio-info	full name	今年70岁的刘某某，为了一句诺言，一辈子踏踏实实做一名"小村大医生"。
	age	古亭村77岁的老人蓝某某为了归还欠银行的一笔500元死账。
	living place	蓝某某出生在遂昌县云峰街道古亭村。
	profession	这位一天两次捡到钱包的"好运人"就是蒙阴一中的的英语教师耿某某。
Social class	low	一个清贫的普通农家，父亲、儿子、孙女毫无怨言地赡养一位无任何血缘关系的"外人"。
	middle	陈某的妻子说，他们家里也就是普通家庭，上有老下有小。
	high	这句话时常在内蒙古明泽集团董事长王某某的心里翻腾着。
Sacrifice for the common interest	material sacrifice	他隔天检查药柜，受潮的药直接销毁，损失的药费自己承担。
	non-material sacrifice	每天为他做三餐，每天打针吃药，就连端屎端尿的活也揽下来。
Rewards	reputational rewards	他被评为全国农村青年创业致富带头人、北京市优秀农村实用人才。
	material reward refusal	钱包的主人一个劲地要给她塞钱。肖某某坚决地拒绝了。
Virtue cascade	trustworthy and honest	为了不让养殖户遭受损失，彭某某把风险留给自己，仍按照回收合同原价收回了养殖户的肉鸭。
	hardworking	虽然有时一天连饭都顾不上吃，还帮助菜农一起装菜卸菜，忙到了深夜还要了解市场信息、掌握蔬菜的价格趋向。
	self-discipline	虽然银行减免并注销了这笔贷款，但放在我私人账户的钱一定要还上。
	helpful	积极参加协会组织的慰问残疾人、资助贫困大学生活动。
	care-taking	他们一家三代几十年如一日地照顾着丁某某老人。
	sense of responsibility	她以当好水资源质量的守望者为己任。

Table 1: Coding scheme for "positive" cases. All "positive" cases included biographical information of the individual and indicated his or her social class. Other codes described the individual's sacrifice for the common interest, the rewards obtained, and the further attribution of other virtues (virtue cascade).

Pattern	Definition	Example
Bio-info	anonymous (for individuals, surname only)	当宁陵县法院执行干警在被执行人郭某家的楼顶将其抓获时，郭某无奈地低下了头。
	anonymous (company name not provided)	原告北京某装饰工程有限公司为被告北京某文化有限公司所有的房屋进行建设、装修。
Implementing Agency	the court	海淀法院3月6日出动执行法官、法警等共计50余人，对15起案件进行集中强制执行。
	Public Security Bureau	华龙区法院的执行法官远赴拉萨，与当地公安机关通力合作。
	telecommunication company	由商南法院向中国移动、联通、电信三大通信运营公司出具协助执行通知书，对失信被执行人实行彩铃和短信曝光。
Causes for punishment	refusing to repay individuals	当地法院判决吕某赔偿梦某医疗费、残疾赔偿金等损失46万元。吕某拒不履行赔偿义务，甚至远走他乡。
	refusing to repay banks	岫岩法院判决某食用菌公司偿还银行贷款本金380万元及相应利息。判决生效后，食用菌公司一直没有履行。
	refusing to repay companies	原告北京某装饰工程有限公司为被告北京某文化有限公司所有的房屋进行建设、装修，施工结束后，被告拖欠原告工程款400余万元。
Reasons to fulfill obligations	actions taken by the court	在中牟法院执行干警的全力配合下，成功将被执行人吕某拘留。
	threatened to be placed on Blacklist	法院将肖某纳入了"老赖"名单里，将他的大头照向社会公布。

Table 2: Coding scheme for "negative" cases. All cases provided anonymized biographical information, an entity implementing the punishment, justification of the punishment, and descriptions on why the obligations were fulfilled in the end.

participating in the securities market (Securities Market Entry Prohibition, 422 entries) and one listing companies with debts (Blacklist of Company Debtors, 1,116,707 entries = 98.2% of all Blacklist entries). For the Blacklist of individuals, all 422 entries included extensive explanations for the punishment (e.g., length of ban) referencing financial law (see Figure 3). Apart from the censored ID card number, the full names of all individuals were published.

Due to the large amount of company records we found on the Blacklist, the Special Attention List, and Administrative Punishment, we crawled the first 1000 pages for these lists. For the Blacklist of companies with financial debt, this resulted in a total of 131,485 entries all of which featured information on why an entity had been blacklisted (see Figure 4). Out of these 131,485 entries, 128,006 entries specified that the financial obligation had not been fulfilled

Figure 3: An entry from the Blacklist of "Securities Market Entry Prohibition". The first column, from top to down: the first arrow points to "name of punishment" and the second points to "content of punishment". The table on the right side of the second arrow shows the detailed explanation of the punishment.

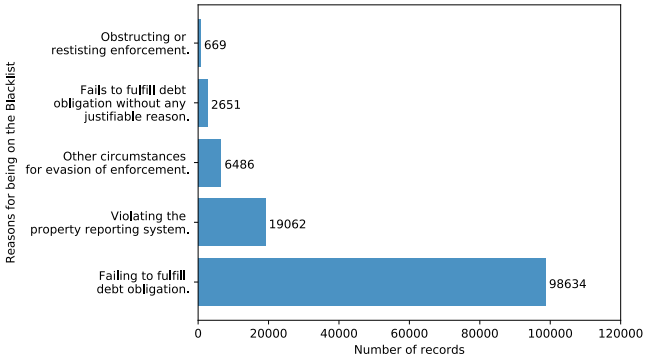


Figure 4: The top 5 reasons for being on the Blacklist of company debtors.

at the time of crawling (corresponding field not shown). Entries included a reference to legal regulation and specified the full name of the company (see Figure 5). Note that some companies listed had multiple entries corresponding to multiple breaches. Together with these explanations, we crawled the date of publication on the Blacklist for each entry. We found that on one day in June 2018, 95.6% of all entries (125,747) had been published on the Blacklist for companies (see Figure 6). This probably indicates that these records had already been collected and processed by another entity before being transferred to and published on the Blacklist.

For the Special Attention List, we collected 30,625 entries containing information on companies that had violated business operation regulations. For all records collected, companies had been blacklisted for providing various types of false information to the authorities (see Figure 7).

Finally, our crawler returned 32,719 entries for the Administrative Punishment register that contained information on both

数据来源：	高法
数据类别：	失信被执行人名单
案号：	（2018）鲁1092执47号
主体名称：	威海久登建材有限公司
企业法人姓名：	
组织机构代码：	558932122
执行法院：	威海经济技术开发区人民法院
地域名称：	山东
执行依据文号：	（2016）鲁1092民初1085号
作出执行依据单位：	威海经济技术开发区人民法院
法律生效文书确定的义务：	被执行人给付申请人昭共计3023657.49元
被执行人的履行情况：	全部未履行
失信被执行人具体情形：	有履行能力而拒不履行生效法律文书确定义务
发布时间：	20180525
立案时间：	20180118
已履行部分：	暂无
未履行部分：	暂无
最新更新日期：	2018-06-11

Figure 5: Screenshot of a company's Blacklist entry. Left column, the first arrow points to a field explaining the specific context of the case, the second arrow points to the date of publication of this entry on the Blacklist.

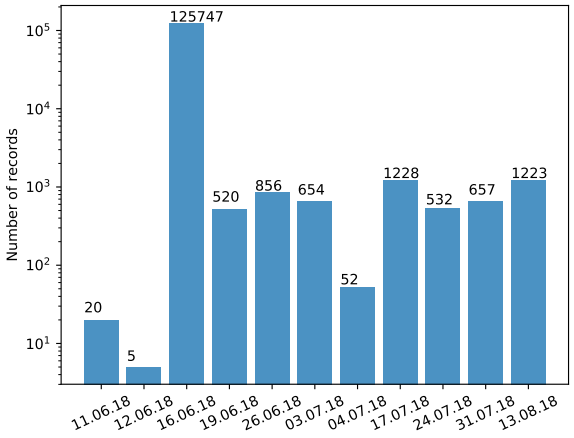


Figure 6: Publication dates of Blacklist entries for company debtors.

individuals and companies (see Figure 8). As Figure 9 shows, the majority of records of the Administrative Punishment register reported traffic rule violations.

Correspondingly, fines were the most widely used measure (see Figure 10). We also found that only company entries of the Administrative Punishment register and the Blacklist consistently featured the Unified Social Credit Code.

On the national SCS information platform "China Credit", we found another Blacklist issued by the Civil Aviation Administration of China (中国民用航空局)[13]. This list, which is updated every month, publishes information on individuals that are excluded from aircraft travel for a period of one year due to misbehavior on airplanes or airports (data collected on August 10, 2018; see Figure 11). According to the list published in August, 2018, 946 individuals were banned from air travel for one year. Among others, the list provided full name, censored ID number, and explanations why individuals had been punished (see three arrows in the first row of Figure 11). Being banned from air travel resulted from taking illegal objects on airplanes, smoking on airplanes, or boarding airplanes

[13]https://hmd.creditchina.gov.cn/, last accessed on November 5, 2018.

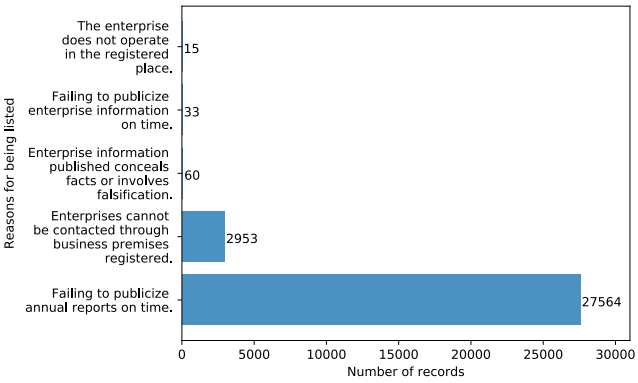


Figure 7: The top 5 reasons for companies to be on the Special Attention List.

主体名称：	赵某
统一社会信用代码：	
处罚类别1：	罚款
处罚事由：	过度疲劳仍继续驾驶的
处罚依据：	暂未入库
处罚名称：	过度疲劳仍继续驾驶的
处罚类别2：	
组织机构代码：	
工商登记码：	
税务登记号：	
法定代表人名称：	
处罚结果：	交通警察总队武清支队对赵文选进行罚款的处罚
处罚期限：	
处罚机构：	交通警察总队武清支队
处罚部门：	
处罚决定日期：	2017/01/03
当前状态：	正常
严重程度：	
地方编码：	120000
创建时间：	2018-07-11 00:00:00

Figure 8: A record of the Administrative Punishment register. The first column, from top to down: the first arrow points to the field "type of punishment" and the second points to the field "reasons for punishment".

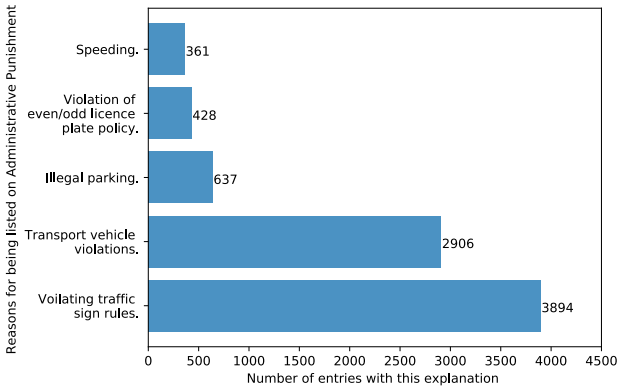


Figure 9: The top 5 reasons why individuals or companies are placed on the Administrative Punishment register.

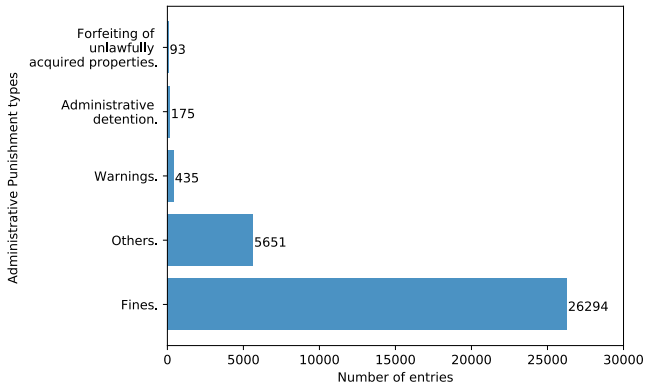


Figure 10: The 5 types of Administrative Punishments.

with a fake passport. The figure also indicates that the list contained names and ID numbers of non-Chinese citizens providing some evidence that foreigners were not excluded from the SCS.

4.2 Redlist

We found one type of list documenting information on "good" behavior - the Redlist. It contained a total of 1,206,944 entries distributed across 24 categories (3 categories for redlisted individuals, 21 categories for redlisted companies). The categories for individuals, translated from Chinese, are: 1) Taxi Star (1557 entries), 2) Top Ten Tour Guides (14 entries), and 3) Five-Star Volunteer (603 entries). For all entries, the full name of the person and his or her partially censored ID number were given. The Five-Star Volunteer category displayed the gender of the person as well as the amount of volunteering hours carried out per person. The lowest amount of volunteering hours documented was 1500 (which was probably the necessary threshold to be listed) and the highest was 25,400. None of the entries we collected from the Redlist provided an explanation justifying why such a honorary title had been awarded to that person (see Figure 13). Thus, we cannot report any observations about justifications on "good" behavior from our Beijing Redlist analysis.

Company categories referred either to tax awards (e.g., A Class Taxpayer) or to other honorable statuses such as Harmonious Labor Relations or Excellent Contributor to Developing Chinese Socialism. Just like the Redlist entries of individuals, there were no justifications explaining why a honorable title had been awarded to a company. No Redlist entry contained the Unified Social Credit Code. Generally, Figure 13 shows a single record of an entity that can display several "positive" and "negative" entries. Thus, there is reason to believe that the interface shown in Figure 13 functions as the governmental SCS information template: recording and making transparent information on rewards and/or sanctions to the public.

Importantly, every Blacklist and Redlist record we collected featured a "Disagreement/Correction (异议/纠错)" function (see Figure 12). This function allowed citizens to object to a Blacklist or Redlist decision by providing a statement of up to 2000 Chinese characters (submission required 18-digit ID number).

4.3 Coding results for "positive" cases on "good" behavior

News reports on "good" behavior were introduced as "Stories of Integrity (诚信人物/故事)" posted under the section of "Integrity Culture (诚信文化)" on the national SCS information platform "Credit China". All of the 20 "positive" cases selected described how a protagonist sacrificed his or her self-interest (both material and non-material) for the common good. Moreover, all cases centered on "trustworthiness" and "honesty" as key SCS virtues. The stories all followed the same narrative structure: they first provided detailed biographical information of a person (full name, social class, profession, family status), followed by a dilemma: the protagonist could either engage in "dishonest" behavior winning him or her an immediate small reward or get a large future reward by being "honest". Once the person had enacted the "honest" behavior, which happened in all the "positive" reports we analyzed, the narratives ended with a virtue cascade.

Take, for example, cases in which individuals found and returned lost property to an owner. Here, all four cases assigned to the topic"return lost property to owner" ended by further attributing "self-discipline", "helpfulness", "care-taking for others", and a "sense of responsibility" to the protagonist as part of a virtue cascade. Another commonality across the selected cases was that all protagonists were morally "praised" by their social environment. Also, the protagonist was recognized for his or her "good" behavior by official agencies or the media in the form of "honors", "decorations", or a "cute nickname". On the other hand, when a material reward was offered for the "good" behavior, as in all cases with topics "family and community relationship and repayment", "return lost property to owner", and "social entrepreneurship to help people out of poverty", the protagonist refused the material reward at all times.

4.4 Coding results for "negative" cases on "bad" behavior

Reports about "bad" behavior were labeled as "Typical Cases (典型案例)" on the homepage of "Credit China" with the sources being both local newspapers and the platform itself. The 26 selected "negative" cases relating to 7 topics all revolved around one common theme, the "Laolai (老赖)": a term specifically referring to individuals and companies refusing to repay debts. These cases were presented in two ways. The 4 cases with the topic "public shaming" were about the courts' actions in solving repayment problems. The remainder of the stories were about specific individuals or companies. All individuals and companies were anonymous in the selected cases. Local courts collaborated with local telecommunication companies in all 4 cases with the topic "public shaming", and the Public Security Bureau played an important enforcement role in all cases with topic "public transport regulation violation". In these reports, both the compulsory actions taken by the court and the threat of being placed on the Blacklist forced the "Laolai" to fulfill the stated obligation. Generally, both "positive" and "negative" case studies we analyzed were homogeneous in structure, framing, and content. This could indicate that they had been deliberately formulated to propagate the SCS's conceptualization of "good" and "bad" behavior.

256	刘[redacted]	420281****0919691X	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	湖北省公安厅机场公安局直属分局航站区派出所 行政处罚决定书	鄂机公直航【行罚决】字（2018）第49号
257	姜[redacted]	370103****12182531	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	北京首都国际机场公安分局公安行政处罚决定书	京机公分（法）决字[2018]第296号
258	李[redacted]	340121****0918915X	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	安徽省公安厅机场公安局合肥机场派出所行政处罚决定书	直公（机）行罚决字[2018]64号
259	赵[redacted]	532331****06090635	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	云南省公安厅民用机场公安局直属公安分局 当场处罚决定书	0548462018050702
260	SAURAM[redacted]	AA7242****	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	深圳市公安局机场分局 行政处罚决定书	深公（机）行罚决字【2018】00280号
261	SHOTAYRVNUR[redacted]	N11668****	随身携带或托运国家法律、法规规定的危险品、违禁品和管制物品的；在随身携带或托运行李中故意藏匿国家规定以外属于民航禁止、限制运输物品的	民航新疆机场公安局 行政处罚决定书	民航新机公（候派）行罚决字[2018]009号

Figure 11: A screenshot of the Blacklist for individuals that are banned from flying on commercial airplanes. In the first row, from left to right, the first arrow points a field containing the full name of the individual; the second to censored ID number; and the third to explanations why individuals have been punished. Two arrows at the bottom left indicate entries of two foreign passengers.

Figure 12: Example of a company's Blacklist entry. The black circle on the upright corner indicates the "Disagreement/Correction (异议/纠错)" function.

5 ANALYSIS

The results of our content analysis demonstrate that there are currently multiple informational asymmetries in both datasets.

5.1 Listed companies versus listed individuals

Currently, companies make up the majority of entries on both the Blacklist and Redlist of Beijing's SCS platform. We found that companies which are involved in the construction of the SCS were also included in the list. For instance, Alibaba (with Zhima Credit) and Tencent (with Tencent Credit) were both granted permission to start individual pilot credit service programs in 2015 and have provided digital data collected from online shopping and social media to the SCS. Both Alibaba and Tencent were listed as A-level Taxpayers on the Redlist. Since we only crawled the Beijing SCS platform, we cannot make any claims about the transparency of other SCS Blacklist and Redlist websites.

Our analysis of "positive" and "negative" cases demonstrates the opposite: here, the majority of reports on either "good" or "bad" behavior focuses on individuals' behaviors. For our manually coded sample, only 15.4% of "negative" reports and 30.0% of "positive" reports featured companies. In both "negative" and "positive" cases that featured companies, however, reports centered on the person in charge of the company typically highlighting the CEO's virtues and vices. In other words, it is not the company as such that is "blamed" or "praised," but rather the person responsible for the company. Such portraits, therefore, signal that individuals are not shielded by large institutions but can be made responsible for their "good" or "bad" decision-making.

5.2 Justifying punishments versus justifying rewards

All entries of the Blacklist explain why a person or company is currently registered on the Blacklist. Moreover, Blacklist explanations include legal terms and refer to laws and regulations. In other words, Blacklist explanations make transparent the mechanism of punishment by specifying a causal link between behavior and consequence. This is perhaps best illustrated by the Blacklist on individuals excluded from air travel (see Figure 11). The legal threat contained in the entries of the Blacklist could furthermore signal that a specific "dishonest" behavior can be detected and sanctioned.

On the other hand, not a single entry of the Redlist includes a formulated explanation on why a person or company has been awarded a honorary title. We found that fulfilling legal obligations (Class A Taxpayer), performing professional (Taxi Star) or volunteering (Five-Star Volunteer) duties can result in reputational gain in the current SCS. However, the mechanisms or criteria determining when an individual or a company secures a place on the Redlist are not further explained. Taken together, the current SCS makes behaviors leading to punishments more transparent than behaviors

Figure 13: Example of a Redlist entry for an individual with the honorary title Five-Star Volunteer. The record does not justify why the honorary title was awarded.

resulting in rewards. More generally, our study could not identify publicly available information associating specific behaviors to a scoring or rating mechanism.

5.3 Types of punishments versus types of incentives

The most common reason for a company to be placed on any of the "negative" lists is failure to pay back debt (the second most common reason is informational misconduct). Failure to pay back debt is also the most prominent reason given for why protagonists of the "negative" cases are registered on the Blacklist. The Chinese term for "Laolai" appeared 481 times in the 789 "negative" reports we collected. All "negative" stories we manually coded report on the activities of a "Laolai" person (either as an individual or as the legal representative of a company). In terms of punishment, individuals and companies face both the material loss specified in the corresponding legal regulation as well as the consequences of being publicly shamed on the Blacklist. In more than 40% of the narratives on "negative" behavior, an individual is threatened to be placed on the Blacklist leading to the immediate compliance of the individual.

On the other hand, individuals and companies on the Redlist receive moral "approval" and reputational gain. Similarly, "positive" cases report on individuals that gain reputational rewards, while at the same time rejecting material incentives when offered as a consequence of their "role-model" behavior. Still, being listed on the Redlist is not mentioned or even indicated by any individuals as a motivational factor for their behaviors. All stories we analyzed emphasize that a morally "praiseworthy" activity is "praiseworthy" when it is "genuinely" moral rather than instrumental in obtaining a material reward. Furthermore, all "positive" stories feature a virtue cascade: once an individual is described as "genuinely honest" or "trustworthy", he or she is attributed other "positive" virtues as a consequence.

6 DISCUSSION & CONCLUDING REMARKS

In this first study of key websites of the Chinese SCS, our goal was to understand how transparent the SCS currently is in providing information on "good" and "bad" behavior. To this end, we collected and analyzed 194,829 Blacklist and Redlist entries from the Beijing SCS website "beijing.gov.cn/creditbj" and applied a machine learning topic modeling algorithm to almost 1000 reports on "positive" and "negative" behavior crawled from the national SCS information platform "www.creditchina.gov.cn". Finally, we manually coded a sample of these texts to understand what kind of specific behavioral information they contain.

The main question arising from our findings, we believe, is whether the degree of the current SCS's transparency is intentionally engineered or whether it is simply a manifestation of work in progress. Is there a purpose in explicitly describing and publishing the causal link between behavior and sanction while leaving information on getting rewards deliberately vague? First, the asymmetries in information provided between the Redlist and the Blacklist could be motivated economically: while an infinite amount of people can be excluded from valuable material resources, only a finite amount can be given valuable resources (e.g., a first-class train ticket). Detailed instructions on how to win rewards could therefore lead to distribution problems since many individuals could implement them. On the other hand, another explanation for the current informational asymmetries of the SCS might be that already existing records of legal offenses were used to start filling Blacklists. Consequently, these records entail more justifications since they refer to specific legal articles or regulations.

The degree of transparency of the SCS observed in this work could also be motivated by behavioral engineering goals. Let's imagine for the moment the system were completely inscrutable (i.e., the system did not justify a score increase or decrease and eventually a given punishment or reward, respectively). In this case, individuals would have little possibility to understand when the SCS rewarded and when it sanctioned specific types of behaviors. Moreover, besides being oblivious to the moral code of conduct, individuals would not have the ability to contest the system's decision-making process (again, to negotiate a norm one must have the necessary epistemic resources to do so). Note that this issue is also debated in the context of the "Right to Explanation" of the European Union's General Data Protection Regulation [27, 31]. A fully transparent scoring system, on the other hand, would precisely map behaviors to rewards or sanctions. Indeed, in the context of a nationwide

digitally-implemented scoring system, full transparency must account for the mechanism that leads to the distribution of rewards or sanctions. This degree of transparency would offer individuals the possibility to understand the system's decision-making procedures at least to a certain extent. In our analysis of SCS Blacklist and Redlist records, we did not identify an explicit SCS scoring mechanism. We have shown, however, that the SCS already enables citizens to dispute single Blacklist and Redlist records. On the other hand, a fully transparent SCS would possibly create other problems: if the SCS became fully transparent in regard to its scoring mechanisms, complying to a norm would likely become a market transaction. In fact, research on intrinsic and extrinsic motivation suggests that introducing an external reward to a norm-guided behavior turns this behavior into a commodity that can be bought [10, 11]. This phenomenon, termed "crowding-out effect", results in fewer people engaging in this behavior since the consequences of failing to act can simply be compensated by financial means [2, 9, 26]. For example, if one reliably receives monetary compensation for being honest, being honest will no longer be evaluated as a moral behavior for both the actor and the recipient. As this line of research suggests, individuals will likely stop attributing a genuine moral character to individuals with a high score in a fully transparent SCS.

Our analysis provides evidence that the currently implemented SCS possibly attempts to counter such a transformation of moral behavior into market transactions. All of the "positive" case studies unambiguously emphasize that norm conformity is "good" because it is "morally valuable" for both average citizens as well as CEOs. None of the Redlist entries describe a connection between moral behavior and external material reward. Rather, they contain virtue signals and reputational gains by awarding symbolic honorary titles (e.g., Five-Star Volunteer). On another sub-page of the national SCS website, we found the publication of 32 ancient Chinese fables (not shown) also promoting self-concepts comprising virtues of being a morally "good" Chinese citizen. In contrast, our analysis on the corpus of "negative" case studies demonstrates the propagation of a "negative" self-concept ("Laolai") attributable to a specific offense (i.e., intentionally not paying back debt). Taken together, our analysis suggests that degrees of transparency can serve different behavioral engineering goals in the context of a digital scoring system.

REFERENCES

[1] Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Security & Privacy* 3, 1 (2005), 26–33.

[2] Roland Bénabou and Jean Tirole. 2006. Incentives and prosocial behavior. *American Economic Review* 96, 5 (2006), 1652–1678.

[3] Yongxi Chen and Anne Cheung. 2017. The transparent self under big data profiling: Privacy and Chinese legislation on the Social Credit System. *The Journal of Comparative Law* 12, 2 (2017), 356–378.

[4] Jason Chuang, Christopher Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 74–77.

[5] Asli Demirguc-Kunt, Leora Klapper, Dorothe Singer, Saniya Ansar, and Jake Hess. 2018. *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution.* The World Bank.

[6] Charles Ess. 2006. Ethical pluralism and global information ethics. *Ethics and Information Technology* 8, 4 (2006), 215–226.

[7] Luciano Floridi. 2013. *The Ethics of Information.* Oxford University Press.

[8] Bernard Gert and Joshua Gert. 2017. The definition of morality. In *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[9] Uri Gneezy, Stephan Meier, and Pedro Rey-Biel. 2011. When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives* 25, 4 (2011), 191–210.

[10] Uri Gneezy and Aldo Rustichini. 2000. A fine is a price. *The Journal of Legal Studies* 29, 1 (2000), 1–17.

[11] Uri Gneezy and Aldo Rustichini. 2000. Pay enough or don't pay at all. *The Quarterly Journal of Economics* 115, 3 (2000), 791–810.

[12] Chris Hoofnagle. 2003. Big Brother's little helpers: How ChoicePoint and other commercial data brokers collect and package your data for law enforcement. *North Carolina Journal of International Law and Commercial Regulation* 29 (2003), 595–637.

[13] Yanzhong Huang. 2012. Why is not there a bottom line for food security issue in China (Zhongguo Shipin Anquan Weihe Meiyou Dixian?) (in Chinese). https://cn.nytimes.com/opinion/20120821/c21huang/

[14] Ipsos Public Affairs. 2017. What worries the world? https://www.ipsos.com/sites/default/files/2017-08/What_worries_the_world-July-2017.pdf.

[15] Nicola Jentzsch. 2006. *The Economics and Regulation of Financial Privacy: An International Comparison of Credit Reporting Systems.* Springer Science & Business Media.

[16] Genia Kostka. 2018. *China's Social Credit Systems and Public Opinion: Explaining High Levels of Approval.* Technical Report. Free University of Berlin. Available on SSRN: https://ssrn.com/abstract=3215138.

[17] Jianzhou Liu. 2011. Building the Social Credit System: The content, the model, and the trajectory (Shehui Xinyong Tixi Jianshe: Neihan, Moshi yu Lujing Xuanze; in Chinese). *Journal of the Party School of the Central Committee of the C.P.C.* 15, 3 (2011), 50–53.

[18] Yao-Huai Lü. 2005. Privacy and data privacy issues in contemporary China. *Ethics and Information Technology* 7, 1 (2005), 7–15.

[19] Andrzej Marczewski. 2017. The Ethics of Gamification. *XRDS* 24, 1 (2017), 56–59.

[20] Mirjam Meissner and Jost Wübbeke. 2016. IT-backed authoritarianism: Information technology enhances central authority and control capacity under Xi Jinping. *China's Core Executive: Leadership Styles, Structures and Processes under Xi Jinping* (2016), 52–57.

[21] Simina Mistreanu. 2018. Life inside China's Social Credit laboratory: The party's massive experiment in ranking and monitoring Chinese citizens has already started. https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/.

[22] Nature. 2018. China sets a strong example on how to address scientific fraud. https://www.nature.com/articles/d41586-018-05417-1

[23] Mareike Ohlberg, Shazeda Ahmed, and Bertram Lang. 2018. Central Planning, Local Experiments: The Complex Implementation of China's Social Credit System. https://www.merics.org/en/microsite/china-monitor/central-planning-local-experiments.

[24] Zahy Ramadan. 2018. The gamification of trust: The case of China's "Social Credit". *Marketing Intelligence & Planning* 36, 1 (2018), 93–107.

[25] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 399–408.

[26] Richard Ryan and Edward Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology* 25, 1 (2000), 54–67.

[27] Andrew Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (2017), 233–242.

[28] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70.

[29] State Council. 2014. Notice of the State Council on Issuing the Outline of the Plan for Building a Social Credit System (2014-2020); (in Chinese). http://www.gov.cn/zhengce/content/2014-06/27/content_8913.htm.

[30] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics of Information Technology* (2009), 105–112.

[31] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[32] Pak-Hang Wong. 2012. Dao, harmony and personhood: Towards a Confucian ethics of technology. *Philosophy & Technology* 25, 1 (2012), 67–86.

[33] Yanfang Wu, Tuenyu Lau, David Atkin, and Carolyn Lin. 2011. A comparative study of online privacy regulations in the US and China. *Telecommunications Policy* 35, 7 (2011), 603–616.

[34] Jinghong Xu. 2015. Evolving legal frameworks for protecting the right to Internet privacy in China. *China and Cybersecurity: Espionage, Strategy, and Politics in the Digital Domain* (2015), 242–259.

[35] Zhiling Zhao and Feng Ding. 2007. Shanghai, Zhejiang, Shenzhen Social Credit System models, problems and revelation (Shanghai, Zhejiang, Shenzhen Shehui Xinyong Tixi Jianshe Moshi, Wenti yu Qishi) (in Chinese). *Wei Shi* 10 (2007), 70–73.

[36] Jarrett Zigon. 2008. *Morality: An Anthropological Perspective.* Berg.

Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness

Severin Engelmann, Mo Chen, Lorenz Dang, Jens Grossklags
Professorship of Cyber Trust
Department of Informatics
Technical University of Munich
Munich, Germany

ABSTRACT

The Chinese Social Credit System (SCS) is a novel digital socio-technical credit system. The SCS aims to regulate societal behavior by reputational and material devices. Scholarship on the SCS has offered a variety of legal and theoretical perspectives. However, little is known about its actual implementation. Here, we provide the first comprehensive empirical study of digital blacklists (listing "bad" behavior) and redlists (listing "good" behavior) in the Chinese SCS. Based on a unique data set of reputational blacklists and redlists in 30 Chinese provincial-level administrative divisions (ADs), we show the diversity, flexibility, and comprehensiveness of the SCS listing infrastructure. First, our results demonstrate that the Chinese SCS unfolds in a highly diversified manner: we find differences in accessibility, interface design and credit information across provincial-level SCS blacklists and redlists. Second, SCS listings are flexible. During the COVID-19 outbreak, we observe a swift addition of blacklists and redlists that helps strengthen the compliance with coronavirus-related norms and regulations. Third, the SCS listing infrastructure is comprehensive. Overall, we identify 273 blacklists and 154 redlists across provincial-level ADs. Our blacklist and redlist taxonomy highlights that the SCS listing infrastructure prioritizes law enforcement and industry regulations. We also identify redlists that reward political and moral behavior. Our study substantiates the enormous scale and diversity of the Chinese SCS and puts the debate on its reach and societal impact on firmer ground. Finally, we initiate a discussion on the ethical dimensions of data-driven research on the SCS.

CCS CONCEPTS

• **Social and professional topics** → **Government technology policy**; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

China's Social Credit Systems; Reputation Systems; Digital Socio-Technical Systems; China.

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462535

ACM Reference Format:
Severin Engelmann, Mo Chen, Lorenz Dang, Jens Grossklags. 2021. Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462535

1 INTRODUCTION

In 2014, the Chinese government published the *Planning Outline for the Construction of a Social Credit System (2014-2020)* as part of its 12th five-year plan [32]. Following its release, media and research have offered various perspectives on the Chinese Social Credit System (SCS, 社会信用体系). Some Western media have characterized the SCS as a mass surveillance apparatus, with the purpose of calculating a digital "sincerity score" for each Chinese citizen based on a wide range of personal data [3, 23, 28]. Below a certain point level, citizens would face multiple restrictions, such as exclusion from air travel and high-speed trains. A positive score, on the other hand, would lead to discounts and preferential treatment for a variety of products and services. This "dystopian perspective" sees the unification of an authoritarian regime's policies and artificial intelligence (AI) to enforce social order by means of a sincerity score. Some media outlets have since revised their original viewpoints regarding such comprehensive sincerity scoring [16, 26].

Academic scholarship on the SCS has largely been theory-driven, which has led to the independent development and discussion of different conceptualizations. The SCS has been defined as a novel administrative policy program with the main goal of strengthening compliance of citizen and organizations with laws and regulations [1, 7]. The novelty consists in the public (at least temporary) disclosure of already existing citizen and organizational records on so-called digital blacklists and redlists. Blacklists publicly showcase non-complying individuals and organizations, while redlists, as their normative counterpart, show complying entities. In this perspective, the SCS deploys reputational tools with some similarity to company rankings or background checks on individuals in Western economies.

Other authors have called the SCS a big data empowered system that collects, processes, and evaluates vast amounts of personal data [6]. These data are ultimately aggregated and published as public credit information (PCI) on digital platforms. This line of research argues that PCI creates transparent citizens, not least due to the lack of a sufficient legal framework that protects personal data in China [22]. Some scholars have noted an all-encompassing application of credit to society's political, economic, and social activities. Thereby, the SCS marks the emergence of a so-called reputation state [9, 24].

As a governance tool, the SCS seeks to harness reputational information for purposes that go beyond neoliberal notions of regulating market failure. Still other perspectives frame the SCS as a social management program [36]. Drawing on concepts from systems engineering, a social management program considers society to be a complex system that can be optimized using digital technologies.

While these accounts disagree in many important regards, three points of agreement can be identified: first, multiple independent initiatives have been labelled as "SCS" [35]. One SCS is driven by the apps and services of big data companies (e.g., Sesame Credit) that distribute scores to consumers in voluntary promotion programs [5, 18]. Here, "voluntary" denotes consenting to the terms and conditions of the service. Second, *local* governments have tested SCSs that integrate different scoring systems in "prototype cities" (社会信用体系建设示范城市), such as Rongcheng and Suzhou. Participation in these local "credit scoring experiments" is mandatory for residents in these areas. Such policy experiments [14] can serve as models for other local SCS implementations but they are not necessarily a model for national implementation. Third, government-led SCS measures have been realized nationally. There are various types of blacklists (黑名单) and redlists (红名单) run by government agencies at different levels of administrative divisions (ADs) including municipalities and provinces, but also government departments at the national level. These platforms publicly display information to "shame"[1] or "praise" natural and legal persons (e.g., companies) for non-compliance or compliance with a variety of legal and social norms [10, 15, 19, 22, 30]. No entity can opt out from being listed. Depending on the type of list, entities are subjected to different types of reward or punishment over a wide range of areas, a process that has been termed "joint reward and punishment mechanism" (JRP) by the Chinese government [32]. Both natural and legal persons on specific blacklists or redlists will be punished or rewarded under the rules defined in Memoranda of Understandings (MoUs). Different government agencies have jointly signed and started enforcing these MoUs [8].

To summarize, the government-run SCS operates blacklists and redlists throughout the entire country. It enforces regulations with reputational and material means and requires mandatory participation. *This* SCS has regulatory "teeth". However, no research has conducted an empirical analysis of this nationwide SCS blacklist and redlist infrastructure.

This lack of knowledge is troubling, as the SCS will likely shape the behavior of about 1.4 billion Chinese citizens and all companies doing business in China. Further, important international long-term technology policy challenges are dependent on the success of systems such as the SCS, as highlighted by Antony Blinken in his confirmation hearings, when he argued that "*whether techno democracies or techno autocracies are the ones who get to define how tech is used (…) will go a long way toward shaping the next decades*" (2021 U.S. Secretary of State confirmation hearings [11]).

This study investigates the design and technical implementations as well as the number and types of blacklists and redlists across 30 Chinese provincial-level ADs. Our exploratory study shows the diversity of SCS lists in granular detail and outlines the informational consistency between social credit records of the same type of list on different SCS platforms. We find that SCS listings focus on economic activities but also capture reputational rewards for moral and political behavior. Moreover, we show that the SCS listing infrastructure is flexible, as observed in a second round of data collection during the COVID-19 outbreak: when necessary, new types of lists can regulate novel forms of transgression and thereby help accomplish new policy goals.

2 STUDY PROCEDURE

2.1 Policy-making in China: Provinces implement blacklists and redlists

SCS implementation is largely left to regional rather than central government, a common trait of China's policy-making process that tends to follow a principle of "centralized planning, decentralized implementation" [12, 13]. As a planning polity, central policy-makers outline policy goals in top-level policy documents valid for a specific policy-making cycle. Commonly, a first policy document (called *jianyi*/建议) includes general guidelines for a new cycle of policy-making. A second, more refined, but still broad, policy outline (called *gangyao*/纲要) sets more specific policy goals [14].[2] Importantly, the *implementation* of the policy goals outlined in top-level policy documents is left to provincial, county, and city governments. This also applies to the SCS: provincial-level administrative authorities (i.e., those in charge of provinces, municipalities under the direct administration of central government, and autonomous regions) are, to some extent, free to determine *how* they implement nationwide policy goals for their AD [27, 31].

The SCS's *gangyao* includes vague instructions regarding social credit record applications for broadly defined commercial and social sectors (e.g., [6, 8, 22]). SCS implementation rests on the commitment of provincial-level ADs[3] to realize general instructions laid out in top-level policy documents. As such, understanding the nationwide SCS listing infrastructure requires an empirical assessment of all SCS platforms at the provincial level. As each province is responsible for the implementation of its own SCS blacklist and redlist, we expected to find differences in the technological setup, interface design, and list types (i.e., differences in types of rewards and sanctions) between the provincial-level SCS platforms.

We conducted two rounds of data collection. First, between June 2019 and December 2019, we collected data on blacklists and redlists from 30 Chinese provincial-level ADs comprised of 22 provinces, 5 autonomous regions and 4 municipalities under the direct administration of central government. Second, in February 2020, we started collecting data on blacklists and redlists related to the coronavirus outbreak.

As we describe in more detail in the methodology section, our study approach is fundamentally *exploratory*. Data collection and analyses were intended to understand SCS implementation with regard to three high-level research questions, as follows.

[1] The authors use quotation marks to communicate a neutral standpoint towards SCS-specific normative concepts (e.g., "positive", "negative", "reward", "sanction/punishment"). For the remainder of the article, quotation marks will be omitted for the sake of reader-friendliness.

[2] Generally, policy-making in China is accompanied by a multitude of other policy documents. Engaging in a comprehensive description of Chinese policy-making would go beyond the scope of this study.

[3] In China, provincial-level ADs comprise provinces (e.g., Sichuan), municipalities under the direct administration of central government (e.g., Beijing, Shanghai) and autonomous regions (e.g., Inner Mongolia, Tibet).

Figure 1: Screenshot of an overview of the SCS information platforms of the different ADs listed on the national SCS platform "creditchina.gov.cn". Taiwan, Hong Kong and Macao were previously listed together with other ADs on the landing page of the "Credit China" website, but without a valid link. The listings were then removed in July 2019. Data collection was conducted via the SCS platform of each AD. Color-coding: orange represents municipality under the direct administration of central government; blue represents provinces; purple represents autonomous administrative regions; green represents the Xinjiang production and construction corps (Bingtuan), an economic and paramilitary organization in the Xinjiang Uyghur Autonomous Region, which is not included in our analysis due to an insignificant amount of credit data. Translations of AD names added by the authors.

- RQ1: Are there technological and design differences in credit lists and records between the provincial SCS platforms?
- RQ2: How do provincial SCS platforms differ in the number and types of blacklists and redlists?
- RQ3: How do SCS blacklist and redlist *records* of the same type of list differ in terms of the information displayed across provincial SCS platforms?

2.2 Methodological approach

2.2.1 Data. Our analysis pertains to blacklists and redlists implemented at the AD level from June 2019 to December 2019. Data collection was aimed at provincial-level blacklists and redlists from 31 ADs (22 provinces, 5 autonomous regions, 4 municipalities under the direct administration of central government) listed on China's national SCS platform "creditchina.gov.cn" (Figure 1).[4] For the follow-up study of coronavirus-related lists, we inspected the same SCS platforms again between February 2020 and April 2020.

Data collection primarily refers to a) the types of lists implemented in each AD (RQ2) and b) retrieving individual credit records from the most commonly implemented blacklist and redlist across all 31 ADs (RQ3). Collecting list types and credit records enabled an analysis of the technical realization and interface designs of SCS platforms and credit records (RQ1).

Our data collection was organized to produce a *descriptive* study of SCS implementation. Our core analyses focused on the diversity of list types across ADs and the structural differences between list records, in particular, their interface designs and the information provided in individual credit records. For several reasons, we did not conduct a quantitative analyses on published records. First, during data collection, we observed that the number of published SCS records changed on a day-to-day basis for all SCS platforms. We refrained from drawing general inferences on SCS credit records based on a onetime quantitative analysis. Second, when we began to scrutinize different SCS platforms, we observed large differences in the amount of credit records uploaded. Some SCS platforms had not published any credit records, while some displayed multiple millions (note that only a few SCS platforms indicated the total number of credit records). Third, given the early stage of SCS development, a comprehensive quantitative analysis of the economic and societal impacts of credit records was not possible at the time of data collection. This impact may need several years to materialize as SCS measures begin to influence the economy, government administration, and social processes at large. Fourth, as we discuss in the next subsection, we encountered challenges in accessing and retrieving public credit information from SCS platforms.

2.2.2 Data collection obstacles. The first obstacle was obtaining access to the 31 AD SCS platforms. Access from our location was severely impeded, so we tested the accessibility of different SCS websites from various locations. To accomplish this, we sent web requests from 44 servers spread around the world to each AD's SCS website.[5] SCS server accessibility from outside China was generally possible but unstable.[6] To investigate SCS platforms, we used a virtual private network of servers located in China. Requests from China provided more stable access to SCS servers than from other locations. All SCS servers, apart from the SCS server of the municipality of Chongqing, responded to requests from a Chinese server. For the server of the municipality Chongqing, no data could be retrieved at any time, as the server did not respond to requests for the entire data collection period from any location. Thus, our final data collection represented 30 ADs. Overall, it took 6 months

[4]This list also included the Xinjiang production and construction corps (Bingtuan). However, we did not include these data in our analysis for two reasons: first, Bingtuan is a unique state-owned economic and paramilitary organization in Xinjiang and, second, at the time of data collection, Bingtuan's SCS platform had published only a very small amount of credit information (9 blacklist and 7 redlists entries).

[5]The analysis was conducted with the Uptrends online monitoring service (www.uptrends.com). Data available from the authors.

[6]The most frequent return values were: HTTP connection failure, HTTP protocol error, HTTP timeout, and TCP connection failure.

to access all SCS platforms and to document the different types of blacklists and redlists, verify them through revisits, and collect credit records for each AD.

While documenting the different types of lists for each province, we observed that each AD operated a different web server with different implementations of front-end, back-end and database design. Moreover, we did not find a public API on any of the AD SCS platforms. Taken together, this made data collection for *credit records* complicated, as each AD SCS platform required the programming of a unique web crawler and scraper.

The systematic sampling of public credit records from each blacklist and redlist on all SCS platforms was not possible for several reasons. First, the number and therefore types of lists implemented varied between the ADs. Some ADs had more than 10 types of lists, while others only displayed a single list (see Results). We saw that some ADs with only a single implemented blacklist or redlist used this list to present different types of sanctions or rewards. Second, some ADs had only one list but no records to show at all. Third, SCS platforms differed in how credit records were displayed. For example, some SCS platforms displayed a number of credit records on a single page and offered page tabs that opened the next page, displaying the next set of credit records. This interface style allowed page visitors to go through all available credit records. Other SCS platforms only showed a selection of credit records and instead of page tabs provided a search bar for specific queries. Here, visitors could not see all available credit records. Finally, some AD SCS platforms deployed captchas and bot blockers that sometimes led to time-out denials such as temporary or even permanent IP address suspension.

Given these restrictions on the collection of credit records, systematic and unbiased sampling of credit records across all SCS platforms was not possible. However, the goal of our study was not to measure effects between credit record samples to generalize to the SCS as a single system. Instead, for the credit record analysis, our research goal was to explore informational differences in credit records across the SCS platforms. For this purpose, homogeneous convenience sampling was sufficient to compare the information provided on credit records on the same list between SCS platforms. Homogeneous convenience sampling differs from conventional convenience sampling by constraining sampling by one factor (see e.g., [17]). We did not sample any credit record on any type of list (i.e., we did not conduct conventional convenience sampling). We directed the analysis of credit records toward the most frequently implemented type of blacklist and redlist across all SCS platforms. Consequently, different crawling and data extraction (scraping) robots were programmed to extract pre-specified information on credit records from the most common type of blacklist and redlist.[7] The two main frameworks and tools used for the crawling and scraping process were ThoughtWorks Limited open source headless browser Selenium and Scrapinghub Limited open source framework called Scrapy. The extracted data were eventually pushed into a noSQL database (MongoDB) as a horizontally scaling non-relational database was the better solution given the different SCS platform implementations.

[7]We provide a code example of a crawler and a spider in the Auxiliary Material.

Figure 2: Shanghai's "Dishonest legal persons subjected to enforcement" (Lao Lai) blacklist of companies only displayed 10 record entries, requiring visitors to make a targeted search query. Translations by the authors.

Finally, the obstacles described above naturally led to credit record samples of varying size. On some SCS platforms, we managed to retrieve thousands of public credit records. On other platforms we obtained less than a hundred; some platforms did not have *any* credit records at all during the entire data collection period (for an overview of sampling results, see Table 2 in the Auxiliary Material). The differences in sample size were not due to any systematic sampling error committed by us but reflected the arbitrariness of the credit record display across the SCS platforms during the data collection period.

3 RESULTS

3.1 Technical implementation and design of blacklists and redlists

Each SCS platform operated a different web server with its own front-end, back-end and database design. We observed that the designs of the blacklists and redlists differed between ADs but was, overall, simple and plain.

All SCS platforms implemented either a Hypertext Markup Language (HTML) document with classic Cascading Style Sheet (CSS) structure or advanced dynamic scripting technology (JavaScript) for lists and individual records.

The majority of ADs (21) displayed only a selection of records but enabled targeted queries via a search bar. The remaining ADs

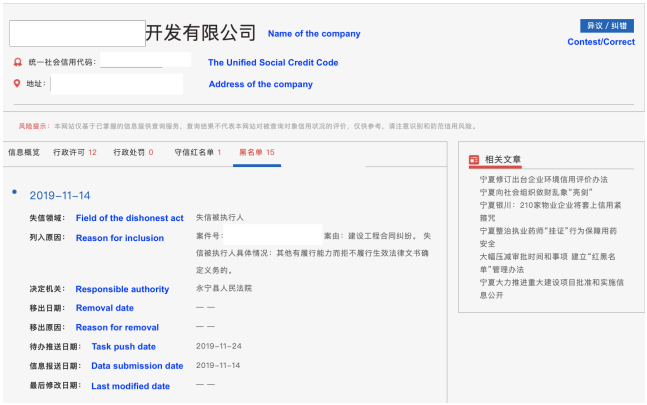


Figure 3: A two-column example credit record of the "Lao Lai" blacklist published on Ningxia's SCS platform. Translations by the authors.

Figure 4: A multi-column example record of Jiangxi's "Lao Lai" blacklist (失信被执行人名单). Translations by the authors.

showed all available social credit records with the help of a page tab. For example, on Guangxi's SCS platform, blacklist records could be accessed via 6852 tabs, each displaying 10 records. By contrast, Shanghai's blacklists showed ten blacklist records with no option to access more entries other than with a targeted query (Figure 2).

The design differences extended to individual credit records. Blacklist and redlist records were either structured as two column tables (Figure 3), multiple column tables (Figure 4) or continuous text documents.

Inner Mongolia and Shandong enabled sharing of blacklist and redlist records through Chinese social media platforms (e.g., Wechat, Sina Weibo, and Baidu Tieba). We found that eight SCS platforms offered citizens and organizations the possibility to contest published social credit records via a standardized interface option (e.g., Figure 3 top right corner). Our data indicate that there are technological and design differences in credit lists and records between provincial SCS platforms (RQ1). The current design and implementation of SCS platforms prioritize the display of social credit records rather than any aspect of their reputational effects. All SCS platforms had a binary rating system for good and bad behaviors – redlists and blacklists. Other than this binary classification, however, ADs did not apply other rating measures, such as numerical or continuous scoring. Indeed, we did not observe any social credit score at all communicated on any provincial-level SCS platform across China. Different types of lists were not put into relation with each other by means of a sorting or ranking. For example, no system of reputational ordering was found between individual records that highlighted severe transgressions more prominently than less severe cases. Five ADs showed numerical aggregation when a citizen or company had multiple social credit records. Entities with additional record entries were not displayed more prominently than entities that had a single credit record entry. Currently, the design of the SCS lists serves as a digitally accessible repository for citizen and company records and does not use any advanced features characteristic of other digital reputation systems [25].

3.2 Diversity and comprehensiveness: Number and types of blacklists and redlists

In response to RQ2, our data provide evidence for substantial differences in the number and types of lists between ADs (compare Figures 5 & 6). This confirms that regional governments determine the number and types of blacklists and redlists for their administrative region. For example, Beijing, Tianjin, Tibet, Guangdong, Hunan, Shanxi and Qinghai each operated more than ten different types of blacklists and redlists. In contrast, Inner Mongolia, Ningxia, Gansu, Guizhou, and Hebei each had implemented only one blacklist *and* one redlist. At present, it is impossible to say why some ADs run multiple lists and some only a single list. The number of lists did not correlate with economic, demographic, or geographic factors (data not shown).

In total, more blacklists (273) were published than redlists (154). We first grouped the 273 blacklists into 41 categories and the 154 redlists into 45 categories. We then created a taxonomy consisting of eight types of blacklists and eight types of redlists that currently make up the entire SCS AD listing infrastructure (Table 1). Note that different types of lists emphasize compliance with the legal and social norms that an AD wants to improve on. Thereby, the SCS influences behavior through two common reputation strategies [2]. With a minimum threshold strategy, blacklisting stresses the need for conformism. This technique tries to bring all entities to the same level of compliance. Redlisting, on the other hand, highlights praiseworthy performers that are intended to serve as behavioral role models.

The majority of blacklists displayed companies and citizens that have not fulfilled a court order, have committed commercial or transactional fraud, or have not complied with specific industry regulations. *All* ADs had implemented a "List of Dishonest Persons subject to Enforcement" also called the "Lao Lai" blacklist. This

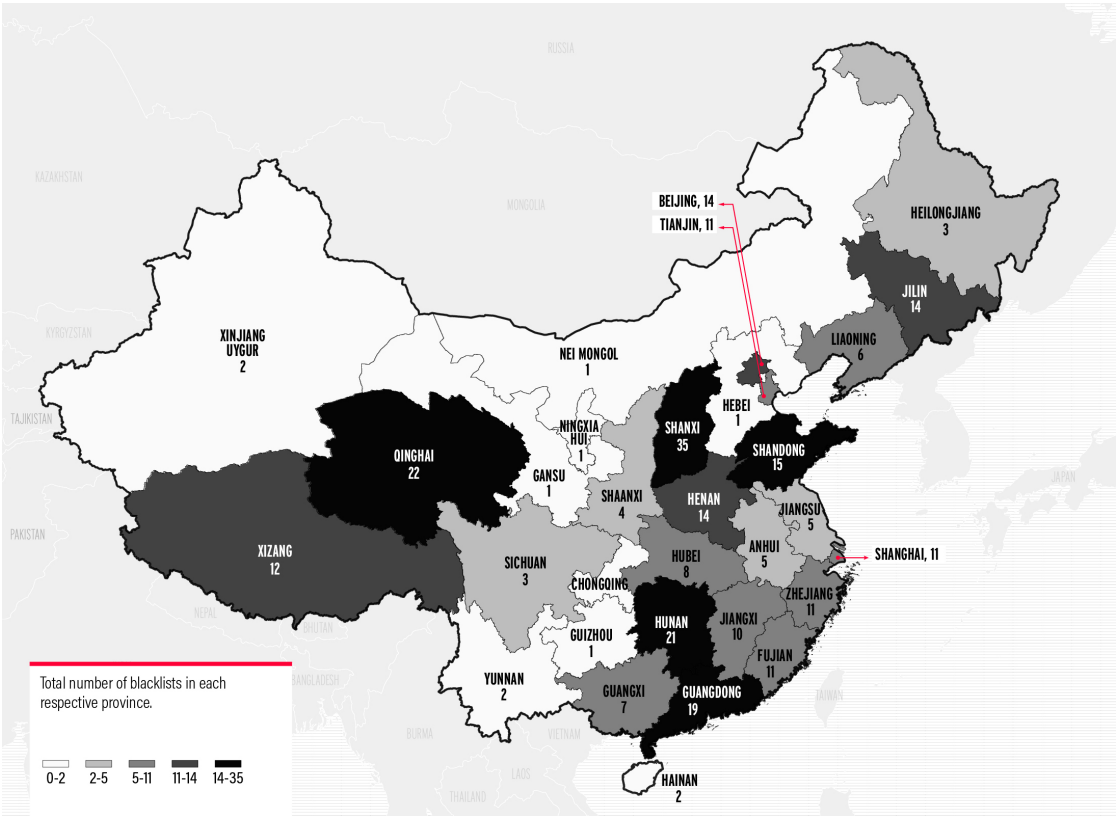


Figure 5: The number of blacklists implemented across 30 ADs. Shanxi had implemented most blacklists (35), followed by Qinghai (22), Hunan (21), Guangdong (19) and Shandong (15).

blacklist published information on citizens and companies that have failed to fulfill a court order. The "Lao Lai" blacklist aims to tackle China's court order enforcement problem [8, 9]. It forms a critical part of the JRP by which listed citizens face multiple restrictions, such as being banned from taking flights and high speed trains. Restrictions for "Lao Lai" companies include denial of licenses, reduced possibility to win bids for public contracts, or being subject to additional requirements for mandatory government approval for investments in sectors where market access is usually not regulated. Beyond the "Lao Lai" blacklist, we did not find any other type of blacklist implemented on all SCS platforms. The other types of blacklist most commonly found targeted non-compliance in tax payment (12 out of 30 ADs), untrustworthy behavior in financial activities (9/30), illegal import or export of products (8/30), delay or failure to compensate migrant[8] workers (8/30, companies only), or failure to protect the environment (7/30, companies only). We found blacklists that sanctioned fraud in marriage registrations or charity donations (social fraud), companies that had failed to comply with product quality standards (especially in food and drug production), or companies that had bad employment relationships.

The most frequently implemented redlists displayed entities that complied with tax law (18 out of 30 ADs) and import and export regulations (10/30). Usually, redlists serve to reward particularly "praiseworthy" behaviors. We made the surprising observation that many types of redlists highlighted regular compliance with laws and regulations. Some redlists, however, showcased individuals and companies that distinguished themselves politically or morally. For example, Beijing's SCS platform published a list called "4th Beijing Excellent Builders of Socialism with Chinese Characteristics", and Jiangxi and Tianjin listed citizens that had been rewarded the "May Fourth Medal". Tianjin had implemented two lists titled "Tianjin Good Man" and "Tianjin Ideological and Moral Model". Tibet had a similar redlist called "Moral Models & Good Political Ideology" (Figure 7). Other redlists were dedicated to citizens that had volunteered, given to charity or won awards in education, science or technology. Overall, the redlist infrastructure was less elaborate than its blacklist counterpart: not a single type of redlist existed in all ADs. Three ADs had published a single redlist with no data (Xinjiang, Gansu, and Jilin).

3.3 Informational consistency on credit records of the most common blacklist and redlist

To address RQ3, we explored the informational differences among the credit records of the most frequently implemented types of lists: the "Lao Lai" list (blacklist) and the "Class A Taxpayer" list

[8] "Migrant" here refers to rural citizens moving into urban centers for employment.

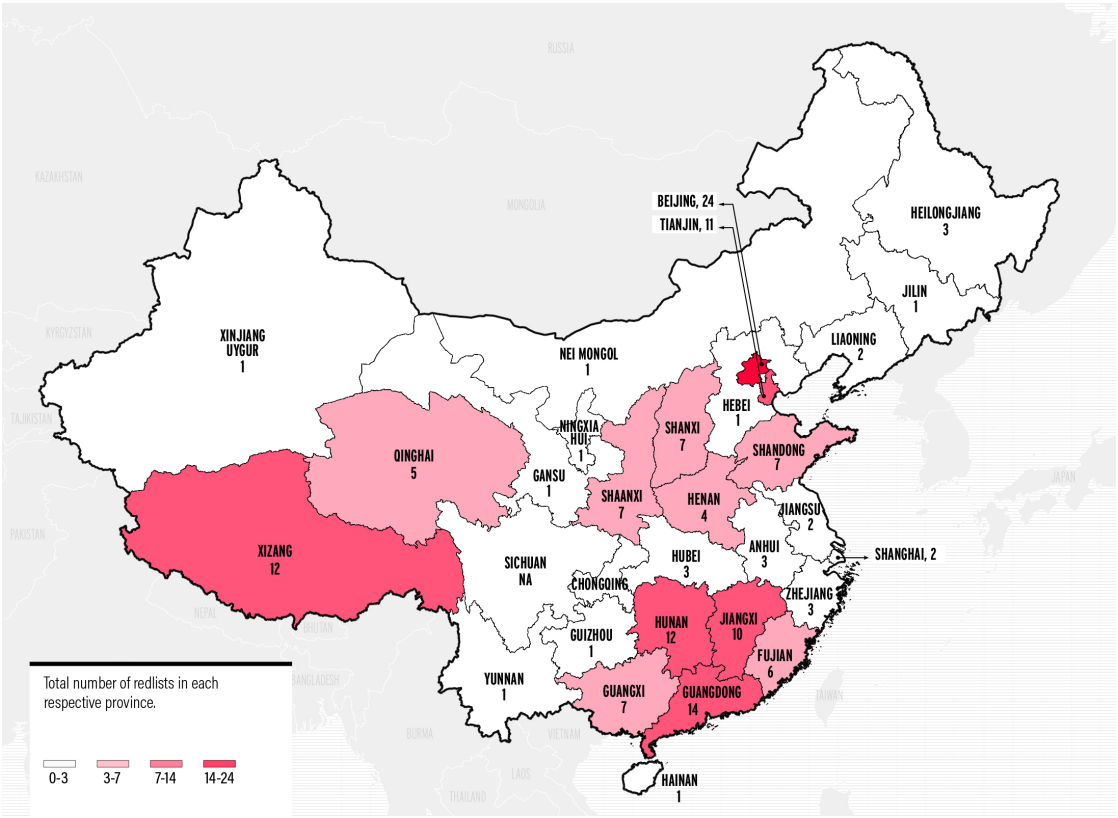


Figure 6: The number of redlists implemented across 30 ADs. Beijing had implemented the most redlists (24), followed by Guangdong (14), Xinjiang (12), Hunan (12), Tianjing (11), and Jiangxi (10).

(redlist). With the exception of Jilin and Tibet, the remaining 28 ADs had published credit records in their "Lao Lai" lists. We compared ADs based on the provision of five types of information in "Lao Lai" credit records: 1) the unified social credit code (companies) or identification number (natural persons), 2) specification of a data source or responsible authority, 3) reasons for listing (i.e., a justification), 4) information on the fulfillment of the requirements, and 5) information on a future removal date of the record (see Figure 8).

3.3.1 Information on "Lao Lai" blacklist credit records. Based on the samples of credit records obtained, out of the 28 different ADs, only 14 ADs had provided either the unified social credit code (8/28) or the natural person's identification number (6/28). The remaining ADs either listed an organization code (3/28) for companies or simply the name of the natural person listed (3/28). 23 ADs specified the data source of the record (i.e., where the data had been generated), the name of the executive court (12/28) or a responsible agency. In all, 24 ADs provided at least some explanation for why an entity had been listed. In the majority of cases, the credit records referred to a specific law that was to be enforced. Finally, 12 ADs indicated whether the requirement had already been fulfilled or not, and only 6 ADs displayed the removal date of the record.

3.3.2 Information on "Class A Taxpayer" redlist credit records (including unspecified redlists). For ADs without a "Class A Taxpayer" list, we inspected records from the only list available. 25 ADs had provided redlist records on their SCS platforms. 17 ADs had explicitly used the term "unified social credit code" in their records, and 7 listed a "taxpayer identification number". The remaining ADs simply presented the name of the listed entity. All ADs that published redlist records provided some form of identifying information. Of these, 21 ADs indicated the responsible authority for the case in question, and 16 ADs included a justification for being listed (commonly termed "reason for inclusion" or "honor content"). 6 ADs indicated the record's expiration date. An example record of a Class A Taxpayer List is shown in Figure 9.

3.4 Flexibility: Blacklists and redlists regulate behavior during the COVID-19 epidemic

Finally, we found that novel types of norm transgression can be quickly subjected to blacklisting and redlisting. Between February 27 and March 30, 2020, we collected data from the same SCS platforms to understand whether blacklisting and redlisting were used to regulate social behavior in an exceptional state of emergency. During this second round of data collection, we had access to 25 of

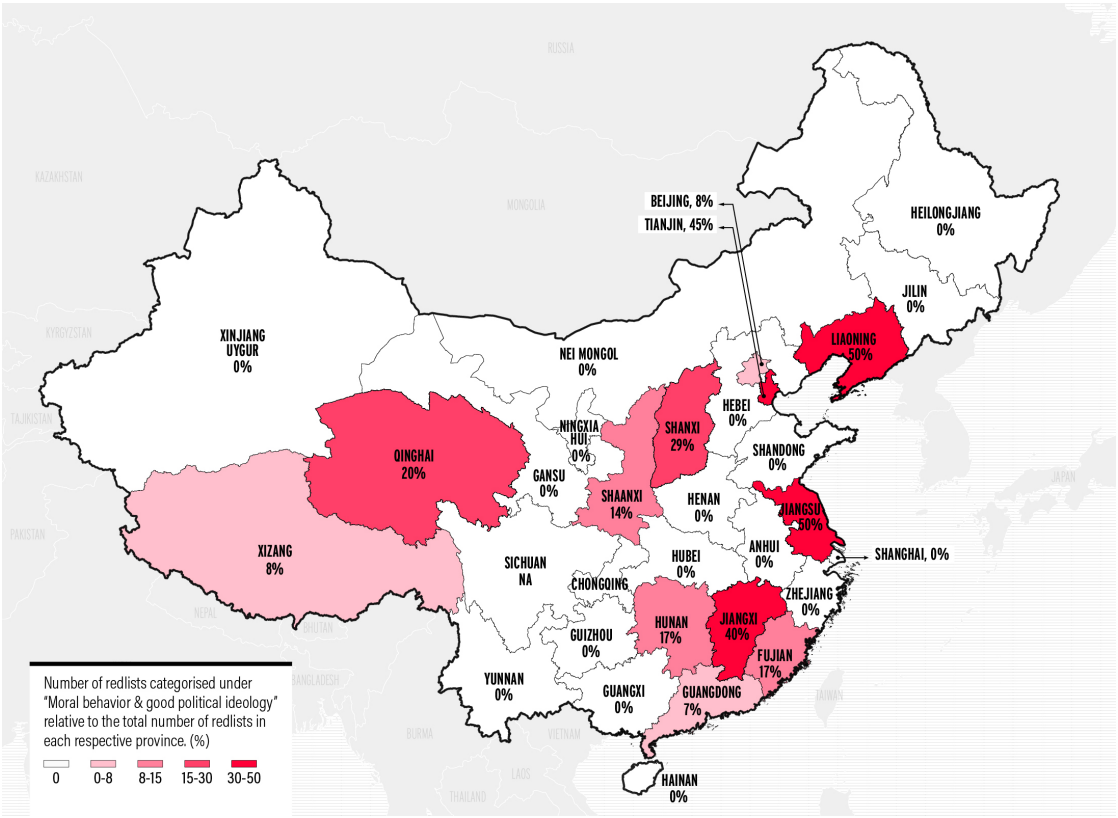


Figure 7: Ratios of redlists for moral behavior and good political ideology to total redlists across the 30 listed Chinese ADs.

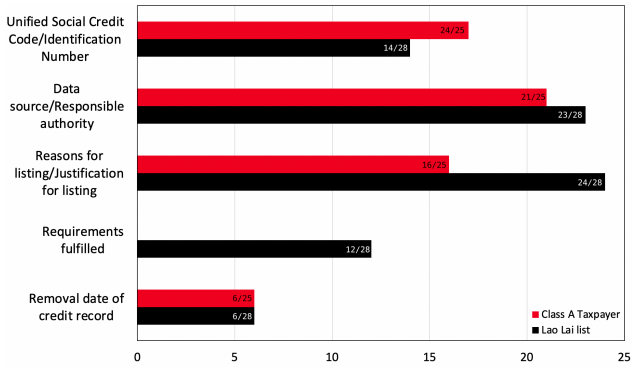


Figure 8: A comparison of the information provided on credit records collected from the most frequently implemented type of blacklist and redlist across all ADs.

the 31 ADs.[9] We identified coronavirus-related blacklists in 15 ADs and redlists in 10 ADs. Pursuant to our first analyses, blacklist and redlist records targeted natural persons and companies. We found that coronavirus blacklists included entities for selling fake preventive health products, violating quarantine regulations, organizing or participating in gatherings during lockdown, or illegally operating transport vehicles as ambulances. Blacklists were presented in different formats across the 15 ADs: they were either given in a row-and-column format (5) or in narrative-like news reports (10) (see Figure 10). Coronavirus redlists reported on devoted professionals such as doctors, nurses, volunteers, and border control officials, as well as on companies and individuals that had donated health products. All coronavirus redlist records were presented as narrative news reports.

4 SUMMARY AND CONCLUDING ANALYSIS

We conducted an empirical investigation on the diversity, flexibility, and comprehensiveness of provincial-level SCS blacklists and redlists in China.

Overall, we highlighted that SCS listing designs facilitate public access to social credit records. The majority of SCS platforms display a selection of credit records and enable targeted queries. SCS platforms serve as digital reputation systems because redlists and blacklists digitally showcase entities' good and bad behaviors. However, with the exception of a few ADs that aggregated credit records for a single entity or allowed sharing of credit records to

[9]We did not have access to the SCS platforms of Jilin, Beijing, Fujian, Qinghai, Chongqing, and Hainan.

Types of blacklists	Blacklists of untrustworthy entities (general)	Blacklists of commercial and transactional activities	Other blacklists	Employment relationship blacklists	Product quality blacklists	Blacklists of financial fraud	Industry blacklists	Blacklists of social fraud	Coronavirus blacklists
Beijing	2	4	3	0	1	4	0	0	NA
Shanghai	4	2	2	0	0	2	1	0	1
Tianjin	6	1	0	1	0	3	0	0	1
Hebei	1	0	0	0	0	0	0	0	1
Shanxi	5	5	5	1	3	2	12	2	1
Liaoning	2	1	3	0	0	0	0	0	0
Jilin	1	6	3	1	2	1	0	0	NA
Heilongjiang	3	0	0	0	0	0	0	0	1
Shandong	2	6	2	1	1	2	1	0	1
Jiangsu	2	0	2	0	0	1	0	0	1
Zhejiang	3	4	2	0	1	1	0	0	1
Anhui	4	0	1	0	0	0	0	0	1
Fujian	3	2	2	2	0	1	1	0	NA
Jiangxi	2	2	1	1	0	2	2	0	1
Henan	5	4	1	0	1	1	2	0	1
Hubei	2	5	0	0	1	0	0	0	0
Hunan	5	5	3	1	4	0	3	0	0
Sichuan	2	1	0	0	0	0	0	0	0
Guangdong	1	5	5	4	2	1	1	0	0
Gansu	1	0	0	0	0	0	0	0	0
Hainan	2	0	0	0	0	0	0	0	NA
Qinghai	3	4	2	1	3	2	7	0	NA
Guizhou	1	0	0	0	0	0	0	0	1
Yunnan	2	0	0	0	0	0	0	0	1
Shaanxi	1	1	1	1	0	0	0	0	1
Tibet	1	4	0	6	0	0	1	0	0
Inner Mongolia	1	0	0	0	0	0	0	0	1
Guangxi	3	1	1	2	0	0	0	0	0
Ningxia	1	0	0	0	0	0	0	0	0
Xinjiang	2	0	0	0	0	0	0	0	0

Types of redlists	Redlists of trustworthy entities (general)	Redlists of commercial and transactional activities	Redlists of moral behavior & good political ideology	Employment & customer relationship redlists	Product quality redlists	Industry redlists & redlists of professions	Redlists of intellectual property & awards	Other redlists	Coronavirus redlists
Beijing	1	5	2	2	2	4	8	0	NA
Shanghai	1	1	0	0	0	0	0	0	0
Tianjin	0	3	5	0	0	0	3	0	0
Hebei	1	0	0	0	0	0	0	0	1
Shanxi	0	3	2	0	0	1	0	1	0
Liaoning	0	1	1	0	0	0	0	0	0
Jilin	1	0	0	0	0	0	0	0	NA
Heilongjiang	1	2	0	0	0	0	0	0	1
Shandong	1	3	0	0	1	1	0	1	1
Jiangsu	1	0	1	0	0	0	0	0	0
Zhejiang	2	1	0	0	0	0	0	0	1
Anhui	0	3	0	0	0	0	0	0	1
Fujian	1	3	1	1	0	0	0	0	NA
Jiangxi	0	2	4	0	0	3	1	0	0
Henan	0	3	0	0	0	0	1	0	1
Hubei	1	2	0	0	0	0	0	0	0
Hunan	1	2	2	0	0	3	4	0	0
Sichuan	0	0	0	0	0	0	0	0	0
Guangdong	0	7	1	0	0	1	5	0	0
Gansu	1	0	0	0	0	0	0	0	1
Hainan	1	0	0	0	0	0	0	0	NA
Qinghai	0	2	1	0	0	1	1	0	NA
Guizhou	1	0	0	0	0	0	0	0	1
Yunnan	1	0	0	0	0	0	0	0	1
Shaanxi	1	3	1	0	0	2	0	0	1
Tibet	0	4	1	0	2	2	1	2	0
Inner Mongolia	1	0	0	0	0	0	0	0	0
Guangxi	0	1	0	0	0	0	6	0	0
Ningxia	1	0	0	0	0	0	0	0	0
Xinjiang	1	0	0	0	0	0	0	0	0

Table 1: The different types of blacklists and redlists implemented by ADs in China. Shading indicates the number of blacklists or redlists for a given type. N/A denotes no access to the SCS platform.

Figure 9: A screenshot of a redlist record from the "Class A Taxpayer List" published on the Fujian SCS platform. Translations by the authors.

social media platforms, we did not observe any automated classification, ranking or scoring on any of the current SCS listings.

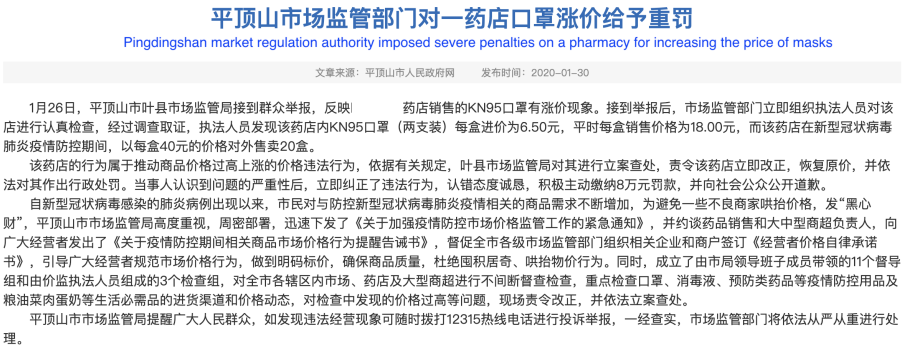
The SCS comprises hundreds of blacklists and redlists across provincial-level ADs. Currently, the majority of these types of lists target compliance with a wide range of laws and regulations. Thereby, SCS blacklists focus on "Lao Lai" entities, which are citizens and companies that have not fulfilled a court order. The SCS first displays "Lao Lai" on its digital listings and hence excludes them from future cooperative opportunities through its JRP. Based on these two mechanisms, the SCS seeks to turn "Lao Lai" into cooperators by attaching an exceptionally high cost to defection. We also observed redlists that highlight praiseworthy political and moral behaviors. Further development of lists that go well beyond legal or regulatory norms could substantially increase the social control characteristics of the SCS.

We have exemplified the flexibility of SCS listings by a case study on the COVID-19 outbreak. Digital blacklists and redlists might be a particularly powerful regulatory measure because they can be adapted to help accomplish novel policy goals quickly and at relatively low costs.

There are several outstanding questions for future research. For example, will SCS platform design incorporate more reputational affordances? Will the governmental and commercial branches (i.e., big data apps) of the SCS cooperate to share and analyze different data streams? Will SCS mechanisms really produce their intended regulatory effects? We believe that asking such questions is crucial and we hope to have laid a useful foundation for future empirical and conceptual studies on the SCS.

5 ETHICAL DIMENSIONS OF THE STUDY

We now turn to initial ethical considerations of data-driven research on SCS implementation. First, our analysis was based on publicly available data found on key platforms of China's SCS. These data are posted to enable public scrutiny. Our paper includes screenshots from the currently available implementations (see Figures 1, 2, 3, 4, 9, 10). Our data collection and analyses are privacy-preserving: we blurred any personally identifiable data to protect the privacy of



平顶山市场监管部门对一药店口罩涨价给予重罚

Pingdingshan market regulation authority imposed severe penalties on a pharmacy for increasing the price of masks

文章来源：平顶山市人民政府网　　发布时间：2020-01-30

1月26日，平顶山市叶县市场监管局接到群众举报，反映[redacted]药店销售的KN95口罩有涨价现象。接到举报后，市场监管部门立即组织执法人员对该店进行认真检查，经过调查取证，执法人员发现该药店内KN95口罩（两支装）每盒进价为6.50元，平时每盒销售价格为18.00元，而该药店在新型冠状病毒肺炎疫情防控期间，以每盒40元的价格对外售卖20盒。

该药店的行为属于推动商品价格过高上涨的价格违法行为，依据有关规定，叶县市场监管局对其进行立案查处，责令该药店立即改正，恢复原价，并依法对其作出行政处罚。当事人认识到问题的严重性后，立即纠正了违法行为，认错态度诚恳，积极主动缴纳8万元罚款，并向社会公众公开道歉。

自新型冠状病毒感染的肺炎病例出现以来，市民对与防控新型冠状病毒肺炎疫情相关的商品需求不断增加，为避免一些不良商家哄抬价格，发“黑心财”，平顶山市市场监管局高度重视，周密部署，迅速下发了《关于加强疫情防控市场价格监管工作的紧急通知》，并约谈药品销售和大中型商超负责人，向广大经营者发出了《关于疫情防控期间相关商品市场价格行为提醒告诫书》，督促全市各级市场监管部门组织相关企业和商户签订《经营者价格自律承诺书》，引导广大经营者规范市场价格行为，做到明码标价，确保商品质量，杜绝囤积居奇、哄抬物价行为。同时，成立了由市局领导班子成员带领的11个督导组和由价监执法人员组成的3个检查组，对全市各辖区内市场、药店及大型商超进行不间断督查检查，重点检查口罩、消毒液、预防类药品等疫情防控用品及粮油菜肉蛋奶等生活必需品的进货渠道和价格动态，对检查中发现的价格过高等问题，现场责令改正，并依法立案查处。

平顶山市市场监管局提醒广大人民群众，如发现违法经营现象可随时拨打12315热线电话进行投诉举报，一经查实，市场监管部门将依法从严从重进行处理。

Figure 10: Screenshot of the coronavirus blacklist from the SCS platform for Henan province. Translation: *On January 26, the Market Supervisory Authority of Ye County Pingdingshan City received reports from the public reporting that ** Pharmacy increased the price of KN95 masks. After receiving the report, the authority immediately sent out law enforcement officers to conduct a serious inspection of the store and found that the purchase price of the KN95 masks (2 pieces in one package) was 6.5 RMB for the store and the sale price was usually 18 RMB. However, the pharmacy sold 20 packages of the masks at the price of 40 RMB during the epidemic period. The pharmacy was thus in violation of the price regulation. Following relevant regulations, the Market Supervisory Authority filed a case for the investigation and ordered the pharmacy to restore the price to its original level. The authority also imposed administrative penalties on the pharmacy according to law. The pharmacy realized the seriousness of the problem and immediately halted the illegal behavior, admitted its misconduct, proactively paid a fine of 80,000 RMB, and apologized to the public.* Translations by the authors.

listed companies and citizens. Our methodological approach does not result in any unfavorable consequences or costs for any of the data subjects. We are transparent in our methodology and provide a representative code example of a web crawler and spider we used in this study (see Auxiliary Material).

Second, our account adheres to the principles of ethical web crawling and scraping [20, 29, 33, 34]. For each SCS platform, we checked for a specified *robots.txt* file. At no point during our data collection did we find a robots.txt file that specified rules for web crawlers. Accordingly, when platforms make data publicly available, do not specify a robots.txt file, and do not provide a data collection interface (e.g., API), then robots are free to gather data (see, e.g., [29, 33]).

Third, the purpose of our study is ethically justifiable on its own. In the absence of systematic empirical accounts, uncertainty will inevitably help foster misconceptions about the SCS (whether overly positive or negative). Given China's geopolitical prominence, governments of other countries may be inspired to copy China's SCS [24]. This is particularly likely for neighboring countries [37]. Data-driven research on SCS implementation can help prevent hasty SCS adaptations by other governments based on false assumptions. Empirical and conceptual analyses on the SCS allow for a more informed public debate about the development of digital socio-technical systems. As our data indicate, *currently*, there is little evidence that blacklists and redlists operate as AI-driven reputation systems. Apart from two SCS platforms that enable sharing of credit records to social media platforms, at the moment, there is no evidence that credit records are subjected to other means of digital reputation mechanisms such as classification, ranking, or profiling based on AI. It is possible that future developments might implement AI-based reputation mechanisms. As we have argued, additional empirical work on the SCS is necessary given that Chinese policy-making rests on often vaguely formulated policy goals. We show a considerable diversity of SCS blacklist and redlist implementation that cannot be concluded from policy analysis alone. Our study raises important questions that also matter for non-Chinese citizens and organizations. For example, is stable access to blacklists and redlists from outside China justifiable when non-Chinese citizens and companies are listed [4, 10]? Should China distribute licenses or special APIs to allow non-Chinese entities to ascertain whether they are listed? Or will Chinese authorities directly notify non-Chinese entities when they are listed?

The Chinese SCS is already one of the most comprehensive reputation systems in the world. Given that the government generates the reputation signals, we believe that SCS blacklisting and redlisting could have a strong influence on societal behavior at large.

Finally, this research extends growing calls for more open data in computational social science [21] with a case for more data availability *in China*. As this body of research has shown, open government data can significantly improve our understanding of societies' most important challenges in the context of equality, health, or employment. Even if data collection obstacles are likely to persist, we hope that our study underlines the importance of future data-driven research on the Chinese SCS.

ACKNOWLEDGMENTS

We thank Rogier Creemers, Bilge Kobas, and Marianne von Blomberg for their helpful input. We also thank the anonymous reviewers for their constructive comments. We gratefully acknowledge funding support from the Bavarian Research Institute for Digital Transformation (bidt). Mo Chen further thanks for the support through a postdoc research stipend of the Fritz Thyssen Foundation. Responsibility for the content of this publication rests with the authors.

REFERENCES

[1] Shazeda Ahmed. 2019. The messy truth about social credit. *Logic* 7 (2019).

[2] Geoffrey Brennan and Philip Pettit. 1993. Hands invisible and intangible. *Synthese* 94, 2 (1993), 191–225.

[3] Charlie Campbell. 2019. How China is using "Social Credit Scores" to reward and punish its citizens. *Time* (2019). Available at: https://time.com/collection/davos-2019/5502592/china-social-credit-score/.

[4] Mo Chen, Kristina Bogner, Joana Becheva, and Jens Grossklags. 2021. The transparency of the Chinese Social Credit System from the perspective of German organizations. In *Proceedings of the 29th European Conference on Information Systems (ECIS)*. Completed research paper.

[5] Mo Chen and Jens Grossklags. 2020. An empirical analysis of the commercial arm of the Chinese Social Credit System. *Proceedings on Privacy Enhancing Technologies* 4 (2020), 89–110.

[6] Yongxi Chen and Anne Cheung. 2017. The transparent self under big data profiling: Privacy and Chinese legislation on the Social Credit System. *The Journal of Comparative Law* 12, 2 (2017), 356–378.

[7] Martin Chorzempa, Paul Triolo, and Samm Sacks. 2018. *China's Social Credit System: A mark of progress or a threat to privacy?* Policy Briefs, Peterson Institute for International Economics No. PB18-14.

[8] Rogier Creemers. 2018. *China's Social Credit System: An evolving practice of control.* SSRN Working Paper Nr. 3175792.

[9] Xin Dai. 2018. *Toward a reputation state: The Social Credit System project of China.* SSRN Working Paper No. 3193577.

[10] Severin Engelmann, Mo Chen, Felix Fischer, Ching-Yu Kao, and Jens Grossklags. 2019. Clear sanctions, vague rewards: How China's Social Credit System currently defines "good" and "bad" behavior. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. 69–78.

[11] Karen Hao. 2021. The Biden administration's AI plans: what we might expect. *MIT Technology Review (22 January 2021)* (2021).

[12] Sebastian Heilmann. 2008. From local experiments to national policy: The origins of China's distinctive policy process. *The China Journal* 59 (2008), 1–30.

[13] Sebastian Heilmann. 2009. Maximum tinkering under uncertainty: Unorthodox lessons from China. *Modern China* 35, 4 (2009), 450–462.

[14] Sebastian Heilmann. 2018. *Red Swan: How Unorthodox Policy-Making Facilitated China's Rise.* Chinese University Press.

[15] Samantha R. Hoffman. 2017. *Programming China: The Communist Party's autonomic approach to managing state security.* Ph.D. Dissertation. University of Nottingham.

[16] Jamie Horsley. 2018. China's Orwellian social credit score isn't real. *Foreign Policy* 16 (2018).

[17] Justin Jager, Diane Putnick, and Marc Bornstein. 2017. II. More than just convenient: The scientific merits of homogeneous convenience samples. *Monographs of the Society for Research in Child Development* 82, 2 (2017), 13–30.

[18] Genia Kostka. 2019. China's Social Credit Systems and public opinion: Explaining high levels of approval. *New Media & Society* 21, 7 (2019), 1565–1593.

[19] Theresa Krause and Doris Fischer. 2020. An economic approach to China's Social Credit System. In *Social Credit Rating*. Springer, 437–453.

[20] Vlad Krotov and Leiser Silva. 2018. Legality and ethics of web scraping. In *Proceedings of the Twenty-fourth Americas Conference on Information Systems (AMCIS)*.

[21] David Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science* 369, 6507 (2020), 1060–1062.

[22] Fan Liang, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain. 2018. Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure. *Policy & Internet* 10, 4 (2018), 415–453.

[23] Alexandra Ma. 2018. China has started ranking citizens with a creepy 'Social Credit' system – Here's what you can do wrong, and the embarrassing, demeaning ways they can punish you. *Business Insider* 29 (2018).

[24] Daithí Mac Síthigh and Mathias Siems. 2019. The Chinese Social Credit System: A model for other countries? *The Modern Law Review* 82, 6 (2019), 1034–1071.

[25] Sergio Marti and Hector Garcia-Molina. 2006. Taxonomy of trust: Categorizing P2P reputation systems. *Computer Networks* 50, 4 (2006), 472–484.

[26] Louise Matsakis. 2019. How the West got China's Social Credit System wrong. *Wired (29 July 2019)* (2019).

[27] Andrew Mertha. 2009. "Fragmented authoritarianism 2.0": Political pluralization in the Chinese policy process. *The China Quarterly* 200 (2009), 995–1012.

[28] Simina Mistreanu. 2018. Life inside China's Social Credit Laboratory: The party's massive experiment in ranking and monitoring Chinese citizens has already started. *Foreign Policy* (2018). Available at: https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/.

[29] Zeina Mneimneh, Josh Pasek, Lisa Singh, Rachel Best, Leticia Bode, Elizabeth Bruch, Ceren Budak, Pamela Davis-Kean, Katharine Donato, Nicole Ellison, et al. 2021. *Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data.* Working Paper.

[30] Mareike Ohlberg, Shazeda Ahmed, and Bertram Lang. 2018. Central planning, local experiments: The complex implementation of China's Social Credit System. *Mercator Institute for China Studies* 43 (April 2018), 1–15.

[31] Gunter Schubert and Björn Alpermann. 2019. Studying the Chinese policy process in the era of 'top-level design': The contribution of 'political steering' theory. *Journal of Chinese Political Science* 24, 2 (2019), 199–224.

[32] State Council. 2014. Notice of the State Council on issuing the outline of the plan for building a Social Credit System (2014-2020); (in Chinese).

[33] Yang Sun, Isaac Councill, and Lee Giles. 2010. The ethicality of web crawlers. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 668–675.

[34] Mike Thelwall and David Stuart. 2006. Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology* 57, 13 (2006), 1771–1779.

[35] Alexander Trauth-Goik. 2019. "Constructing a culture of honesty and integrity": The evolution of China's Han-centric surveillance system. *IEEE Technology and Society Magazine* 38, 4 (2019), 75–81.

[36] Marianne von Blomberg. 2018. The Social Credit System and China's rule of law. In *Social Credit Rating*, Oliver Everling (Ed.). 111–137.

[37] Yau Tsz Yan. 2020. *Exporting China's Social Credit System to Central Asia.* Retrieved Apr 28, 2021 from https://thediplomat.com/2020/01/exporting-chinas-social-credit-system-to-central-asia/

A AUXILIARY MATERIAL FOR "BLACKLISTS AND REDLISTS IN THE CHINESE SOCIAL CREDIT SYSTEM: DIVERSITY, FLEXIBILITY, AND COMPREHENSIVENESS"

A.1 Documentation: Example crawler and spider for Guangdong province

The following code sections are an excerpt of the crawling and scraping methodology to systematically collect data from public blacklists and redlists of the Chinese Social Credit System. The crawler for collecting relevant data and the spider for extracting specific information from the data are demonstrated for the example of the Guangdong province below. Please note that the collection methodology may have to be adjusted, if the collection site is undergoing changes. You also may want to revisit the discussion on the ethics of data crawling in our paper (see Section 5).

Crawler example Guangdong province:
This section shows how the link lists are created, in particular, the methodology to collect the deep links that lead to the entry records of blacklists and redlists. A headless browser (like Selenium) is used, which is basically a normal web browser remotely controlled by a programmed robot.

In the following, an example of a web crawler is given:

```
class GuangdongSelenium():
    def crawl_red(self):
        link = 'https://credit.gd.gov.cn/opencreditAction!getOpencreditList_new.[...]&tbType=1'
        print_start("Guangdong␣Redlist")
        linkliste = []
        file = open("linklist_guangdong_red.txt", "a")

        driver.get(link)
        driver.find_element_by_css_selector('#newtype␣>␣option:nth-child(8)').click()
        driver.find_element_by_css_selector('label.search_button').click()

        while '下一页' in driver.page_source:
            try:
                categorylist = driver.find_elements_by_css_selector('tbody␣>␣tr:nth-child(1)␣>␣td␣>␣div␣>␣a')
                for i in categorylist:
                    print(i.get_attribute('href'))
                    s = i.get_attribute('href')
                    linkliste.append(s)
                driver.find_element_by_css_selector('a.next').click()
                time.sleep(10)
            except():
                print ("Error,␣no␣next␣page␣available!")
                break

        print("Length␣of␣final␣linklist:␣", len(linkliste))
        linkliste = list(dict.fromkeys(linkliste))
        print("This␣is␣the␣lenght␣of␣the␣list␣after␣removing␣all␣duplicates:␣", len(linkliste))
        for e in linkliste:
            file.write(e + "\n")

        print("Crawled␣links␣are␣written␣into␣the␣final␣file.")
        print("File␣created")
        file.close()
        driver.close()
        sys.exit()

    def crawl_black(self):
        link = 'https://credit.gd.gov.cn/opencreditAction!getOpencreditList_new.[...]&tbType=2'
        print_start("Guangdong␣Blacklist")
```

```
        linkliste = []
        file = open("linklist_guangdong_black.txt", "a")
        driver.get(link)
        driver.find_element_by_css_selector('#newtype > option:nth-child(2)').click()
        driver.find_element_by_css_selector('label.search_button').click()
        try:
            while '下一页' in driver.page_source:
                wait = WebDriverWait(driver, 10)
                wait.until(ec.visibility_of_element_located((By.CSS_SELECTOR, 'a.next')))
                time.sleep(10)
                categorylist = driver.find_elements_by_css_selector('tbody > tr:nth-child(1) > td > div > a')
                for i in categorylist:
                    print(i.get_attribute('href'))
                    s = i.get_attribute('href')
                    file.write(s + "\n")
                    linkliste.append(s)
                driver.find_element_by_css_selector('a.next').click()
                time.sleep(5)
        except:
            pass
            print("Error, no next page available!")
        print("File created")
        file.close()
        driver.close()
        sys.exit()
```

The desired output should be a collection of links stored in corresponding files 'linklist_guangdong_black.txt' or 'linklist_guangdong_red.txt'.

```
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=FF89EED12BC14E21BF36360E9044FC45
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=FF89EED12BC14E21BF36360E9044FC45
[...]
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=FF89EED12BC14E21BF36360E9044FC45
https://credit.gd.gov.cn/infoTypeAction!getAwardAndGruel.[...]id=FF89EED12BC14E21BF36360E9044FC45
```

Spider example Guangdong province:
This section shows a web scraping spider, a methodology that follows the web crawling process. A web scraper's task is to sequentially work through the web crawler's link list and extract specific data.

In the following, an example of a web scraper is given:

```
import scrapy, re

class GuangdongSpider(scrapy.Spider):
    name = "guangdong"
    file = open("linklist_guangdong_black.txt", "r")
    start_urls = [i.replace("\n", "") for i in file]

    def parse(self, response):
        table = response.css('table > tr > td')
        yield{
            'case_number' : table[1].css('::text').extract_first(),
            'lost_trustee_name' : table[3].css('::text').extract_first(),
            'gender' : table[5].css('::text').extract_first(),
            'age' : table[7].css('::text').extract_first(),
```

```
        'ID_number_desensitization__organization_code' : table[9].css('::text').extract_first(),
        'corporate_legal_person_name' : table[11].css('::text').extract_first(),
        'executive_court' : table[13].css('::text').extract_first(),
        'execution_basis_number' : table[15].css('::text').extract_first(),
        'basis_for_execution' : table[17].css('::text').extract_first(),
        'obligation_established_by_the_law' : table[19].css('::text').extract_first(),
        'implementation_of_the_person_being_executed' : table[21].css('::text').extract_first(),
        'untrustworthy_enforcer' : table[23].css('::text').extract_first(),
        'release_time' : table[25].css('::text').extract_first(),
        'filing_time' : table[27].css('::text').extract_first(),
        'fulfilled_part' : table[29].css('::text').extract_first(),
        'unfulfilled_part' : table[31].css('::text').extract_first(),
        'hyperlink' : response.url
    }
```

A.2 Table: Summary of credit record collection for blacklists and redlists

AD	No. of blacklist records	Avg. size blacklist record	No. of variables	No. of redlist records	Avg. size redlist record	No. of variables
Municipalities						
Beijing	100	1700 B	35	50	776.9 B	27
Shanghai	10	156.5 B	3	10	157.8 B	3
Tianjin	1501	1100 B	5	2000	306.6 B	5
AR						
Guangxi	30281	265.7 B	8	27692	547.5 B	15
Inner Mongolia	10	795.9 B	15	10	319.5 B	5
Ningxia	20	853.3 B	12	19	714.5 B	12
Xinjiang	3	1100 B	12	no data	-	-
Tibet	no data	-	-	no data	-	-
Provinces						
Anhui	190	926.5 B	15	190	315.8 B	6
Fujian	99	477.6 B	9	78	380.5 B	7
Gansu	20	1200 B	21	no data	-	-
Guangdong	160	1900 B	17	90	476.1 B	6
Guizhou	38	1600 B	6	39	2900 B	6
Hainan	40	817.3 B	17	40	654.6 B	13
Hebei	311	663.9 B	11	652	515.2 B	11
Heilongjiang	24	804.2 B	6	7	939.7 B	14
Henan	180	218.0 B	2	180	218.0 B	2
Hubei	50	588.4 B	11	50	465.5 B	8
Hunan	20	174.1 B	4	79	129.9 B	3
Jiangsu	50	1700 B	26	50	440 B	8
Jiangxi	2413	1600 B	16	482	1300 B	13
Jilin	no data	-	-	no data	-	-
Liaoning	4	1100 B	14	8	356.1 B	8
Qinghai	19	1000 B	15	18	928.6 B	15
Shaanxi	49	1100 B	15	47	748.6 B	15
Shandong	100	672.3 B	14	100	361.5 B	7
Shanxi	53	2100 B	21	73	1100 B	21
Sichuan	320	226.4 B	10	10	650.9 B	10
Yunnan	50	752.0 B	9	42	516.8 B	9
Zhejiang	1950	163.0 B	4	5580	217.0B	5
$\sum$	38065			37596		

Table 2: The "No. of blacklist records" and "No. of redlist records" indicate the number of credit records retrieved from each AD SCS platform for the most commonly implemented type of blacklist and redlist, respectively. Numbers show varying sample sizes due to several data collection obstacles (see Section 2.2). "Avg. size blacklist record" denotes the average byte size of a blacklist record for each sample. "No. of variables" indicates the number of informational variables on each credit record in the sample.

Ordinary People as Moral Heroes and Foes: Digital Role Model Narratives Propagate Social Norms in China's Social Credit System

Mo Chen
Technical University of Munich
Garching, Bavaria, Germany
mo.chen@tum.de

Severin Engelmann
Technical University of Munich
Garching, Bavaria, Germany
severin.engelmann@tum.de

Jens Grossklags
Technical University of Munich
Garching, Bavaria, Germany
jens.grossklags@in.tum.de

ABSTRACT

The Chinese Social Credit System (SCS) is a digital sociotechnical credit system that rewards and sanctions economic and social behaviors of individuals and companies. As a complex and transformative digital credit system, the SCS uses digital communication channels to inform the Chinese public about behaviors that lead to reward or sanction. Since 2017, the Chinese government has been publishing "blameworthy" and "praiseworthy" role model narratives of ordinary Chinese citizens on its central SCS information platform creditchina.gov.cn. Across many cultures, role model narratives are a known instrument to convey "appropriate" and "inappropriate" social norms. Using a directed content analysis methodology, we study the SCS-specific social norms embedded in 100 "praiseworthy" and 100 "blameworthy" role model narratives published on creditchina.gov.cn. "Blameworthy" role model narratives stress social norms associated with an "immoral" SCS identity label termed "Lao Lai" — a "moral foe" that fails to repay debt. SCS role model narratives familiarize Chinese society with SCS-specific measures such as digital surveillance, public shaming, and disproportionate punishment. Our study makes progress towards understanding how a state-run sociotechnical credit system combines digital tools with culturally familiar customs to propagate "blameworthy" and "praiseworthy" identities.

CCS CONCEPTS

• **Applied computing → Sociology**.

KEYWORDS

China, Social Credit System, moral education, narrative study

ACM Reference Format:
Mo Chen, Severin Engelmann, and Jens Grossklags. 2022. Ordinary People as Moral Heroes and Foes: Digital Role Model Narratives Propagate Social Norms in China's Social Credit System. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3514094.3534180

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534180

1 INTRODUCTION

In the past eight years, the Chinese government has made efforts to reshape its domestic power structure. The government removed the term limits for the Chinese presidency, created an anti-corruption ministry, and launched a "propaganda" app called "Xuexi Qiangguo" (学习强国, literally translated as "study and make the country strong").[1] Further, after four decades of rapid economic growth, *domestic* demand-driven models aim to consolidate economic sustainability [34, 60].

In 2014, the government published a *Planning Outline for the Construction of a Social Credit System (2014 - 2020)*; a high-level policy document that mandates a nationwide digital social credit system referred to as the Chinese Social Credit System (社会信用体系, SCS). The SCS's purpose is to evaluate, reward and punish the behavior of individuals, as well as commercial and societal organizations [57]. The outline describes two key SCS-specific regulatory measures: first, a digital "shaming"[2] and "praising" reputation system and a "joint punishment and reward mechanism" that distributes disproportionate "punishments" and "rewards", respectively [9, 10, 17, 19, 20, 39]. The Chinese SCS is a novel regulatory instrument enforcing reputational *and* material incentives and sanctions with the help of a large-scale digital infrastructure. The regulatory idea of the SCS rests on a broad conceptualization of "credit" that covers economic and social behaviors. SCS policy documents specify 14 different economic (e.g., production safety, finance, construction, e-commerce, etc.) and 10 different social sectors (e.g., health care, social security, and labor and employment) for credit application [57]. This "credit everywhere" directive subjects Chinese society to an all-encompassing concept of metrics with the aim to build a "socialist harmonious society" without "social contradictions" [57].

The establishment of a large-scale digital SCS to enforce social norms[3] corroborates the government's efforts to govern society through mechanisms that go beyond common legal and regulatory practices. In order for citizens to comply with SCS-specific social norms, the government must create awareness and understanding

[1] Civil servants, and employees of state-owned enterprises, particularly party members, are "encouraged" to use the app [56].
[2] Throughout this paper, the authors use quotation marks to communicate a neutral standpoint towards SCS-specific normative concepts (e.g., "praiseworthy", "blameworthy", "shaming", "praising").
[3] This paper uses the term *social norm* in a purely functionalist manner (see, e.g., [22]). A functionalist account defines social norms as deliberate measures by one party or group to establish social order over another. While other accounts of social norms study their natural emergence in individual or group interaction (see, e.g., [15]), a functionalist account puts emphasis on the exogenous dimensions of social norms attributable to the Chinese SCS.

of these norms. This research focuses on the central SCS platform "Credit China" (creditchina.gov.cn). Run by the National Center for Public Credit Information, the platform functions as the main SCS platform on all SCS-related developments. "Credit China" provides public access to official policy documents of the SCS, presents different types of reputational blacklists, and publishes SCS role model narratives and SCS news reports; as such, the platform also propagates SCS-specific social norms to the Chinese public.

While previous research on the SCS has largely focused on policy document analysis, here, we contribute to a more precise understanding of how the Chinese government makes SCS social norms intelligible to society at large. SCS policy documents describe vague instructions on SCS development, a common trait of policy documents issued by the central government [25]. Moreover, the broad public does not tend to engage with policy documents. Second, the SCS digitally publishes credit records on citizens, companies, and other organizations on so-called SCS blacklists (displaying "blameworthy" behavior) and redlists (displaying "praiseworthy" behavior). While blacklist and redlist records provide some information on why an entity was listed (i.e., punished or rewarded) [19], such justifications are written in legal and technical jargon. They do not offer causal or contextual clarifications for the sanctioned or rewarded behaviors [20]. We observe that, since 2017, the national SCS platform "Credit China" (creditchina.gov.cn) has been regularly publishing *SCS role model narratives* on "praiseworthy" and "blameworthy" behaviors. SCS role model narratives explicitly convey SCS-specific social norms to a broad audience. They vividly illustrate how ordinary Chinese citizens comply with or transgress SCS-specific norms and what consequences they experience. Narratives, stories, or folklore are as old as civilization. In the Chinese SCS, narratives on ordinary citizens are integrated into a digital infrastructure. They are published online and readers can share narratives to Chinese social media platforms amplifying the messages they seek to convey.

China has a long cultural tradition of propagating social norms through narratives, stories, and portraits of *model individuals* (e.g., [1]). First, Chinese ethical scholarship formulates principles through narratives, rather than through abstract principles. Second, besides a plethora of ancient moral narratives that still profoundly influence moral education in China today,[4] the Chinese government today uses narratives to showcase moral exemplars through reader-friendly stories and portraits (e.g., famous and popular narratives on moral heroes such as Huang Jiguang and Lei Feng). In the context of the SCS, we find that the government employs a similar strategy. Consequently, their analysis enables a more substantive understanding of the specific social norms the Chinese government wants the public to comply with and internalize with regard to SCS implementation.

We apply a directed content analysis methodology to systematically study the SCS-specific social norms embedded in 200 "blameworthy" and "praiseworthy" role model narratives on creditchina.gov.cn. Our study exemplifies how socio-cultural traditions influence and resurface in the implementation of a large digital sociotechnical system. Role model narratives on creditchina.gov.cn represent a prime example of how "…state actors appropriate technologies to support broader ideological shifts in their discourse" [36]. In addition, digital narratives present the biographical information and moral judgments of ordinary Chinese citizens that, as we show, can be distributed to large social media networks. SCS narratives demonstrate the problematic coupling of traditional values and socio-political policy plans by large digital infrastructures.

2 BACKGROUND

In Western media, the SCS has been linked to a national metric system assigning social credit scores to individuals (e.g., [6, 21]). While this perspective needs clarification, the SCS allows for more government supervision of individuals, companies, and institutions through digital information technologies. First, big information technology companies contribute to the construction of the SCS and distribute trustworthiness scores to individuals in promotion programs (e.g., Zhima Credit) [10, 32]. Second, *local* governments have tested different rating systems in "prototype cities" such as Rongcheng and Suzhou. Here, social credit ratings grant or deny citizens access to various public services and products [41]. Participation in these local "credit scoring experiments" is mandatory. However, these local "policy experiments" do not necessarily serve as a model for national policy implementation.

A number of SCS-specific measures operate at the national level. Early research accounts noted the existence of different types of SCS blacklists and redlists. With these lists, the SCS uses digital platforms to publicly "shame" or "praise" natural and legal persons for non-compliance or compliance, respectively, with a variety of legal and social norms (e.g., [7, 19, 20, 27, 35, 45]). Another national SCS-specific measure is the SCS joint punishment and reward mechanism. Thereby, "praiseworthy" or "blameworthy" behavior in one specific area leads to "reward" or "punishment" in different areas of life. To give just one example, blacklisted individuals have been barred from booking 26.8 million flights and nearly 6 million high-speed train trips since June 2019 (according to the National Development and Reform Commission).[5] Scholars note that public "shaming" and "praising" platforms as well as joint punishment and reward mechanisms differentiate the Chinese SCS from other social credit systems [17, 20].

2.1 SCS implementation as a digital transformation of culturally and politically familiar customs

Social science and legal scholarship has mainly focused on the privacy implications that result from the surveillance measures of the Chinese SCS (e.g., [28, 35]). A key observation is that the Chinese SCS is able to collect, process, and analyze personal data for a broad range of different purposes [12, 47]. As a "surveillance system", the SCS is a critical stepping stone for the government not only to monitor, but also to regulate and shape people's behaviors [39]. However, prior research seems to indicate that Chinese citizens do not primarily associate the SCS with the dangers of surveillance [32]. Compared to the astonishment and criticism from some Western media (e.g., [7, 14, 42]), Chinese citizens appear to perceive

[4] For example, the *Twenty-four Stories about Filial Piety* written by Guo in the Yuan Dynasty.

[5] Refer to http://www.sohu.com/a/327229387_120054409, last accessed on May 21, 2022.

the SCS favorably rather than critically [32]. The high approval levels can partially be explained by the effort of the government to base SCS mechanisms on culturally familiar customs and practices. For example, blacklists and redlists are common modes of shaming and praising schemes in Chinese society. In kindergarten, it is not uncommon for children to receive "praise" and "blame" via so-called "Honor Rolls" and "Critique Rolls", respectively. Beyond kindergarten, "praise" and "blame" mechanisms include public presentation of photos of individuals on banners at the entrance of buildings such as hospitals, schools, and companies. The distribution of reputational "reward" and "punishment" by institutions represents a culturally accepted regulatory instrument.

Second, according to survey research, Chinese citizens voice little doubt regarding the political legitimacy of the government to ensure social order through surveillance and monitoring systems [32]. Characteristics of what has been referred to as the "surveillance tradition" of the government date back to the "personal file system" *dang'an* [17, 35]—a national archive system that was set up in 1949 to systematically collect, record, and store information on citizens' and organizations' attitudes and behaviors [43]. Similar to the *dang'an*, SCS measures apply to individual citizens, companies, and social organizations. Given the longstanding surveillance practices represented by the *dang'an* system, Chinese society is unlikely to perceive the implementation of data-rich digital reputation lists by the government as an illegitimate political measure. This is not to say that Chinese citizens attach a low value to their privacy in principle. When it comes to using corporate digital services such as WeChat, for example, Chinese citizens do raise concerns about their privacy but are less likely to take corresponding privacy actions [11]—this "privacy paradox" is prevalent among users in Western societies, too [13].

2.2 Narratives as instruments for propagating ethical norms and political propaganda

Across cultures, stories, poems, and plays are an indispensable and prevalent source of ethical principles [4, 26, 51, 53]. Narratives naturally raise ethical questions and present possible model behaviors, good and bad. The narrative format is particularly suitable to illustrate complex ethical scenarios in a comprehensible manner. In William Shakespeare's *King Henry V* soldiers face the moral trade-off whether to fulfill the king's demands for war when they believe that the king's motivation for war is irrational and unjust. Or take Mark Twain's *The Adventures of Huckleberry Finn*. The story illustrates the moral tensions of Huckleberry Finn who decides to protect his escaped enslaved friend Jim rather than returning Miss Watson's "lost property". Narratives are powerful media for ethical deliberations, they place moral choices in specific, real-world contexts. The narrative format may not be suitable for generalizing abstract principles, but it vividly reveals the conditional trajectories that cause protagonists to face moral trade-offs or dilemmas [53].

Deontological and utilitarian ethics are typically concerned with the conceptual development of ethical principles. These ethical traditions justify a moral imperative conceptually and take them to be universally valid across contextual conditions. In contrast, Chinese ethics has a practical focus and demands practical solutions to specific ethical conflicts [62], and is *"skeptical that highly abstract theories will provide a response that is true to the complexities of that problem"* [61]. As such, Chinese moral philosophy takes a predominantly virtue ethics approach. Its emphasis lies on the development and presentation of a particular moral character in the face of a particular problem [61]. Here, the narrative format plays an indispensable role in conveying ethical deliberation and decision-making in Chinese ethics. Examples of Chinese role model narratives abound. The *Biographies of Exemplary Women*, compiled two millennia ago, is the earliest extant book of Confucian ethics solely devoted to the education of women. It includes 125 biographical accounts of exemplary women in ancient China. Well-known to the Chinese today is the famous *Twenty-four Stories about Filial Piety*. Written about 700 years ago, this collection of stories aims to educate the public on the virtue of Confucian filial piety. In Confucian ethics the virtue of filial piety represents a constitutive element of "communitarianism". Narrated scenarios illustrate virtuous acts that cover moral conflicts. For example, the passage *7A35* in the book *Mencius*, places the protagonist in the following situation: would one hand over one's own father to the state if he has committed a murder? Another "virtuous exemplar" of filial piety—perhaps better known to the Western world—is the young girl called Mulan. An entire collection of poems called the *Ballad of Mulan* documents her courage and sense of duty in China 1500 years ago.[6]

In the 20th century, the Chinese government has used role model narratives to underline "praiseworthy" moral dispositions. For instance, Huang Jiguang is highly decorated as a revolutionary martyr for "sacrificing" himself during the Korean War in the 1950s. Another example is the story of Lei Feng—a socialist hero during the 1960s and a famous hero in contemporary Chinese society [49]. He is glorified for his "unconditional loyalty" to the Chinese Communist Party (CCP). More recently, stories praising and blaming citizens regularly appear on Chinese television. In 2016, the state's television station China Central Television produced a special program called "Role Model/榜样". In each season, the program presents the "stories" of ten CCP members, praising their dedication and steadfastness in their faith as CCP members. The Chinese public is familiar with the use of narrative portraits of role models that propagate political and ideological ideals. Narratives published on creditchina.gov.cn follow this tradition and instill a representation of everyday moral life in citizens' minds [38]. This work presents evidence that the Chinese SCS uses narratives of ordinary Chinese citizens to familiarize society with digital surveillance practices and digital reputation listings to enforce SCS-specific norms.

3 DATA AND METHODS

3.1 Data

In September 2017, the national SCS platform creditchina.gov.cn started the regular publication of "blameworthy" role model narratives about "dishonest"/"untrustworthy" natural and legal persons. These "blameworthy" role model narratives can be accessed on the landing page of creditchina.gov.cn (titled "representative cases/典型案例)"[7]. In November 2017, the platform also started publishing

[6] For a comprehensive overview of narratives in Chinese ethics, see [61].

[7] This section only included "blameworthy" narratives when we crawled the data in August 2018. Now, this section includes both "blameworthy" and "praiseworthy" narratives.

Title: *Quzhou Court: For the first time, capturing a "Lao Lai" using the measure of "temporary control"!*

The charged person, [redacted] (last name), is a "contractor". In 2017, he hired three people ([redacted], [redacted] and another person) to work for a steel company in Fengnan District, Tangshan City. He did not pay the workers their salary which amounted to 14,300 RMB and was subsequently sued by the court. The court of Quzhou County ordered [redacted] to pay 14,300 RMB for labour remuneration to [redacted], [redacted], and others. After the verdict came into effect, [redacted] refused to fulfill his obligation, and the case entered the enforcement process.

The court of Quzhou County dealt this case as one involving people's livelihood and tried to educate and persuade [redacted] directly or through his family members. But [redacted] still refused to fulfil his obligation. He went out to work and played the game of "hide and seek" with court executives. [redacted] was then put on the "List of Dishonest Persons Subject to Enforcement" according to law, and became a "Lao Lai". Due to [redacted]'s long-term concealment and evasion of execution of assets, in July this year, the court of Quzhou County applied "temporary control" in accordance with the law with the help of the public security bureau in July. Only three days later, the Yonghongqiao Police Station, Lunan Branch of Tangshan Public Security Bureau came with the good news: [redacted] was successfully captured. The court of Quzhou County dispatched executives who drove more than 1,200 kilometres overnight to take [redacted] back. Frightened of the strong enforcement, [redacted] contacted his family on his way back to the court. Finally, his family then sent the money to the court.

Figure 1:
Translation of a "blameworthy" role model narrative from creditchina.gov.cn. This is an excerpt of the complete role model narrative. The narrative also provided the following information: publication date (July 30, 2018), original source of the role model narrative (Jiaotong Wang), and the category of the role model narrative (Representative Cases); as well as a sharing function with links to the platforms of Wechat, Weibo, Baidu Tieba, and Renren.

Title: [redacted] *Twenty years of upholding "honesty and trustworthiness" and giving back to the home village*

[redacted], born in November 1963, is a member of the Communist Party of China, secretary of the party branch of [redacted] Village, Shuitun Town, Yicheng District, Zhumadian City, Henan Province, general manager of [redacted] Human Resources Co., Ltd. and [redacted] Technology Co., Ltd. [redacted] has always adhered to the life tenet of "honesty and trustworthiness is gold and virtuous". He has set up his own "Poverty Alleviation Convoy" with the idea of "facilitating labour with passenger transport, promoting poverty alleviation with labour". For 20 years, he has behaved according to the virtues of honesty and trustworthiness, exempting transport fares for migrant workers from his home village for over 50 million yuan, sending more than 1.6 million people to the south for employment, and helping more than 3,000 families to get rid of poverty. He tried every means to persuade five companies to settle in [redacted] Village, fulfilling the dream of poor households seeking employment and poverty alleviation at the doorstep of his home. He is enthusiastic about public welfare and donated more than 3 million yuan to roads and bridges construction, education, earthquake relief, and supporting students in need. In recent years, [redacted] has been awarded more than 30 titles including "National Outstanding Migrant Workers", "Outstanding Migrant Workers from Henan Province" and Zhumadian City "May 1st Labour Medal".

Figure 2:
Translation of a "praiseworthy" role model narrative from creditchina.gov.cn. This is an excerpt of the complete role model narrative. The web-page also provided the following information: publication date (April 2, 2018), original source of the role model narrative (Credit China), and the category of the role model narrative (Trustworthy Figures); as well as a sharing function. It also featured an image of the protagonist and an audio recording of the narrative.

"praiseworthy" role mode narratives of "honest" and "trustworthy" individuals and representatives of companies. These "praiseworthy" role model narratives can be accessed on the sub-page "credit culture (诚信文化)" under the headline "integrity characters/stories (诚信人物／故事)". Both "praiseworthy" and "blameworthy" narratives are either created and published by creditchina.gov.cn itself or selected and taken from city, provincial, and other national government-associated news outlets.

We crawled and scraped publicly available "blameworthy" and "praiseworthy" narratives on creditchina.gov.cn. This resulted in a corpus of 798 "blameworthy" and 156 "praiseworthy" role model narratives. To generate comparable datasets, we used the random number method (e.g., [18]) to select 100 "praiseworthy" and 100 "blameworthy" role model narratives. We found that protagonists in all "praiseworthy" narratives were individuals and their full names were provided. In contrast, 11 out of 100 "blameworthy" narratives (11%) portrayed companies. Only in 2 "blameworthy" cases (2%), a full name of the protagonist was included, while in the remaining 98 cases the protagonist's name was partly anonymized (only the family name was provided). In the process of coding, we obscured the protagonist's name, living address and related companies' names to reduce the risk of re-identification. Translations of a "blameworthy" and a "praiseworthy" narrative can be found in Figures 1 and 2, respectively.

3.2 Research ethics

Our analysis is built on publicly available data from key sites of the Chinese SCS, which is posted with the intent of public scrutiny. The two main frameworks and tools used for the crawling and scraping process were ThoughtWorks Limited open source headless browser Selenium and Scrapinghub Limited open source framework called Scrapy. Our methodological approach conformed to the legal and ethical principles of web scraping [33]. Moreover, our research adheres to ethical guidelines on crawling publicly available SCS data raised in [19]. These include protecting the privacy of data subjects at all times and checking for robots.txt files before crawling.

3.3 Method

We applied a directed content analysis to map out social norms propagated through role model narratives published on creditchina.gov.cn. Directed content analysis draws on existing research when identifying appropriate codes for textual analysis (see, in particular, [29]). We developed four codes based on Tappan and Brown's work on the analysis of narratives about individuals that experience a moral conflict [58]. A first code termed "moral conflict" (Code 1) documented the moral conflict of an individual in a given role model narrative. Next, we developed codes that helped us explore the nature of the moral experience of the protagonist when confronted with the moral conflict. Tappan and Brown suggest that the moral experience of an individual in the context of moral conflict requires analysis of the *cognitive*, *affective*, and *conative* dimensions of the protagonist's experience [58]. These codes allowed us to pose the following questions: given the moral conflict, *what does the protagonist think?* (Code 2); *what does the protagonist feel?* (Code 3); and *what does the protagonist do?* (Code 4). Codes 2, 3, and 4 made the reflective, emotional, and behavioral dimensions of the moral experience intelligible.

We also wanted to understand whether the assignment of a single virtue or vice led to the attribution of other virtues or vices, respectively. We termed this code "virtue/vice cascade" (Code 5). First, being attributed multiple virtues for carrying out a specific virtuous act indicates a special importance of this virtue. Second, this code allowed us to define the broadness and specificity of the SCS conceptualization of its key virtues "honesty and trustworthiness" (as outlined in the official SCS documents, see [57]).

Furthermore, we took into account social norm messages that have proven to be effective in nudging individuals into a desired behavior [5, 16]. Two types of social norm messages are typically distinguished: *injunctive* and *descriptive* social norm messages. Injunctive norms refer to behavior other individuals approve of (e.g., 80% of individuals think activity x is morally good), while descriptive norms directly refer to the desirable behavior of others (e.g., 80% of individuals engage in desirable activity x) [16, 22, 50]. To avoid redundancy in our analysis (see Code 1 "moral conflict" and Code 4 "the protagonist's actions"), we only used injunctive norms for our analysis (Code 6).

Finally, we applied a code to understand how the author of a role model narrative interpreted the overall moral identity of the protagonist. In role model narratives, authors construct moral identities [3, 37]. A particular interpretation of the individuals' moral experiences (see Codes 2, 3, 4) by the authors signals the virtues and vices a model citizen, company, or organization is supposed to conform to. As is common in Chinese ethics, virtues and vices tend to be connected to a particular identity ("the moral exemplar"). In order to capture such a moral identity in the role model narratives, we created a code termed "identity labeling" (Code 7). Our final coding scheme included three categories with seven codes in total (for the coding schemes for "praiseworthy" and "blameworthy" narratives, respectively) (see, e.g., Table 1).

4 RESULTS

Text lengths and SCS keywords: The average length of "praiseworthy" narratives was 1,423.27 Chinese characters, more than two times longer than that of "blameworthy" narratives (544.77 Chinese characters). "Praiseworthy" but not "blameworthy" narratives featured either a real photo of the protagonist (46 narratives) or an audio recording of the narrative (50 narratives).

A word frequency analysis revealed the terms "honest/诚实", "trustworthiness/守信" and "honest and trustworthy/诚信" were mentioned altogether 348 times in "praiseworthy" narratives. In "blameworthy" narratives, the contrary concept "untrustworthy/失信" was mentioned only 145 times. However, we found that the term "Lao Lai/老赖" appeared 198 times across "blameworthy" narratives and at least once in every "blameworthy" narrative in our sample. "Lao Lai" refers to individuals or companies that do not repay debt and is commonly known as a substitute of "dishonest person subject to enforcement (失信被执行人)".

Finally, we wanted to understand the occurrence of different SCS-specific and non-specific sanction and detection measures in "blameworthy" role model narratives (see Figure 3). 36 "blameworthy" narratives included the term "blacklist". "Public shaming" was

Table 1: Coding scheme for "blameworthy" role model narratives.

Categories	Codes	Examples
Narrative context	(1) Decision scenario	Owing debts of 30 million RMB
	(2) The protagonist's thoughts	"It is only 2000 RMB. I do not have to repay."
	(3) The protagonist's feelings	"I feel deeply regretful".
	(4) The protagonist's actions	Refusing to repay debt with various excuses.
Virtue/Vice	(5) Vice cascade	He fails to repay debt, ..., he lied.
Social norm expression	(6) Injunctive norm	"Neighbors will not come into contact with the Lao Lai."
Identity	(7) Identity labeling	"Lao Lai (老赖)"
		Owing debts of 30 million RMB... still lives a luxury life.

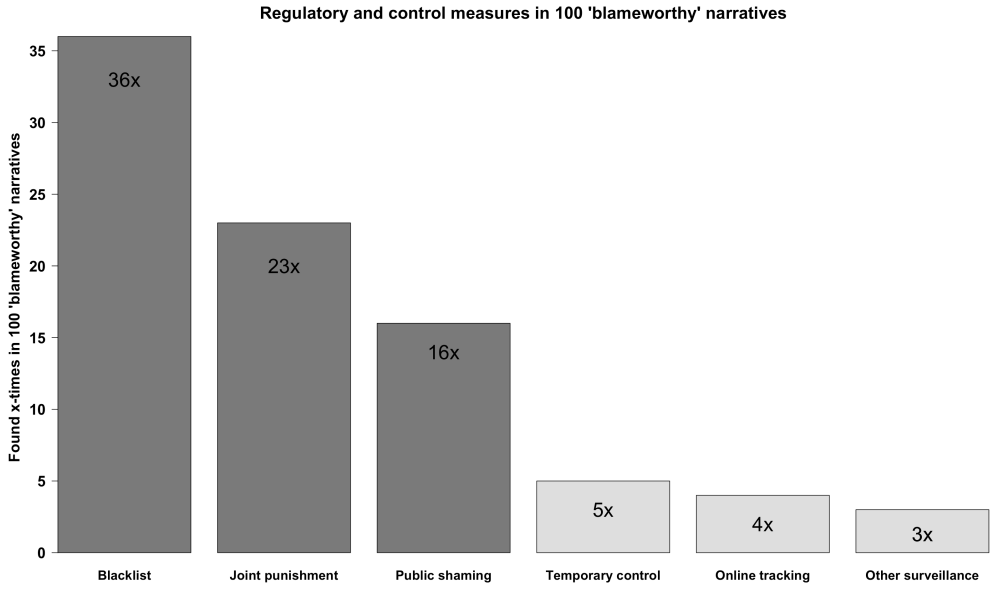


Figure 3: Number of different regulatory and control mechanisms in "blameworthy" narratives. Dark gray: SCS-specific mechanisms. Light gray: three other types of regulatory and control mechanisms including online tracking (e.g., social media tracking).

explicitly mentioned in 16 of the "blameworthy" narratives. Here, the protagonist's personal information (e.g., passport photo) was posted either online (e.g., social media) or offline at bus stops in the protagonist's living area. 23 "blameworthy" narratives used the term "joint punishment". In these narratives, the protagonist failed to repay debt and was subsequently banned from taking high-speed trains, boarding flights, participating in village elections, departing from and entering China, applying for loans from the bank, gaining job promotions as a public servant, and/or indulging in luxury consumption. In five narratives, the "joint punishment" mechanism sanctioned the protagonist's family members. For example, the protagonist's child could not go to a private school (with high tuition fees) due to the father's transgressions (a measure that is also formulated in the relevant SCS policy document).

Other narratives described how the government was capable of effectively capturing "Lao Lai". "Temporary control" (临控)" is an online or offline surveillance measure operated by the public security organs to monitor an individual's activities. Online accounts and information taken from social media were collected to track the protagonist in four narratives. In three narratives, other surveillance strategies were applied such as video surveillance. "Blameworthy" narratives also highlighted data sharing practices between public security services, hotel registries, and train ticket booking sites for surveillance purposes.

Biographical information of protagonists: Protagonists in "praiseworthy" narratives were individuals. 11 "blameworthy" narratives portrayed companies; eight described a legal representative of the company.

In our sample, 99 "praiseworthy" narratives communicated the gender of the protagonist (75 males, 24 females), 73 "praiseworthy" narratives indicated the age. For "blameworthy" narratives, 49% of the sample indicated the gender of the protagonist (39 males, 5 females). The protagonist's living location was given in 94 "blameworthy" narratives.

4.1 Qualitative content analysis

4.1.1 The narrative's storyline. "Praiseworthy" narratives covered a variety of different moral conflicts. These dealt with ostensibly incommensurable trade-offs between protagonists' interests and the interests of the collective (see Figure 4). Protagonists were confronted with a moral conflict that tempted them to further their own self-interests at the expense of civic honesty. Protagonists in the "praiseworthy" narratives always chose to be honest towards other members of society. In "praiseworthy" narratives, we identified 141 decisions in total (narratives could include multiple conflicts). 31 of these decisions were about paying back debt or salary. The protagonist typically repaid his or her debt faithfully; often despite modest financial possibilities. 29 scenarios showed protagonists helping others financially or non-financially. In another 19 narratives, businessmen guaranteed product quality at the cost of their own economic interest. Other scenarios included taking care of both family and non-family members in various contexts (15), returning lost property of others under various circumstances (13), giving back to one's home village financially and non-financially (12), and working diligently for the public good (11).

All "blameworthy" narratives portrayed an individual who deliberately failed to fulfill a financial obligation, i.e., a repayment of debt—ranging from 300 USD to about 16 million USD. A typical "blameworthy" narrative explained how a Chinese court used various surveillance technologies to identify and sanction "Lao Lai". Across the "blameworthy" narratives, the list of sanctions included exclusion from high-speed trains and any form of political participation, public shaming, detention, and imprisonment.

4.1.2 The protagonists' moral experiences. What the protagonist thinks (cognitive): 95 "praiseworthy" narratives described the cognitive experience of the protagonist when facing the moral conflict (see Table 2). Protagonists either reflected on the importance of being trustworthy in the role they had in society (e.g., as a citizen, lawyer, or doctor) or on the general well-being of others (e.g., "the owner of the lost wallet must be worried").

In contrast, only 27 "blameworthy" narratives described the protagonist's thinking. "Blameworthy" narratives showcased the protagonist's *misrepresentation* of the moral scenario. For example, a "Lao Lai" falsely believed that he was not responsible for the debt and therefore not obligated to repay. In another narrative, a "Lao Lai" with debt falsely thought that the court could not take effective measures against him because of his low economic status. After being threatened with detention he paid back the debt. In another example, an individual owed a relatively small amount of money to another citizen (2000 RMB, around 300 USD) and thought the court would not enforce any sanctions, which turned out to be false.

What the protagonist feels (affective): 63 "praiseworthy" narratives described the emotional state of the protagonist. The most common emotive attitude displayed by protagonists was a "rewarding sense of responsibility" and "satisfaction" as a result of being "honest" toward other citizens.

38 "blameworthy" narratives described how protagonists felt about their behavior. "Lao Lai" either felt "apologetic" or "regretful" for their actions or feared the consequences of being punished: for example, being detained by the police or being publicly shamed on blacklists. The emotions of "Lao Lai" were described only after their misbehavior had been revealed.

How the protagonist acts (conative): In all "praiseworthy" narratives, individuals acted according to what they believed was expected of them by society: A "good" citizen returns the lost property of another citizen, a "good" doctor treats everybody regardless of their financial background, and a "good" entrepreneur pays employees on time.

In "blameworthy" narratives, protagonists escaped debt obligations by moving to another province, hiding in another family's home, or secretly transferring assets to another person. After the court had taken a certain enforcement action, "Lao Lai" fulfilled the debt obligation. For example, one protagonist lived a luxury life based on debt and frequently showed his wealth on social media. When the individual was identified and punished by public shaming he was reported to have paid back the debt immediately.

4.1.3 Virtue & vice cascade. In our sample, 88 "praiseworthy" narratives featured a "virtue cascade": when protagonists were reported to be "honest" or "trustworthy", protagonists were attributed multiple other virtues. These included diligence, kindheartedness or benevolence, filial piety, and a sense of responsibility to the society.

In contrast, only 16 "blameworthy" narratives featured a corresponding "vice cascade". 11 of them highlighted that a "Lao Lai" was also a "liar". Two "blameworthy" narratives told the story of a "Lao Lai" that was "dishonest" to his friends that had previously helped him.

4.1.4 Injunctive norm expression. 79 "praiseworthy" narratives incorporated multiple different injunctive norms such as positive comments from co-workers and villagers, friendly nicknames given by members of the social circle (e.g., "the secretary for children"), and official honorary awards (e.g., "Good People in Anhui Province").

Only 9 "blameworthy" narratives used an injunctive norm. In one "blameworthy" role model narrative, the injunctive norm was expressed by the protagonist: "My neighbours would not come into contact with me once they knew that I am a Lao Lai". In five "blameworthy" narratives, injunctive norms were propagated through the activities and words of relatives who fulfilled debt obligations for the "Lao Lai".

4.1.5 Identity. "Praiseworthy" role model narratives did not include a specific label that served to emphasize a morally ideal identity. In contrast, "blameworthy" narratives fostered a strong link between a specific "immoral" behavior (i.e., deliberately avoiding to repay debt) and a specific "blameworthy" identity, the "Lao Lai". In only one narrative, the individual himself expressed explicitly that

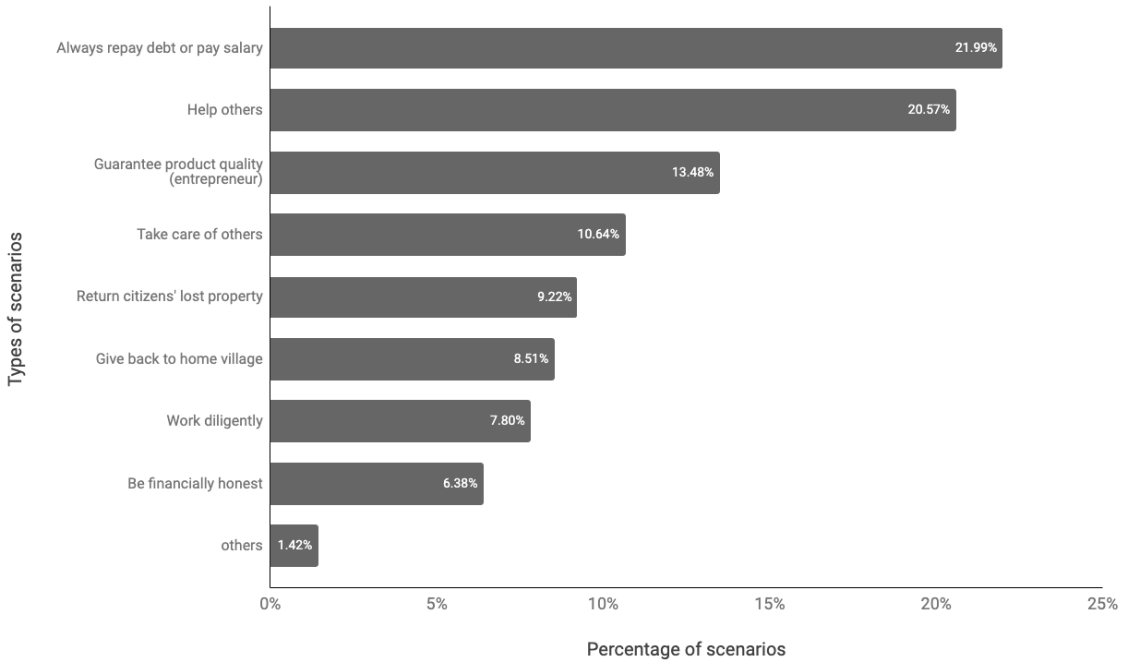


Figure 4:
Scenario analysis for "praiseworthy" narratives. "Other" mostly referred to various economic virtues: pay employees on time, take care of consumers' rights, and obey the CCP under any circumstances. Numeric values represent the percentages of texts that feature a given scenario.

Table 2: Coding results.

	"Praiseworthy" role model narratives, frequencies (%)	**"Blameworthy" role model narratives, frequencies (%)**
Narrative context of the moral story		
Moral conflict	100% about voluntary sacrifice for public good	100% about debt obligation and the court's action
The protagonist's thoughts	95%	27%
The protagonist's feelings	63%	38%
The protagonist's actions	100% about sacrifice of self-interest	100% about the escape from debt obligation
Virtue & vice cascade		
Virtue cascade	88%	/
Vice cascade	/	16% about vice cascade
Social norm		
Injunctive norm	79%	9%
Identity	100% about honest and trustworthy; 100% justified	100% about "Lao Lai"; 41% justified

he was a "Lao Lai". In all other "blameworthy" narratives (99), the identity "Lao Lai" was attributed to the protagonist by the authors of the role model narratives. 41 narratives provided a justification for assigning the identity label "Lao Lai" to the protagonist. For example, a "Lao Lai" went on luxurious trips and lived in a high-end hotel while refusing to pay back debt. In the remaining 59 "blameworthy" narratives, however, the authors of the narratives did not justify the attribution of the "Lao Lai" label.

5 ANALYSIS

5.1 Role model narratives underline the SCS's priority for "sincerity" in economic activities

The SCS national platform propagates social norms through narratives focusing on transgressions in the context of economic activities. Across the narratives, businessmen and businesswomen were the most represented profession. Business activities ranged from selling breakfast on the street to producing an annual output worth over 100 million RMB (15 million US dollars). As such, different from traditional Chinese ethical narratives that cover a wide range of virtues, the SCS narratives have a specific focus—moral behaviors in an economic context. In addition, all "blameworthy" narratives reported on an individual or a company that failed to repay debt. This indicates the importance of economic development as a goal of the SCS: China's corporate defaults hit a record high of 62.59 billion RMB (9.67 billion USD) in the first half of 2021.[8] The ratio of household debt to GDP hit an all-time high of 62.4% in September 2021.[9] Investigating individual households, one can observe that the thriftiness culture and the tradition of savings are fading in China [48, 55]. Preventing debt defaults is a pressing economic issue in China and the SCS purports to be part of its solution. The strong focus on the detection and subsequent punishment of "Lao Lai" provides evidence that the SCS makes financial dishonesty very costly. In addition, the SCS represents a new measure to evaluate the creditworthiness of individuals and companies. The broad conceptualization of "credit" enables evaluation of businesses based on *trustworthiness* rather than on *financial* creditworthiness. Here, SCS redlists and blacklists further aim to decrease informational asymmetry between cooperating entities [23, 30].

5.2 SCS role model narratives use ordinary people as moral heroes and familiarize the public with SCS-specific surveillance

A result of reading "blameworthy" narratives is that the readership inevitably becomes familiar with the different forms of technological and administrative surveillance measures. Here, the narrative format allows authors to introduce the state's range of surveillance tools: online tracking, digital blacklisting, temporary control. Narratives clarify the purpose for which they can be used and showcase the near unconditional success of surveillance technologies in finding those that have not complied with laws. Narratives on creditchina.gov.cn are able to accomplish what neither the SCS policy documents nor the SCS blacklists or redlists achieve: they combine empirical with fictional elements to portray the power of the state's surveillance apparatus in sanctioning defectors and transgressors. They can be swiftly accessed on the platforms and are easy to read.

Role model narratives use *ordinary* people rather than heroes as moral exemplars. The one-sided emphasis on ordinary people echoes what Turner has referred to as "demotic turn" [59]. It denotes an increasing visibility of ordinary people in mass media. The media not only celebrates ordinary people through reality TV, journalism, radio, and user-generated content but actively creates culturally intelligible identities around them. Scholars of narratives have argued that life stories of ordinary citizens are a *"marker for a society that is losing faith in the more established sacred narratives of religion, preferring more prosaic accounts for advice and guidance"* [44]. In China, there has been an increasing use of ordinary public idols such as socialist heroes and other non-elite figures since the 1950s [31]. Popular Chinese television programs such as *Touching China* (感动中国) and *Civilian Heroes* (平民英雄) illustrate this transformation.[10] However, *currently*, we cannot find a TV program focusing on the SCS specifically. SCS narratives are potentially powerful instruments for propagating SCS-specific social norms to a broad audience. Their sharing to all relevant Chinese social media platforms effectively increases their visibility.

5.3 The emergence of the "Lao Lai" as an "immoral" SCS identity

The strict categorization into "praiseworthy" and "blameworthy" role models corresponds to the two ideal moral role models in Confucianism, one of the most prominent traditions of Chinese ethics. In Confucianism, the *Junzi* represents the gentleman (literal translation), while the *xiaoren* literally refers to a "small man" [8]. In the *Analects, Book 4.16*, for instance, Confucius stated that *"The gentleman comprehends righteousness; the small man comprehends profit"*. In traditional Chinese narratives, a particular virtue is exemplified across different social scenarios by the *junzi*, or in contrast, by the *xiaoren*. Such an exemplary person displays virtuous or immoral acts for the public to imitate or to refrain from, respectively. It is for this reason that Chinese ethics is often referred to as "exemplarism" [46], whereby ethical judgment is fundamentally based on "analogical reasoning" [54, 62]. The communication of such "exemplarism" unfolds best in the narrative format: stories inspire an audience to strive for the moral character of the *junzi* or to refrain from being labeled as the *xiaoren*.

Authors of role-model narratives deliberately use stylistic features to strengthen the distinction between "praiseworthy" and "blameworthy" moral characters. "Praiseworthy" narratives attempt to create sympathy and empathy with protagonists when they illustrate the reflective and emotional dimensions of virtuous intentions and convictions. The presentation of a photograph and the detail of biographical information further emphasize that protagonists are worthy of moral emulation in "praiseworthy" narratives. In contrast, the lack of a visual depiction and the informational reduction to a stereotypical label "Lao Lai" of protagonists in "blameworthy" narratives aim to produce a dissuasive effect. The attribution of the label "Lao Lai" lacks justification. In "blameworthy" narratives, protagonists' intentions and beliefs are revealed retrospectively, concealing the reasons that led to the borrowing of money and the subsequent failure to repay. Generally, "blameworthy" narratives do not specify why the protagonist is in a debt situation in the first place. While there are many—perfectly justifiable—reasons why a person can end up in a debt situation (e.g., sickness, loss of

[8] Data source: *Reuters* at https://www.reuters.com/world/china/chinas-corporate-bond-defaults-touch-record-high-2021-07-09/, accessed on May 26, 2022.
[9] Data source: *CEIC* at https://www.ceicdata.com/en/indicator/china/household-debt--of-nominal-gdp, accessed on May 26, 2022.
[10] Both TV programs focus on the moral lives of ordinary Chinese citizens.

employment), authors of "blameworthy" narratives only attended to the reflective and emotional experience of protagonists after they have been captured and sanctioned. An insufficiently justified identity label likely creates stereotypying and possibly discrimination against members of this group [2, 52]. Labels function as external identity markers, constituting an influence on an individual's identity beyond the individual's control [24]. Being assigned such a label may carry a number of negative connotations, treating an individual as if they were generally rather than specifically in the wrong. Subsequently, such individuals could be gradually cut off from participation in more conventional (group) activities, denied ordinary means of carrying out the routines of everyday life, and may eventually find themselves in social isolation. As is illustrated by the "blameworthy" narratives, reports on "Lao Lai" regularly appear on TV news programs, in newspapers, on websites, on social media, or in public areas such as train stations and bus stops.

In a recent study on the relationship between folklore and economic prosperity in 958 societies, Michalopoulos & Xue find that the depiction of "tricksters" or "cheaters" is among the most common archetypes in narrative traditions around the world [40]. Importantly, cultures with more narratives on tricksters that are unsuccessful and that get punished for their antisocial behavior are more trusting and prosperous today than cultures with narratives in which tricksters often get away. The authors argue that such *"folklore-based measures of historical attitudes are robust predictors of contemporary values and economic choices"* [40]. Observing that "Lao Lai" are always identified, captured, and sanctioned in the role model narratives we studied, leads us to believe that SCS narratives could work as powerful portraits of antisocial behavior in Chinese society nowadays.

6 CONCLUDING REMARKS

We analyze 100 "blameworthy" and 100 "praiseworthy" role model narratives on creditchina.gov.cn. We find that these narratives help to instill a sense of "folk morality", showcasing, partly empirically and partly fictionally, how individuals comply with social norms, how they transgress them, and what consequences they experience. By authorial choice, narratives are rich in biographical detail, which helps readers believe in their presented realities. They are short stories and, as such, everything they contain is there for a reason. Indeed, SCS role model narratives are not "just-so stories" that are first and foremost entertaining in nature. They effectively model "blameworthy" and "praiseworthy" social norms in an epistemically viable manner: they explain a particular causal trajectory in the past, reconstructing specific episodes of moral decision-making coherently and vividly. They reflect the author's perceptions on the moral ills of social life in China.

Over time, social norms change, in particular, when societies face enormous challenges. We found that, in May 2020, creditchina.gov.cn started publishing narratives on "praiseworthy" and "blameworthy" social norms "necessitated" by the emergence of the coronavirus pandemic.[11] The SCS's *Planning Outline* [57] specifically mandates the application of the concept of "credit" to health care, health services, and public health. When we revisited the platform, we found that it displayed three types of narratives that can be translated into "positive role models/正面典型", "exposure of dishonest conducts/失信曝光", and "how wonderful you are/你有多美". Narratives on "positive role models" appeared to portray companies that have produced and distributed epidemic prevention materials to help fight the crisis. In contrast, narratives on the "exposure of dishonest conducts" focused on companies that—in response to the coronavirus—have jacked up their prices, produced and sold poor-quality or counterfeit epidemic prevention products, posted deceptive advertisements, or committed coronavirus-related tax fraud. These coronavirus-related "blameworthy" narratives also showcased protagonists who have sold wild animals illegally, spread rumors related to the pandemic, and hid or lied about their travel histories to avoid quarantine. The third type of coronavirus narrative "how wonderful you are" portrayed protagonists that have responded to the crisis particularly well as professionals (e.g., doctors, nurses, businessmen, etc.) and non-professionals (various types of volunteers). This shows that SCS narratives on creditchina.gov.cn can be swiftly adapted to address novel demands for moral "praise" and "blame".

SCS narratives fall back on traditional Chinese narratives that convey ethical values and norms. This can be interpreted as an attempt to disguise novel measures of social control as "old wine in new bottles". To say it in Chinese: 新瓶装旧酒 (roughly translated "using a successful strategy that echoes the past"). At least since the 1950s, however, moral education has never only been about cultivating people's morality in China, but has always been closely intertwined with the political agenda of the CCP [49].

Digital role model narratives keep up with the trend of applying digital technologies as tools of social control; they serve as a political instrument promoting policies, spreading ideology, and shaping public discussion. The familiar format of the narrative contributes to the government's efforts to legitimize a new form of social control through a variety of SCS-specific mechanisms such as blacklisting, public shaming, joint enforcement as well as other means of mass surveillance. Narratives on creditchina.gov.cn may seem innocuous to some readers. At the same time, they work as a further building block for the state's increasing surveillance and control over Chinese society.

[11] See https://www.creditchina.gov.cn/xinxingfeiyanyiqing/, accessed on May 26, 2022.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers as well as Marianne von Blomberg for their constructive feedback. We further are grateful for funding support from the Bavarian Research Institute for Digital Transformation (bidt). Mo Chen also received funding from the Fritz Thyssen Foundation for this research. Responsibility for the contents of this publication rests with the authors.

REFERENCES

[1] Børge Bakken. 2000. *The exemplary society: Human improvement, social control, and the dangers of modernity in China.* Oxford University Press.
[2] Daniel Bar-Tal. 2012. *Group beliefs: A conception for analyzing group structure, processes, and behavior.* Springer Science & Business Media.
[3] Roy F. Baumeister. 1998. The self. In *Handbook of Social Psychology* (4th ed.), Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (Eds.). Vol. 1. McGraw Hill, 680–740.
[4] Marvin W. Berkowitz and Fritz Oser. 1985. *Moral education: Theory and application.* Lawrence Erlbaum Associates.
[5] Douglas B. Bernheim. 1994. A theory of conformity. *Journal of Political Economy* 102, 5 (1994), 841–877.

[6] Rachel Botsman. 2017. Big data meets Big Brother as China moves to rate its citizens. *Wired UK* (Oct. 2017).

[7] Charlie Campbell. 2019. How China is using "Social Credit Scores" to reward and punish its citizens. *Time* (Jan. 2019).

[8] Wing-tsit Chan. 1988. Exploring the Confucian tradition. *Philosophy East and West* 38, 3 (1988), 234–250.

[9] Mo Chen, Kristina Bogner, Joana Becheva, and Jens Grossklags. 2021. The transparency of the Chinese Social Credit System from the perspective of German organizations. In *Proceedings of the 29th European Conference on Information Systems (ECIS)*.

[10] Mo Chen and Jens Grossklags. 2020. An analysis of the current state of the Consumer Credit Reporting System in China. *Proceedings on Privacy Enhancing Technologies* 2020, 4 (2020), 89–110.

[11] Mo Chen and Jens Grossklags. 2022. Social control in the digital transformation of society: A case study of the Chinese Social Credit System. *Social Sciences* 11, 6 (2022), Article No. 229.

[12] Yongxi Chen and Anne Cheung. 2017. The transparent self under big data profiling: Privacy and Chinese legislation on the Social Credit System. *The Journal of Comparative Law* 12, 2 (2017), 356–378.

[13] Zhen Troy Chen and Ming Cheung. 2018. Privacy perception and protection on Chinese social media: A case study of WeChat. *Ethics and Information Technology* 20, 4 (2018), 279–289.

[14] Martin Chorzempa, Paul Triolo, Samm Sacks, et al. 2018. *China's Social Credit System: A mark of progress or a threat to privacy?* Peterson Institute for International Economics, Policy Brief 18-14.

[15] Maciej Chudek and Joseph Henrich. 2011. Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences* 15, 5 (2011), 218–226.

[16] Robert B. Cialdini and Raymond R. Reno. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58, 6 (1990), 1015–1026.

[17] Rogier Creemers. 2016. Is big data increasing Beijing's capacity for control? http://www.chinafile.com/conversation/Is-Big-Data-Increasing-Beijing-Capacity-Control/ Last accessed on May 26, 2022.

[18] Johnnie Daniel. 2012. Choosing the type of probability sampling. In *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Sage Publications, Inc, Chapter 5, 125–175.

[19] Severin Engelmann, Mo Chen, Lorenz Dang, and Jens Grossklags. 2021. Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 78–88.

[20] Severin Engelmann, Mo Chen, Felix Fischer, Ching-Yu Kao, and Jens Grossklags. 2019. Clear sanctions, vague rewards: How China's Social Credit System currently defines "good" and "bad" behavior. In *Proceedings of the Second ACM Conference on Fairness, Accountability, and Transparency*. 69–78.

[21] Patrick Farrell and Patrick Tyrrell. 2018. China's 'Social Credit' monitoring: Big Brother's frightening new tool for repression. https://www.dailysignal.com/2018/11/02/chinas-social-credit-monitoring-big-brothers-frightening-new-tool-for-repression/ last accessed on May 26, 2022.

[22] Gerlinde Fellner, Rupert Sausgruber, and Christian Traxler. 2013. Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association* 11, 3 (2013), 634–660.

[23] Walter Garcia-Fontes. 2005. *Small and medium enterprises financing in China*. Central Bank of Malaysia Working Paper.

[24] Erving Goffman. 1959. *The presentation of self in everyday life*. Doubleday.

[25] Sebastian Heilmann. 2008. From local experiments to national policy: The origins of China's distinctive policy process. *The China Journal* 59 (2008), 1–30.

[26] Richard H. Hersh, John P. Miller, and Glen D. Fielding. 1980. *Models of Moral Education: An Appraisal*. Longman.

[27] Samantha Hoffman. 2017. Programming China. *Merics China Monitor* 44 (2017), 1–12.

[28] Samantha Hoffman. 2018. Managing the state: Social credit, surveillance, and the Chinese Communist Party's plan for China. In *Artificial Intelligence, China, Russia, and the Global Order*, Nicholas D. Wright (Ed.). Air University Press, 48–54.

[29] Hsiu-Fang Hsieh and Sarah Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15, 9 (2005), 1277–1288.

[30] Cheng Huang and Zhifei Liu. 2014. Analysis on financing difficulties for SMEs due to asymmetric information. *Global Disclosure of Economics and Business* 3, 1 (2014), 77–80.

[31] Elaine Jeffreys. 2012. Modern China's idols: Heroes, role models, stars and celebrities. *Portal: Journal of Multidisciplinary International Studies* 9, 1 (2012), 1–32.

[32] Genia Kostka. 2019. China's social credit systems and public opinion: Explaining high levels of approval. *New Media & Society* 21, 7 (2019), 1565–1593.

[33] Vlad Krotov and Leiser Silva. 2018. Legality and ethics of web scraping. In *Proceedings of the Twenty-fourth Americas Conference on Information Systems*.

[34] Nicholas R. Lardy. 2016. China: Toward a consumption-driven growth path. In *Seeking Changes: The Economic Development in Contemporary China*, Yanhui Zhou (Ed.). World Scientific, 85–111.

[35] Fan Liang, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain. 2018. Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure. *Policy & Internet* 10, 4 (2018), 415–453.

[36] Silvia Lindtner, Ken Anderson, and Paul Dourish. 2012. Cultural appropriation: Information technologies as sites of transnational imagination. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*. 77–86.

[37] Nina Mazar and Dan Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* 45, 6 (2008), 633–644.

[38] Bryan McLaughlin and John A. Velez. 2019. Imagined politics: How different media platforms transport citizens into political narratives. *Social Science Computer Review* 37, 1 (2019), 22–37.

[39] Mirjam Meissner and Jost Wübbeke. 2016. IT-backed authoritarianism: Information technology enhances central authority and control capacity under Xi Jinping. *China's Core Executive: Leadership Styles, Structures and Processes under Xi Jinping* (2016), 52–57.

[40] Stelios Michalopoulos and Melanie Meng Xue. 2021. Folklore. *The Quarterly Journal of Economics* 136, 4 (2021), 1993–2046.

[41] Simina Mistreanu. 2018. Life inside China's Social Credit laboratory: The party's massive experiment in ranking and monitoring Chinese citizens has already started. https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/ last accessed on May 26, 2022.

[42] Steven Mosher. 2019. China's new 'Social Credit System' is a dystopian nightmare. https://nypost.com/2019/05/18/chinas-new-social-credit-system-turns-orwells-1984-into-reality/ last accessed on May 26, 2022.

[43] William W. Moss. 1996. Dang'an: Contemporary Chinese archives. *The China Quarterly* 145 (1996), 112–129.

[44] Michael Murray. 2003. Narrative psychology and narrative analysis. In *Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design*, Paul M. Camic, Jean E. Rhodes, and Lucy Yardley (Eds.). American Psychological Association, 95–112.

[45] Mareike Ohlberg, Shazeda Ahmed, and Bertram Lang. 2018. Central planning, local experiments: The complex implementation of China's Social Credit System. *Mercator Inst. China Studies* (April 2018).

[46] Amy Olberding. 2008. Dreaming of the Duke of Zhou: Exemplarism and the Analects. *Journal of Chinese Philosophy* 35, 4 (2008), 625–639.

[47] Michael Persson, Marije Vlaskamp, and Fokke Obbema. 2015. *China rates its own citizens – Including online behavior.* https://www.volkskrant.nl/buitenland/china-rates-its-own-citizens-including-online-behaviour~a3979668/. last accessed on May 26, 2022.

[48] Andrew Polk. 2018. Chinese need to learn to save again. *Bloomberg.com* (February 2018).

[49] Gay Garland Reed. 1995. Moral/political education in the People's Republic of China: Learning through role models. *Journal of Moral Education* 24, 2 (1995), 99–111.

[50] Raymond R. Reno, Robert B. Cialdini, and Carl A. Kallgren. 1993. The transsituational influence of social norms. *Journal of Personality and Social Psychology* 64, 1 (1993), 104–112.

[51] David Resnick. 2002. The role of heroes in Jewish education. *Religious Education* 97, 2 (2002), 108–123.

[52] Milton Rokeach. 1960. *The Open and Closed Mind: Investigations into the Nature of Belief Systems and Personality Systems*. Basic Books.

[53] Peter Singer and Renata Singer. 2005. *The moral of the story: An anthology of ethics through literature*. Wiley.

[54] Edward Slingerland. 2011. The situationist critique and early Confucian virtue ethics. *Ethics* 121, 2 (2011), 390–419.

[55] Sohu Business. 2018. The Young Chinese with Zero Saving and Unaffordable Credit Cards Alert China's Economy. http://www.sohu.com/a/249634274_100063243, last accessed on May 26, 2022 (in Chinese).

[56] Philip Spence. 2019. How to cheat at Xi Jinping Thought. *Foreign Policy* (2019). https://foreignpolicy.com/2019/03/06/how-to-cheat-at-xi-jinping-thought/ Last accessed on February 4, 2022.

[57] State Council. 2014. Notice of the State Council on issuing the outline of the plan for building a Social Credit System (2014-2020); (in Chinese).

[58] Mark B. Tappan and Lyn Mikel Brown. 1989. Stories told and lessons learned: Toward a narrative approach to moral development and moral education. *Harvard Educational Review* 59, 2 (May 1989), 182–205.

[59] Graeme Turner. 2010. *Ordinary people and the media: The demotic turn*. Sage Publications.

[60] Rod Tyers. 2016. China and global macroeconomic interdependence. *The World Economy* 39, 11 (2016), 1674–1702.

[61] David Wong. 1995. Chinese ethics. In *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-chinese/ last accessed on May 26, 2022.

[62] David Wong. 2002. Reasons and analogical reasoning in Mengzi. In *Essays on the Moral Philosophy of Mengzi*, Xiusheng Liu and Philip J. Ivanhoe (Eds.). Hackett Publishing Indianapolis, 187–220.

Social Media Profiling Continues to Partake in the Development of Formalistic Self-Concepts. Social Media Users Think So, Too.

Severin Engelmann
Technical University of Munich
Garching, Germany
severin.engelmann@tum.de

Valentin Scheibe
Ludwig Maximilian University of Munich
Munich, Germany
v.scheibe@campus.lmu.de

Fiorella Battaglia
Università del Salento
Lecce, Italy
fiorella.battaglia@unisalento.it

Jens Grossklags
Technical University of Munich
Garching, Germany
jens.grossklags@in.tum.de

ABSTRACT

Social media platforms generate user profiles to recommend informational resources including targeted advertisements. The technical possibilities of user profiling methods go beyond the classification of individuals into types of potential customers. They enable the transformation of implicit identity claims of individuals into explicit declarations of identity. As such, a key ethical challenge of social media profiling is that it stands in contrast with people's ability to self-determine autonomously, a core principle of the right to informational self-determination.

In this research study, we take a step back and revisit theories of personal identity in philosophy that underline two constitutive meta-principles necessary for individuals to self-interpret autonomously: justification and control. That is, individuals have the ability to justify and control essential aspects of their self-concept. Returning to a philosophical basis for the value of self-determination serves as a reminder that user profiling is essentially normative in that it formalizes a person's self-concept within an algorithmic system. To understand whether social media users would want to justify and control social media's identity declarations, we conducted a vignette survey study (N = 368). First, participants indicate a strong preference for more transparency in social media identity declarations, a core requirement for the justification of a self-concept. Second, respondents state they would correct wrong identity declarations but show no clear motivation to manage them. Finally, our results illustrate that social media users acknowledge the narrative force of social media profiling but do not strongly believe in its capacity to shape their self-concept.

CCS CONCEPTS

• **Social and professional topics → User characteristics**; • **Security and privacy → Social aspects of security and privacy**.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534192

KEYWORDS

Ethics of artificial intelligence, user profiling, personal identity, social media, autonomy.

ACM Reference Format:
Severin Engelmann, Valentin Scheibe, Fiorella Battaglia, and Jens Grossklags. 2022. Social Media Profiling Continues to Partake in the Development of Formalistic Self-Concepts. Social Media Users Think So, Too.. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3514094.3534192

1 INTRODUCTION

Social media platforms enable advertisers to create and target user audiences based on the identification, processing, and analysis of several thousand user attributes such as likes, interests, beliefs, behaviors, relationships, moral convictions, and political leanings [3, 18, 50, 55, 65, 69]. User profiling techniques infer identity claims of users based on views and clicks, visual data such as images and videos, or the number and types of "followers" or "friends" [11, 18, 20, 21, 39, 45, 47, 70]. There is growing recognition in user profiling and user modeling communities that such profiling techniques create unique ethical challenges [3, 30, 63].

These challenges typically fall back on the inability of users to access, understand, and contest automatically-generated identity claims based on their personal data. Specifically, they arise from the restricted ability of social media users to exercise their right to informational self-determination, a central right of many privacy laws around the world. The right to informational self-determination rests on the fundamental idea that it is critical for individuals to freely and autonomously "self-determine" or "self-develop" [8, 36, 46, 49, 59]. The right to informational self-determination mandates that it is critical for individuals to be able to exercise control over their personal information. In the face of technologies that analyze the sentiment of users based on speech or visual data [18, 19, 57, 58] or that interpret data that users have shared unintentionally [2], the notion of individual control over personal data as a feasible mechanism for informational self-determination is, however, severely challenged.

In this paper, we offer a partly philosophical and a partly empirical account to address this problem field. From a philosophical perspective, we aim to make the following two contributions. First, we return to scholarship on the fundamental value of autonomous

self-determination offered by philosophical theories of personal identity. Philosophical theories of personal identity conceptualize necessary *procedural criteria* that enable an individual to form a self-concept. Personal identity is an essentially contested concept and, as such, inherently procedural—disputes on the concept's boundaries are essential to the concept itself [42, 54].[1] In contrast, when essentially contested concepts become subjected to digital formalism, they are fixated by definitions that work optimally only under the constraints of computability. The analysis of theories of personal identity can illustrate to us, perhaps again, the enormous power of social media user profiling in determining all procedural elements that exist between personal data and their analysis as declarations of identity: the power to create user profiles over time, the power to change or correct user profiles when needed, as well as the power to change the rules by which user profiles can be generated, changed, or corrected.

Second, the generation of digital representations of personal identity necessarily creates normative trade-offs. We present one normative trade-off by referring to what we call "model fitness." Here we ask whether the digital representation of an individual's self-concept *should* align as much as possible with how a person would self-determine in order to respect that person's autonomy. Social media platforms have the power to decide what types of data and what amounts of data are sufficient to justify an identity claim about a user. Social media platforms control "model fitness." We exemplify this phenomenon by referring to the literature on "window sliding" in learning tasks with concept drift adaptation [26, 38, 41, 71] and collaborative filtering [29].

We further take it that the power of social media profiling to make identity claims about billions of users is a strong argument in favor of usable transparency that allows users to view (understand their justification) and correct (exercise control over) such identity claims. Here, we engage in another trade-off: if people could view and correct identity claims of social media profiling, then such identity claims could influence a person's self-concept. Social media identity claims could undermine a person's autonomy to self-determine under conditions of transparency when people see, reflect on, and internalize "how a machine interprets" them. Transparency could empower social media identity claims rather than people's autonomy to self-determine.

Subsequently, we have conducted an empirical vignette study to understand how individuals (N = 368) evaluate social media's identity claims with regard to accuracy, transparency, and control. We find that people believe social media user profiling can make accurate judgments about essential aspects of their personal identity, but that they prefer privacy over accuracy. Moreover, people show a strong desire for transparency defined as the ability to view and understand declarations of identity by social media platforms. While people state that they want to compare whether such identity declarations align with their own self-concept, they believe that these do not influence their self-concept. Our study provides evidence that people assert that social media identity claims do not feed back into their own self-concept when they are made transparent and intelligible.

With this work, we seek to contribute to scholarship on the relation and interaction between humans and their algorithmically generated identity declarations. We provide a philosophical lens on the value of self-determination as the process to justify and control essential aspects of a person's self-concept. The conceptualization of autonomy through personal identity creates a firm foundation for determining the ethical challenges of social media user profiling. With a vignette survey study, we take a tangible step towards understanding how people actually evaluate algorithmic identity declarations by social media platforms.

Before we move on to the next section, we would like to offer a disclaimer: In this work, we do not claim that social media user profiling *generates* personal identity or suggest that the resulting profiles can be considered as equal to a person's self-concept. We do not engage in arguments that draw an ontological comparison between a user profile and the person behind it. In other words, we do not claim that social media user profiling leads to a user profile that *is* the personal identity of the individual. Rather, we observe that social media user profiling procedures possess a unique, technologically-afforded narrative force that computationally fixates the interpretative potential of a person's self-concept. This fixation creates ethical challenges when user profiling algorithms turn a person's personal data into declarations of identity that a person cannot view, cannot understand, and cannot contest.

2 SOCIAL MEDIA USER PROFILING IS FUNDAMENTALLY NORMATIVE

Our analysis considers user profiling procedures for *social media* advertisement. All major social media platforms offer a marketing page with an interface[2] where marketers can select desirable user attributes[3] for targeted advertising.

Previous work on social media profiling has summarized what kind of user attributes social media profiling generates. Such profiles consist of user inferences based on online data (e.g., user-generated content on the platform) as well as offline data (e.g., data integrated from data brokers) [3, 65]. User profiling for social media generates sophisticated representations of users based on demographic information including age or gender as well as information associated with user behaviors, preferences, and intentions [1, 12, 16, 27, 68]. Inferences are in part based on "explicit identity claims" (e.g., explicitly *stated* profession or sexual orientation) as well as on "implicit identity claims." Implicit identity claims are "given off" by an individual rather than consciously communicated [56, 62]. Implicit identity claims are inferences users communicate indirectly, for example, through their affiliations to certain individuals, social or institutional groups, preferences, and interests expressed in a non-specific manner. Explicit and implicit identity claims can comprise behaviors (e.g., clicks or views) and beliefs (expressions of interest, intentions, convictions, etc.) [3, 65]. Social media targeting tools offer marketers the option to select an audience (a group of users) based on whether they "possess" or do not "possess" a desirable attribute. Aimeur has provided a comprehensive list of the types of attributes (i.e., identity claims) analyzed for user profiling including

[1] Please note that this account focuses exclusively on Western approaches to philosophical theories of personal identity.

[2] See, for example, Meta audience insights or Instagram audience insights.

[3] We refer to such user attributes as "declarations of identity."

name, age, address, identity of friends, sexual orientation, political views, smoker yes/no, pregnancy/wedding, interests, credit score, home value, and others [2]. To understand the normative dimensions of user profiling on social media, the technological *instantiation* of a user's profile, for example as a feature vector [7], is not significant for this analysis. What is relevant is the algorithmic mapping function implemented to assign attributes to users based on their data. Any mapping process from user data to user inference digitally fixates the interpretative potential of an individual user. We refer to this process as the generation of a *formalistic self-concept*. By essentially determining this interpretative potential within an algorithmic frame, mapping functions become normative, for example, when they prioritize user data to constitute an attribute while failing to consider others.

In philosophy, a person's self-concept is procedural, contextual, and contestable [5, 10, 23]. Recent work in Science and Technology Studies has outlined that profiling socially contested concepts through mathematical formalism without accounting for their full meaning creates so-called abstraction and formalism traps [54]. Abstraction and formalization necessarily involve a process of imperfect translation: no model (or profile) is large enough to include all characteristics of an informational object. Similarly, in philosophy, no single theory of personal identity contains all constitutive principles that make up personhood. Indeed, it is the disagreement on fundamental conceptual features that creates the essential demarcations of a contested concept such as freedom, privacy, autonomy and so on [42]. In user profiling for social media advertisement, abstraction is constrained by two core conditions: First, by the purpose for which the object is profiled—here for commercial purposes (marketing)—and, second, by the mathematical constraints of computability. Regarding the latter, not all features of an object can be modeled by computational resources; for example, the phenomenological experience of human consciousness cannot— in principle—be captured by computational means.[4] Overall, philosophical theories of personal identity offer a useful conceptual framework to understand the normativity of generating formalistic self-concepts.

3 JUSTIFICATION AND CONTROL: TWO META-PRINCIPLES OF PERSONAL IDENTITY

In the following section, we detail how three influential theories of personal identity lay out procedural criteria that enable a person to form a self-concept autonomously.[5] Attributable to philosophical scholarship, such procedural requirements are subject to productive dispute. Yet, a body of philosophical scholarship on personal identity [22, 51–53, 60, 61] agrees on two constitutive meta-principles necessary for individuals to self-interpret autonomously: individuals have the ability to *justify* and *control* essential elements of their self-concept.[6] Some philosophers place the source of individuals' abilities to justify and control essential aspects of their self-concept in the individual only (e.g., [10, 37]); other theorists argue that social agents partake in the formation of a self-concept [51, 52, 60, 61].

3.1 Harry Frankfurt's second-order desires

In "Freedom of the Will and Concept of a Person," Harry Frankfurt developed a notion of personal identity grounded in the structure of human will [22]. Humans are capable of evaluating the desirability of their desires. A person also cares about the desirability of their desires. Frankfurt calls such desires "second-order desires" that are desires about desires or wants about wants. The object of a first-order desire is a state of affair, while a second-order desire's state of affair is a first-order desire. The desirability of our desires is ethically significant. For example, a person can want to want to eat in a certain way. Vegetarianism, an ethical principle, governs how a person acts on their first-order desire to eat. Frankfurt argues, "*only humans are capable of reflective self-evaluation manifested in the formation of second-order desires*" [22]. The essence of a person lies in will, however, a person needs to be able to "*become critically aware of their own will*" [22]. Individuals need critical reflection to evaluate which of their desires are desirable. Persons are autonomous in determining which desire they want to be moved by when acting. Repeated identification with a specific second-order desire enables us to truly care for something.

Frankfurt's theory of personal identity clearly presents a strong ideal of what it means to be a person. Individuals are required to engage in reflective justification of their second-order desires to fully qualify as persons. There is little room for ambiguous or even paradoxical desires that clearly constitute human experiences. Frankfurt's conception of personhood is an example of a theory from "within": his principles of personal identity are subjective and can even be criticized as "solipsism." External influences, cultural or social, appear to restrict rather than help strengthen individuals' ability to form a self-concept. Summarizing, Frankfurt's second-order desires stress the need for *justifying* one's self-concept, while the identification with a second-order desire underscores that persons can *control* what principles constitute their self-concept.

3.2 Charles Taylor's weak and strong evaluator

The philosopher Charles Taylor deliberately tries to avoid "solipsistic tendencies" and points to the importance of social interaction for the development of a self-concept. Taylor stresses the significance others have for our capacity to evaluate what we desire [60]. Many of our desires, wishes, hopes, attitudes, goals and so on develop only in dialogue with others. Taylor places personal identity between private and public spheres: Privately, a human being is a person because of their reflective self-evaluative capacities that require qualitative articulacy. Publicly, a person necessarily adopts such qualitative articulacy by interaction with other individuals.

[4] Theories on the phenomenological self by Dan Zahavi [67] develop a notion of personal identity that falls back on phenomenological experience.

[5] Personal identity conceptually differs from theories of *personality*. An account of personality is, for example, the prominent Big-Five (BFM) model of personality [32]. The BFM subscribes to personality theories that suggest personality to consist of context-consistent, quantitatively-assessable, enduring traits. In contrast, personal identity explains *how* individuals come to form a persistent self-concept. While such a self-concept may comprise a set of traits, it is the set of principles by which an individual's self-concept develops that is the focus of philosophical theories of personal identity.

[6] Other conceptualizations of hermeneutic personal identity also highlight—in some way or another—the importance of the two meta-principles of justification and control for a person's self-concept (see, for example, [5, 24, 64]). However, they motivate these principles with a different set of reasons.

Similar to Frankfurt's first-order desires and second-order desires, Taylor distinguishes between so-called "weak" and "strong evaluators"[7]. A weak evaluator simply deliberates different options on the basis of their convenience: their goal is to get the most overall satisfaction. Such an evaluator does not reflect on the qualitative aspects of their choices. Non-qualitative evaluation leads to the selection of a desired object or action because "*of its contingent incompatibility with a more desired alternative*" [61]. A weak evaluator chooses something merely on circumstantial grounds. Their deliberation does not exceed a mere desirability calculation for choices to provide some satisfaction. Taylor claims that persons can evaluate what they are and shape whatever they wish to be on this basis. Different from Frankfurt, however, the freedom to self-interpret takes place between private and social spheres. This freedom (i.e., control) to self-define by evaluation (i.e, by justification) means that persons can be made *responsible* for their self-concept [61].

3.3 Maya Schechtman's narrative self-constitution view

The philosopher Marya Schechtman asserts that an autonomous person has the capacity to psychologically organize a stream of events into a culturally accepted form of a narrative "*by which we will come to think of ourselves as persisting individuals with a single life story*" [52]. The elements of a narrative that a person can articulate constitute the person to a higher degree than those elements that a person cannot articulate.

An individual compares, organizes, and relates experiences by culturally-determined standards. It follows that no time-slice—any momentary event that an individual experiences—is in any way definitive for a person's identity. Only when interpreted in the context of the narrative is such a time-slice a meaningful element of a person.[8] Telling a story is only one element of a person's narrative. Individuals form a narrative, but they also enact it and subsequently criticize it: they are not only the authors of their narrative but their protagonists and critics, too. As an author, a person tries to understand the meaning events have by integrating them into their continuous narrative. A person is the critic of their narrative when they come to reflect, evaluate, and criticize the actions they have carried out. While the order in which these steps take place is certainly dynamic, it demonstrates that a person plays different roles within their own narrative—they are not simply describing what they have experienced as a commentator or storyteller in the literal meaning of the term. For Schechtman, a person's narrative is actively *negotiated* between subjective and objective accounts. A person may have their own interpretation of a certain event; however, their identity will be undermined if claims reach a level of incomprehensibility for other people. A person's choices and actions must "*flow intelligibly from (their) intentions, motives, passions, and purposes...*" [52]. Without our narrative context, other individuals cannot make sense of our choices and actions. The narrative view gives individuals freedom to shape (i.e., control) who they wish to be, re-interpret their past and anticipate their future self-concept (i.e., justification). A person's social environment holds a person accountable for the narrative they articulate.

Summary of philosophical theories of personal identity: While differences exist between the theories by Frankfurt, Taylor, and Schechtman, two meta-principles can be discerned: justification and control. First, a self-concept develops through *reflective justification*. Individuals become persons when they justify their self-concept—through reflective capabilities and in a narrative that is negotiated between subjective and objective accounts. Second, individuals can exert some control over their self-concept. While the theories disagree over the degree of control individuals have in forming an understanding of themselves, fundamentally, they all suggest that personhood is grounded in an individual's autonomy to determine essential aspects of their hermeneutic identity. It is for this reason that persons can justifiably be held responsible for their own identity.

4 TWO NORMATIVE TRADE-OFFS IN USER PROFILING FOR SOCIAL MEDIA MARKETING

We argue that social media profiling generates digitally formalized identity claims of a person by mechanisms that do not sufficiently allow for justification and control. In the following, we discuss two normative trade-offs that result from the inherent normativity of social media user profiling as discussed in Section 2.

4.1 Normative trade-off 1: The privacy versus model fit trade-off

4.1.1 Concept drift challenges. One normative judgment user profiling is necessarily required to make is to determine when enough data (or evidence) has been collected and analyzed to justify the inference of a person's attribute (i.e., an identity declaration). It is a normative undertaking to decide when the amount of personal data is sufficient to ensure proportionality between the user input and the attribute inference. Is the inference proportional to a single activity or expression of belief? Or is its proportionality dependent on multiple consecutive expressions of the belief? Resolving such questions, user profiling necessarily excludes user input from being considered for drawing user inferences. Schechtman asserts that individuals have the capacity to attribute meaning to a selection of experiences that become part of their own unique narrative. However, it is the narrative that is self-constituting, not the single experience. It follows that no time-slice—any momentary event that an individual experiences—is in any way definitive for a person's identity. Such a time-slice is only a descriptive and meaningful element of a person when interpreted in the *context* of the narrative.

Schechtman's concept of a "time-slice" can be compared to the concept of "window sliding" used in learning tasks with concept drift adaption [26, 38, 41, 71]. Concept drift techniques are deployed to gain knowledge from data stream *changes*. Drifts or changes in a data stream can be either sudden or gradual. The former could be a sudden new interest in a new subject, while the latter could be a growing interest in moving to another country. In user profiling, concept drift belongs to a class of challenges called dynamicity problems [48, 66]. Recommender systems apply dynamic user profiles to offer more value to the user, who sees informational resources

[7] Arguably, a person that chooses merely on the basis of Frankfurt's first-order desires corresponds to Taylor's weak evaluator.
[8] "*Whether or not a particular action, experience, or characteristic counts as mine is a question of whether or not it is included in my self-narrative*" [25].

they have only recently become interested in, and to the advertiser that can bid for audiences with the most up-to-date profile.

Machine learning (ML) classifiers are able to respond to concept drift—gradual, sudden, or reoccurring changes often in multiple data streams—without "neglecting" the outdated data [71]. For example, sliding windows of fixed and variable sizes of training data are used to build an updated model [26]. Since both fixed and variable windows are definite in their size, some old data will necessarily be "forgotten." What criteria determine which data are to be forgotten and which ones are to be considered in creating an updated profile of a person? The promise of targeted advertisement rests on the belief that more recent user data corresponds to a more accurate profile of the user. However, model fit, a continuously updated model of a user's profile, requires a potentially uninterrupted flow of user data, raising privacy concerns [13]. The more time-slices are created, the more accurate the representation of the user, but the more user data is needed.

4.1.2 Lookalikes through Neighborhood-Based Collaborative Filtering. Collaborative filtering (CF) is one of the most widely applied user modeling techniques in many recommender systems. For example, as a user profiling technique, *k*-nearest neighbor relies on the assumption of similarity between individuals [29]. Similar profiles presumably react similarly to certain informational items. The advantage of CF is that one only requires a model of one of the two—users or items—to model the other. Consequently, CF uses items to model users and users to model items. The more users evaluate informational resources, the more they help the system for its predictive analysis of other users. Social media (as well as search engines) offer their customers so-called "Lookalike Audiences."[9] With many marketers, Lookalikes are popular since they can use their well-known customer base to target "similar" but potentially new customers. Lookalikes are less privacy-invasive because they use data that is already available to make inferences about a user. Taylor's and Frankfurt's concept of a person, however, stresses the ability of persons to decide what is *desirable for them*. *k*NN-based CF and Lookalikes work in the opposite way. They determine the desirability of one's desires as equal or at least similar to the desirability of other, already "known" individuals' desires, to use Frankfurt's nomenclature.

4.2 Normative trade-off 2: The transparency versus autonomy trade-off

A key question is if people would actually care about model fit—an accurate representation of their *formalistic* data narrative. Perhaps individuals do, after all, live in the best of all possible worlds: they draw enormous benefits from using social media and do not worry about how their data is mapped to a spectrum of attribute inferences. One way forward would be to enable individuals to understand and correct inferences they do not agree with. Here, another normative complication emerges. A person could gain autonomy from having access to their social media's identity declarations. However, these identity declarations could in turn influence a person's self-concept.

Should individuals get access in order to understand and contest their "data narrative"? Providing explanations on "how the systems works" has shown to increase users' trust in many different recommender systems [9, 14, 17, 44]. Usable transparency allows users to tell the system when an inference is presumptuous (or even wrong). For example, a system could show users those identity declarations that have been sold to marketers or that were based on implicit identity claims. However, simply revealing—at least in part—the content behind user profiles could support internalization and conformation to the proposed inferences. Perhaps individuals would welcome such a degree of transparency as a mechanism to "offload" the psychological work necessary to attribute meaning to certain life events posted online [51–53]. Making inferences transparent to the individual means recognizing their semantic power in shaping who individuals are and who they can become. This second normative trade-off arises from the question of whether the autonomy gained from being able to understand such recommended inferences outweighs a potential loss of autonomy when they become part of a person's self-concept. This could mean that, today, a person, their social network (offline and online), *and* social media profiling identity declarations together participate in creating a person's self-concept.

The effect on individuals' self-concept could be enhanced if social media user profiling generates specific identity declarations repeatedly or even permanently. According to Frankfurt's theory of personal identity, a person attempts to form a self-concept that stems from their care for what they desire. Frankfurt recognizes that one can only care about something if it is for extended periods of time. Desires typically last for moments only: if one cared about something for only a moment one could not be distinguished from a person that acted out of impulse. How would users perceive such recommended attribute inferences? Perhaps with little skepticism, since they would acknowledge the algorithmic output as an objective and truthful interpretation of their wishes, wants, and desires?

5 METHODS AND EXPERIMENTAL PROCEDURE: VIGNETTE STUDY

To address the key questions arising from both normative trade-offs, we conducted a vignette study that asked respondents a) whether they believed social media profiling could accurately infer elements of their self-concept, b) whether they considered accuracy of these identity declarations to be desirable, c) whether they had motivation to view and correct identity declarations, and d) whether they believed that social media identity declarations would influence their self-concept if they were made transparent to them. The goal of the vignette study was to take a tangible step towards understanding whether social media users preferred accuracy of social media identity declarations over privacy (trade-off 1) and whether they believed that social media identity declarations would influence their self-concept (trade-off 2). Vignette studies have been extensively used in human computer interaction, psychology, and experimental philosophy to elicit participants' explicit ethical judgments in various hypothetical scenarios [4, 6, 15, 28, 31, 33, 34, 40, 43]. Moreover, with our vignette survey study, we follow calls for more experimentally-informed AI ethics [35].

Our study was a within-subject design, we presented each respondent with the same hypothetical vignette scenario. First, the

[9] See, for example: https://www.facebook.com/business/help/164749007013531 accessed May 30, 2022.

vignette asked respondents to imagine that they are active users on a social media platform (see the hypothetical vignette scenario in Appendix A.1). As an active user, each respondent was told that they regularly engage in typical actions on the social media platform. Participants read that they publish postings, share postings by other users, and react to other users' postings. Second, the vignette introduced examples of data types each respondent shares with the social media platform (gender, location, relationship status, social contacts, content viewed, content clicked, etc.). Respondents were told that the social media platform uses algorithms to draw conclusions about them based on the data they share in order to show them more suitable content and advertisements. Third, the vignette elaborated on the types of conclusions (i.e., identity declarations) that the social media platforms draws about them. The vignette explained that the platform collects data that users actively share to draw conclusions about them. For example: *"...when you provide your real birthday, the platform uses this information to show you content that it takes to be suitable for your age group."* Respondents were also told that the platform draws conclusions about users based on data that users may not be aware that they are sharing. For example, *"...since you share your location data, the platform tries to conclude where you work and live. As another example, the platform also tries to conclude what hobbies you have based on your friends' activities."* Respondents were further told that, using their data, the social media platform attempts to conclude their interests, their political orientation, their religious beliefs, and aspects of their personality (among others). Lastly, we asked respondents to imagine that the platform "combines and stores" all conclusions about them in a so-called "user data profile" (UDP). The vignette explained to users that the social media platform uses the content of their UDP to recommend relevant information and advertisements. The hypothetical vignette scenario ended by telling respondents that the social media platform generates all of its revenues from personalized advertisement. We included two attention checks in the vignette. All participants were active social media users.

After respondents had read the vignette and passed the attention checks, they rated questions using a 7-point Likert scale. Questions were divided into 5 categories and shown to respondents in random order within these categories. The first two categories of questions asked respondents whether they believed social media platforms could make accurate judgments about them and whether the social media platform *should* make accurate judgments about them. We defined accuracy as a) general judgments, b) specific judgments, and c) temporal judgments. The third and fourth set of questions asked whether respondents desired to view and understand social media judgments about them and whether they would change incorrect judgments. Questions on respondents' preference for transparency included a) data collection & use, b) preference for understanding conclusions of the social media platform, and c) preference for transparency of their UDP (i.e., all identity declarations). Finally, a fifth set of questions asked respondents whether social media judgments would have an influence on their self-concept given that respondents could view their UDP. We defined "influence" as respondents' willingness to a) compare elements of their UDP with their self-concept, b) their willingness to reevaluate their self-concept in light of the identity declarations in their UDP, and c) their willingness to integrate elements of their UDP into their self-concept that they would not have associated with their self-concept. All questions are listed in Appendix A.3.

We recruited participants with Prolific. Based on pretests, we set the expected completion time at 20 minutes, with a payout of USD 3.75 (above US minimum wage of 2021). Data collection started on July 26, 2021 and ended on August 8, 2021. We recruited 458 respondents from the United States user base. 59 submissions were excluded for failing one of two attention checks, 10 for duplicate submissions, 9 for an unusually short response time, and 11 for being invalid (e.g., no prolific ID). This resulted in a final sample of 368 respondents (see demographics in the Appendix A.4). The mean time of completion was 15.3 minutes.

Our home institution does not require an ethics approval for questionnaire-based online studies. When conducting the study and analyzing the data, we followed standard practices for ethical research: presenting detailed study procedures, obtaining consent, not collecting identifiable information or device data, and using a survey service[10] that guaranteed compliance with the European Union's General Data Protection Regulation. The study did not include any deceptive practices. Subjects could drop out of the study at any point. All data were fully anonymized, and the privacy of all subjects was maintained at all times during the study.

6 RESULTS

Respondents' beliefs on the ability of social media platforms to make accurate judgments about them (Fig. 1a). A majority of respondents believed that social media algorithms could make accurate and correct judgments about them *in general* (78.2%). While 66.2% of respondents were convinced that social media algorithms could correctly judge "what is valuable to them," just over half of respondents said that social media algorithms can accurately reflect who they are (51.1%). Most respondents believed that their UDP was unique in comparison to other social media users (72.6%). However, only a minority of respondents said that family and close friends would be able to identify them by their UDP (45.5%).

Respondents' beliefs on the ability of social media platforms to make accurate judgments about them on specific attributes (Fig. 1b). Respondents believed that social media algorithms can accurately infer their interests (89.9%), their past (81.3%) and future purchasing behaviors (64.5%), as well as their location (77.4%). Just over half of those surveyed stated that social media algorithms could accurately conclude who they meet (54.8%).

Respondents also said that social media algorithms are able to accurately conclude their political stance (80.5%) and, albeit with less agreement, their religious beliefs (59.5%). Most respondents agreed that social media algorithms can correctly infer their attitudes towards the COVID-19 vaccine (77.9%), climate change (74.6%), and immigration (64.8%). However, respondents did not think that social media profiling was able to differentiate between their private and social self both online (35.5%) and offline (30.7%).

Respondents' beliefs on the ability of social media platforms to make accurate temporal judgments about them (Fig. 1c). Respondents believed that social media algorithms are able to keep their UDP up to date (71.4%). Respondents stated that

[10] SoSci Survey: https://www.soscisurvey.de/

Figure 1: (a) Respondents believe social media platforms (SMP) can make accurate judgements about them. UDP—user data profile. (b) Respondents believe social media (SM) algorithms are able to accurately infer a variety of attributes including their interests, purchases, location, political stance, or religious beliefs. Respondents do not believe SMP is able to distinguish who they are in private vs. who they are in social contexts. (c) Respondents believe SMP is able to keep their UDP up to date, but that their UDP does not tell an accurate story of their life. Note for all figures: results for "strongly agree" and "agree" are shown as "agree," results for "strongly disagree" and "disagree" are shown as "disagree."

their UDP from a month ago still included accurate conclusions (69.1%). However, just over half of respondents thought that their UDP from a year ago was still accurate (51.4%). A majority of respondents said that the social media platform would be able to conclude whether they had changed as a person after several years of being a user (68.9%). In contrast, only a minority of respondents believed that their *entire UDP* would tell an accurate story of their life since they started using the platform (37.9%).

Respondents' beliefs on the normativity of accurate social media judgments (Fig. 2). Most respondents stated that they wanted social media platform operators to ensure that their UDP was accurate (72.4%). Just more than half of respondents wanted social media operators to invest extra resources to make sure their UDP was accurate (56.9%). However, only a minority of 28.8% of respondents were in favor of trading their personal data for the creation of their UDP. Importantly, respondents did not want to

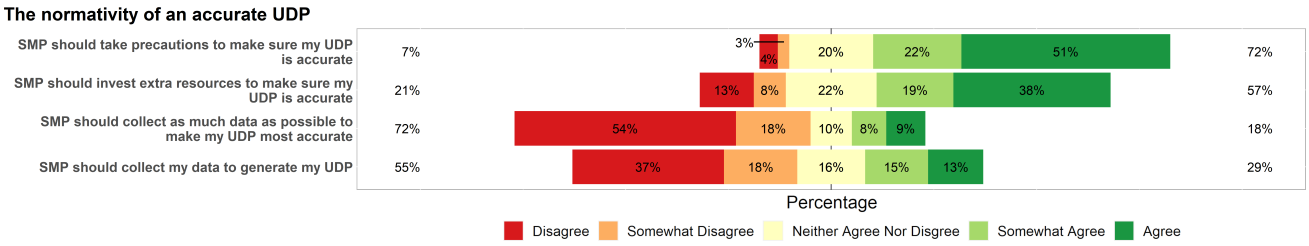


Figure 2: Respondents prefer an accurate UDP but not at the expense of their privacy.

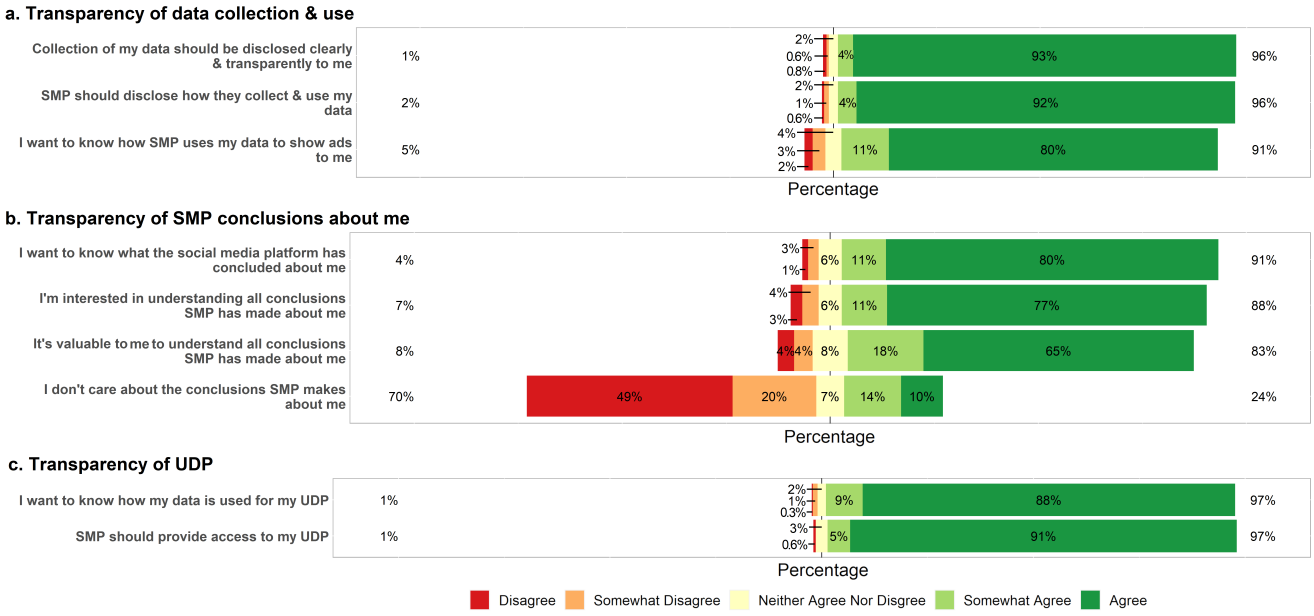


Figure 3: Respondents show great preference for transparency of (a) personal data collection & use, (b) conclusions SMP has made about them, and (c) of their UDP.

trade their personal data for an accurate UDP: only 17.9% agreed that the social media platform should collect as much personal data as possible to ensure that their UDP was as accurate as possible.

Respondents' preference for transparency of data collection & use (Fig. 3a). Respondents expressed their desire for transparency of personal data collection on social media, transparency of conclusions the social media platform made about them based on their data, and transparency of their UDP. Regarding data collection, most respondents stated that procedures of data collection should be disclosed clearly and transparently to them (96.4%) and that the social media platform should disclose how they collected and used their personal data in general (96.1%) and for showing advertisements (91.1%).

Respondents' preference for transparency of conclusions (Fig. 3b & c). Similarly, respondents showed a strong preference to understand what the social media platform has concluded about them (90.1%). Of the respondents, 87.7% stated that they were interested in understanding all conclusions the social media platform had made about them and 83.2% believed that such an understanding would be valuable to them. Only 24% of respondents stated that they do not care about conclusions the social media platform draws about them. Finally, similarly large majorities of respondents expressed their desire to understand how their personal data was used to create their UDP (96.7%). Of the respondents, 96.6% said that they wanted access to their UDP in general (Fig. 3c).

Respondents' preference for control over their UDP (Fig. 4). While respondents showed a clear preference for transparency, their desire to control (i.e., change or otherwise influence) their UDP was mixed. A majority stated that the social media platform should allow them to correct errors in their UDP (90.2%). However, only a small majority said they would be motivated to change wrong conclusions in their UDP (60.2%). When we asked whether correcting and maintaining their UDP would be "too tedious," respondents showed no clear preference (agree: 37.8% vs. disagree: 45.4%, neither: 16.8%). Approximately half of respondents (57.3%) believed they

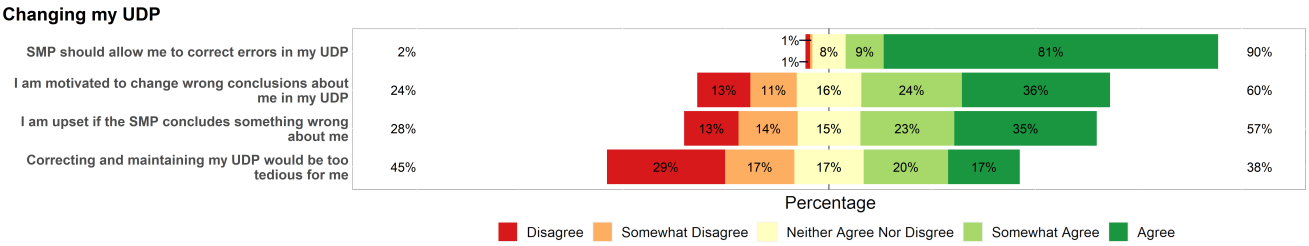


Figure 4: Respondents state that the SMP should allow them to correct errors in their UDP but provide no clear preference on whether they would be willing to correct and maintain their UDP.

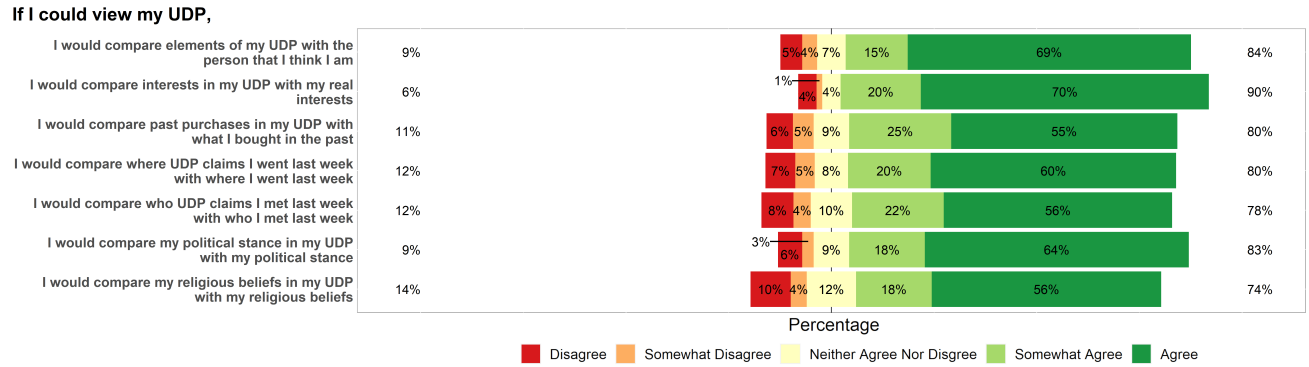


Figure 5: Provided their UDP was transparent, respondents would compare elements of their UDP with a range of personal attributes.

would be upset if the social media platform concluded something about them that they thought was incorrect.

Respondents' beliefs on the influence of the UDP on their self-concept (comparison UDP vs. self-concept, Fig. 5). Provided they had access to their UDP, the majority of respondents maintained that they would compare elements of their UDP with the person they thought they were (84.1%). Most respondents said they would compare interests in their UDP with their real interests (89.7%). Respondents further stated they would compare past purchases (79.9%), past locations (79.9%, "last week"), and past social meetings (77.9%, "last week") with those in their UDP. Among the respondents, 82.7% would compare their political stance with the one registered in their UDP and 74.3% of respondents would compare their religious beliefs with those in their UDP.

Respondents' beliefs on the influence of the UDP on their self-concept (reevaluation of self-concept, Fig. 6). Only a minority of respondents believed that viewing their user data profile would result in a reevaluation of their self-concept (agree: 21.5%). Few respondents stated that they would reevaluate their interests (agree: 17.3%), their future purchases (agree: 26.8%), their political stance (agree: 11.5%) or their religious beliefs (agree: 9.22%) after viewing their UDP.

Respondents' beliefs on the influence of the UDP on their self-concept (meaning of unaware identity declarations in the UDP, Appendix Fig. 7). Respondents were undecided whether social media conclusions were meaningful to them (agree: 47.3% vs. disagree: 35.6%). A small majority of respondents disagreed that conclusions about them in their UDP—that they did not know about—would be meaningful to them (disagree: 53.8%). A small majority of respondents also objected to statements saying conclusions about their political stance (disagree: 55.0%) or religious beliefs (disagree: 59.9%) in their UDP—that they did not know about—would be meaningful to them. Finally, we asked respondents whether their UDP would be a source of inspiration when looking for a *new* interest. Only 41.1% of respondents said that they would look into their UDP for suggestions on new interests. Likewise, respondents believed that when they saw an interest in their UDP that they would not have believed to be their interest, then this "recommended" interest would not become a new interest for them (agree: 36.6%). An even smaller minority of respondents said that predicted purchases in their UDP would influence actual future purchases (agree: 34.6%).

7 DISCUSSION OF RESULTS AND CONCLUDING REMARKS

In this work, we argued that the computability of digital representations of personal identity creates normative trade-offs when social media profiling generates identity claims that work only under the constraints of computability and that people cannot understand, view, or contest. Consequently, one of the key ethical challenges of social media profiling is that it stands in contrast with people's ability to self-determine freely and autonomously. To illustrate the

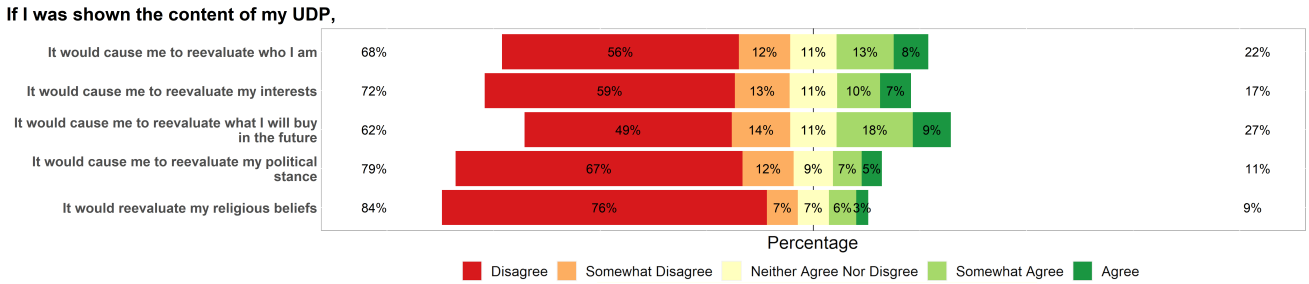


Figure 6: Respondents strongly believe that viewing the content of their UDP would not cause them to reevaluate elements of their self-concept.

inherently procedural nature of autonomous self-determination, we revisited theories of personal identity in philosophy that underline two constitutive meta-principles: justification and control. That is, individuals have the ability to *justify* and *control* essential elements of their self-concept. The return to the philosophical basis for the value of self-determination serves as a reminder that social media profiling represents an inherently normative formalization process of a person's self-concept. Within the interpretative space between data and declaration, social media platforms determine the meaning of views, clicks, posts, and social relationships without offering usable means for understanding or correcting essential parts of this process. As such, social media identity declarations are radically different from the procedural criteria laid out by theories of personal identity in philosophy.

Taking a step toward understanding how "ordinary" social media users view social media identity declarations, we conducted a vignette survey study. We found that people believe that social media platforms can make a variety of accurate judgements about them but that they cannot represent their entire self-concept. For example, respondents thought that social media profiling is able to accurately infer whether they have changed as a person over time, but that it cannot tell an accurate story of their life since signing up to the platform. Thus, respondents defined limits for the ability of social media identity declarations to represent certain aspects of their self-concept. Interestingly, respondents did claim that their own user data profile (UDP) was unique and that other users had a different UDP.

Respondents showed a strong preference for more transparency and stated that they would compare their own self-concept with a variety of social media identity declarations. However, the respondents in our study did not believe that social media identity declarations would be meaningful to them. Respondents also stated they would correct wrong identity declarations but showed no clear motivation to manage them. Taken together, we believe that it is reasonable to assume that social media users have at least some motivation to control essential aspects of their social media identity declarations. Providing such identity controls does present technological as well as design challenges for social media platform operators. However, social media platforms go to great lengths to offer advertisers usable controls to specify which user attributes exactly they wish to include in their custom audiences. In providing usable justification and control, social media platforms give priority to advertisers determining detailed custom audiences for targeted advertisement over giving users the possibility to understand, control, and rectify potential inaccuracies in their user profiles.

Finally, respondents did not believe that social media identity declarations would *influence* their self-concept. Respondents stated that previously unknown identity declarations would be unlikely to become part of their self-concept and they strongly objected that viewing social media identity declarations would cause them to reevaluate their self-concept. Future studies should try to understand whether people's self-concept is resilient to social media identity declarations as participants stated in our study. Perhaps people are overconfident in the immunity of their self-concept against social media declarations? Also, a majority of respondents expressed the desire to compare components of their UDP with their self-concept. Considering our results, we take it that people are, at least, curious to understand how social media platforms interpret them based on their personal information. They acknowledge the narrative force of social media profiling but do not strongly believe in its capacity to shape their self-concept. We encourage future studies to explore whether our findings extend to social media users in other cultures.

To conclude, we have focused on the *process* by which social media generate identity declarations based on personal information through user profiling. In comparison to the large corpus of studies that have focused on the *consequences* of user profiling (e.g., filter bubbles, misinformation), philosophical accounts on the procedural aspects of social media user profiling remain scarce. While our vignette study produces an initial understanding of the relationship between social media users and their identity declarations, we expect that this account provides ample opportunity for follow-up studies on the ethical challenges of social media profiling. Social media will continue to exercise its power to partake in the formation and development of formalistic self-concepts. We provide evidence that social media users think so, too.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their constructive feedback and comments. This research was supported by a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in any part of the research or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

REFERENCES

[1] Ahmad Abdel-Hafez and Yue Xu. 2013. A survey of user modelling in social media websites. *Computer and Information Science* 6, 4 (2013), 59–71. https://doi.org/10.5539/cis.v6n4p59

[2] Esma Aïmeur. 2018. Personalisation and privacy issues in the age of exposure. In *Conference on User Modeling, Adaptation and Personalization (UMAP)*. 375–376. https://doi.org/10.1145/3209219.3209271

[3] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *Network and Distributed System Security Symposium (NDSS)*. 1–15. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_10-1_Andreou_paper.pdf

[4] Kwame Anthony Appiah. 2008. *Experiments in ethics.* Harvard University Press.

[5] Kwame Anthony Appiah. 2010. *The ethics of identity.* Princeton University Press.

[6] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138. https://doi.org/10.1027/1614-2241/a000014

[7] Joseph Blass. 2019. Algorithmic advertising discrimination. *Northwestern University Law Review* 114, 2 (2019), 415–467. https://scholarlycommons.law.northwestern.edu/nulr/vol114/iss2/3/

[8] Edward J. Bloustein. 1964. Privacy as an aspect of human dignity: An answer to Dean Prosser. *New York University Law Review* 39, 6 (1964), 962–1007. https://heinonline.org/HOL/LandingPage?handle=hein.journals/nylr39&div=71

[9] Jens Brunk, Jana Mattern, and Dennis M. Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*. 429–435. https://doi.org/10.1109/CBI.2019.00056

[10] Sarah Buss and Lee Overton (Eds.). 2002. *Contours of agency: Essays on themes from Harry Frankfurt.* MIT Press.

[11] Buru Chang, Yonggyu Park, Donghyeon Park, Seongsoon Kim, and Jaewoo Kang. 2018. Content-aware hierarchical point-of-interest embedding model for successive POI recommendation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*. 3301–3307. https://www.ijcai.org/proceedings/2018/0458.pdf

[12] Jinpeng Chen, Yu Liu, and Ming Zou. 2016. Home location profiling for users in social media. *Information & Management* 53, 1 (2016), 135–143. https://doi.org/10.1016/j.im.2015.09.008

[13] Michela Chessa, Jens Grossklags, and Patrick Loiseau. 2015. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *IEEE 28th Computer Security Foundations Symposium (CSF)*. 90–104. https://doi.org/10.1109/CSF.2015.14

[14] Jong Kyu Choi and Yong Gu Ji. 2015. Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction* 31, 10 (2015), 692–702. https://doi.org/10.1080/10447318.2015.1070549

[15] Cory J. Clark, Jamie B. Luguri, Peter H. Ditto, Joshua Knobe, Azim F. Shariff, and Roy F. Baumeister. 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106, 4 (2014), 501–513. https://doi.org/10.1037/a0035880

[16] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. 2019. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access* 7 (2019), 144907–144924. https://doi.org/10.1109/ACCESS.2019.2944243

[17] Nadia El Bekri, Jasmin Kling, and Marco F. Huber. 2020. A study on trust in black box models and post-hoc explanations. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)*, Francisco Martínez Alvarez (Ed.). Advances in Intelligent Systems and Computing, Vol. 950. Springer, 35–46. https://doi.org/10.1007/978-3-030-20055-8_4

[18] Severin Engelmann and Jens Grossklags. 2019. Setting the Stage: Towards Principles for Reasonable Image Inferences. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP)*. 301–307. https://doi.org/10.1145/3314183.3323846

[19] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What People Think AI Should Infer From Faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533080

[20] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. 171–179. https://doi.org/10.1145/3159652.3159691

[21] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with Instagram pictures. In *Workshop on Emotions and Personality in Personalized Systems*. 7–10. https://doi.org/10.1145/2809643.2809644

[22] Harry G. Frankfurt. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy* 68, 1 (1971), 5–20. https://doi.org/10.2307/2024717

[23] Shaun Gallagher. 2000. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* 4, 1 (2000), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

[24] Shaun Gallagher (Ed.). 2011. *The Oxford handbook of the self.* Oxford University Press.

[25] Shaun Gallagher and Dan Zahavi. 2020. *The phenomenological mind.* Routledge. https://doi.org/10.4324/9780429319792

[26] João Gama, Raquel Sebastião, and Pedro Pereira Rodrigues. 2013. On evaluating stream learning algorithms. *Machine Learning* 90, 3 (2013), 317–346. https://doi.org/10.1007/s10994-012-5320-9

[27] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. 2007. User profiles for personalized information access. In *The Adaptive Web*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Springer, 54–89. https://doi.org/10.1007/978-3-540-72079-9_2

[28] Armin Granulo, Christoph Fuchs, and Stefano Puntoni. 2019. Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour* 3, 10 (2019), 1062–1069. https://doi.org/10.1038/s41562-019-0670-y

[29] Jonathan Herlocker, Joseph Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 230–237. https://doi.org/10.1145/312624.312682

[30] Mireille Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law* 20, 1 (2019), 83–121. https://doi.org/10.1515/til-2019-0004

[31] Michael R. Hyman and Susan D. Steiner. 1996. The vignette method in business ethics research: Current uses, limitations, and recommendations. In *Proceedings of the Annual Meeting of the Southern Marketing Association*. 261–265.

[32] Oliver John, Laura Naumann, and Christopher Soto. 2008. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research* (3rd ed.), Oliver P. John, Richard W. Robins, and Lawrence A. Pervin (Eds.). The Guilford Press, 114–158.

[33] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194.

[34] Joshua Knobe and Shaun Nichols. 2017. Experimental philosophy. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[35] Steven R. Kraaijeveld. 2021. Experimental philosophy of technology. *Philosophy & Technology* 34, 4 (2021), 993–1012. https://doi.org/10.1007/s13347-021-00447-6

[36] Joseph Kupfer. 1987. Privacy, autonomy, and self-concept. *American Philosophical Quarterly* 24, 1 (1987), 81–89. https://www.jstor.org/stable/20014176

[37] John Locke. 1690. *An essay concerning human understanding.* Printed for Thomas Basset.

[38] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2019. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2019), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

[39] Weizhi Ma, Min Zhang, Chenyang Wang, Cheng Luo, Yiqun Liu, and Shaoping Ma. 2018. Your Tweets reveal what you like: Introducing cross-media content information into multi-domain recommendation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*. 3484–3490. https://www.ijcai.org/proceedings/2018/0484.pdf

[40] David E. Melnikoff and Nina Strohminger. 2020. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour* 4, 12 (2020), 1258–1264. https://doi.org/10.1038/s41562-020-00943-3

[41] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. YALE: Rapid prototyping for complex data mining tasks. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 935–940. https://doi.org/10.1145/1150402.1150531

[42] Deirdre K. Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), Article No. 20160118. https://doi.org/10.1098/rsta.2016.0118

[43] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41, 4 (2007), 663–685. https://www.jstor.org/stable/4494554

[44] Kenya Freeman Oduor and Eric N. Wiebe. 2008. The effects of automated decision algorithm modality and transparency on reported trust and task performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 4 (2008), 302–306. https://doi.org/10.1177/154193120805200422

[45] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Seventh International Conference on Language Resources and Evaluation (LREC)*. 1320–1326. http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[46] Carina Prunkl. 2022. Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence* 4, 2 (2022), 99–101. https://doi.org/10.1038/s42256-022-00449-9

[47] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our Twitter profiles, our selves: Predicting personality with Twitter. In *IEEE Third International Conference on Social Computing*. 180–185. https://doi.org/10.1109/

PASSAT/SocialCom.2011.26
[48] Dimitrios Rafailidis, Pavlos Kefalas, and Yannis Manolopoulos. 2017. Preference dynamics with multimodal user-item interactions in social media recommendation. *Expert Systems with Applications* 74 (2017), 11–18. https://doi.org/10.1016/j.eswa.2017.01.005
[49] Antoinette Rouvroy and Yves Poullet. 2009. The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In *Reinventing Data Protection?*, Serge Gutwirth, Yves Poullet, Paul De Hert, Cécile de Terwangne, and Sjaak Nouwt (Eds.). Springer, 45–76. https://doi.org/10.1007/978-1-4020-9498-9_2
[50] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Alan Mislove, and Aaron Rieke. 2019. Algorithms that" Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences. *arXiv preprint arXiv:1912.07579* (2019).
[51] Marya Schechtman. 1996. *The constitution of selves.* Cornell University Press.
[52] Marya Schechtman. 2011. The narrative self. In *The Oxford Handbook of the Self*, Shaun Gallagher (Ed.). Oxford University Press.
[53] Marya Schechtman. 2014. *Staying alive: Personal identity, practical concerns, and the unity of a life.* Oxford University Press.
[54] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Second Annual Conference on Fairness, Accountability, and Transparency*. 59–68. https://doi.org/10.1145/3287560.3287598
[55] Dusan Sovilj, Scott Sanner, Harold Soh, and Hanze Li. 2018. Collaborative filtering with behavioral models. In *Conference on User Modeling, Adaptation and Personalization (UMAP)*. 91–99. https://doi.org/10.1145/3209219.3209235
[56] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* 48, 2 (2018), 204–231. https://doi.org/10.1177/0306312718772094
[57] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129
[58] Luke Stark and Jevan Hutson. 2022. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* 32, 4 (2022), 922–978. https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2/
[59] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet Policy Review* 8, 2 (2019), 1–22. https://doi.org/10.14763/2019.2.1410
[60] Charles Taylor. 1976. Responsibility for self. In *The Identities of Persons*, Amélie Rorty (Ed.). University of California Press.
[61] Charles Taylor. 1989. *Sources of the Self: The Making of the Modern Identity.* Harvard University Press.
[62] José Van Dijck. 2013. 'You have one identity': Performing the self on Facebook and LinkedIn. *Media, Culture & Society* 35, 2 (2013), 199–215. https://doi.org/10.1177/0163443712468605
[63] Jan Van Gemert, Cor Veenman, Arnold Smeulders, and Jan-Mark Geusebroek. 2010. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 7 (2010), 1271–1283. https://doi.org/10.1109/TPAMI.2009.132
[64] J. David Velleman. 2006. *Self to self: Selected essays.* Cambridge University Press.
[65] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M. Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P. Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *The World Wide Web Conference*. 1920–1930. https://doi.org/10.1145/3308558.3313666
[66] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Xiaofang Zhou. 2015. Dynamic user modeling in social media systems. *ACM Transactions on Information Systems* 33, 3 (2015), 1–44. https://doi.org/10.1145/2699670
[67] Dan Zahavi. 2007. Self and other: The limits of narrative understanding. *Royal Institute of Philosophy Supplements* 60 (2007), 179–202. https://doi.org/10.1017/S1358246107000094
[68] Fattane Zarrinkalam, Hossein Fani, and Ebrahim Bagheri. 2019. Extracting, mining and predicting users' interests from social networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1407–1408. https://doi.org/10.1145/3331184.3331383
[69] Qian Zhao, Martijn Willemsen, Gediminas Adomavicius, Maxwell Harper, and Joseph Konstan. 2018. Interpreting user inaction in recommender systems. In *12th ACM Conference on Recommender Systems (RecSys)*. 40–48. https://doi.org/10.1145/3240323.3240366
[70] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. 2019. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1s (2019), Article No. 14. https://doi.org/10.1145/3233184
[71] Indrė Žliobaitė. 2010. Learning under concept drift: An overview. *arXiv preprint arXiv:1010.4784* (2010). https://doi.org/10.48550/arXiv.1010.4784

A APPENDIX STUDY MATERIALS

A.1 Hypothetical Vignette Scenario

Please read the following scenario carefully:

Imagine that you are an active member of a global social media platform. Think of a social media platform that is similar to a handful of prominent examples such as Facebook, Twitter, or Instagram. Imagine that, on this platform, you are an active member and regularly post content. For example, you frequently upload images to the platform. When your friends publish similar posts, you commonly "react" to their posts. Generally, you often consume the content that the platform presents to you in its so-called "news feed".

More specifically, the data you share with the platform includes your real name, your age, and gender. You also share your current location with the platform, your social contacts, your relationship status, the type of device you use, and your activity data: what content you view and click on when you use the platform and at what time you do so. You are aware that the social media platform has developed algorithms that attempt to draw a variety of conclusions about you based on the types of data you share. The social media platform states that it uses such "conclusions about you" in order to show you more suitable content and product advertisement.

Some conclusions may be based on the data you share actively and consciously. For example, when you provide your real birthday, the platform uses this information to show you content that it takes to be suitable for your age group. Some conclusions about you are based on data that you share implicitly, so you may not be aware that you have shared such data about you.

For example, since you share your location data, the platform tries to conclude where you work and live. As another example, the platform also tries to conclude what hobbies you have based on your friends' activities. The platform attempts to conclude your interests (e.g., movies, music, or books you might like) and your behaviors (e.g., what you buy, who you meet). It tries to conclude your religious beliefs (e.g., whether you are part of a religion or an atheist) and your political stance (e.g., whether you consider yourself liberal or a conservative). The social media platform also tries to draw conclusions about who you are as a person more generally; for example, how you might react to certain content, how introverted or extroverted you are, or how sociable you are.

Now, please imagine that the social media platform combines and stores all conclusions about you in your user data profile. Again, the social media platform claims that it needs the content of your user data profile to know what content and advertisement you find suitable. The social media platform generates all of its revenues by showing you advertisements.

To recap, there are two different user profiles on social media: One profile that you use to share posts or share messages, your profile on the social media platform. The other one is generated by the social media platform about you, which will be referred to as your "user data profile" for the rest of the survey. All survey questions relate to your user data profile, not your social media profile.

You were shown a description of a social media platform. You will now be asked questions regarding your personal perception of social media platforms like the one described previously. All questions relate to a social media platform that was introduced to you in the opening text.

Please answer these questions from your own point of view.

A.2 Manipulation Checks (in-text)

(1) Asked prior to vignette text: It is important that you pay attention to this study. Please read the scenario described below carefully.

- *Please confirm this by selecting "Strongly disagree."*

(2) Asked at the end of the vignette text: Please indicate which of the following is true.

My user data profile is:

- *My social media profile that I use to socialize when I log on to the social media platform.*
- *My profile that the social media platform's algorithms generate about me based on the data I share explicitly and implicitly.*
- *I don't know.*

A.3 Survey Questions

7-point-scale, 1 = "Strongly disagree" to 7 = "Strongly Agree," and "I don't want to answer."

Questions were divided into 5 categories and shown to respondents in random order within these categories. Participants did not see headlines of question categories.

Accurate judgments (general)

- *The social media platform is able to draw correct conclusions about me.*
- *I believe that the social media platform is able to know what is valuable to me.*
- *I believe that my user data profile is unique. Other users have a different user data profile.*
- *If close friends and family saw my user data profile, they would be able to identify that it's me.*

Accurate judgments (specifics)

- *I believe that social media algorithms are able to accurately conclude what my interests are (e.g., movies, music, or books I like).*
- *I believe that social media algorithms are able to accurately conclude what I have bought in the past.*
- *I believe that social media algorithms are able to accurately conclude what I will buy in the future.*
- *I believe that social media algorithms are able to accurately conclude where I go.*
- *I believe that social media algorithms are able to accurately conclude who I meet.*
- *I believe that social media algorithms are able to accurately conclude my political stance.*
- *I believe that social media algorithms are able to accurately conclude my religious beliefs.*
- *I believe that the social media platform is able to know where I stand on important issues such as my acceptance of the Covid-19 vaccination.*
- *I believe that the social media platform is able to know where I stand on important issues such as climate change.*
- *I believe that the social media platform is able to know where I stand on important issues such as immigration.*
- *The social media platform is able to distinguish between who I am in private and who I am in social contexts on the social media platform.*
- *The social media platform is able to distinguish between who I am in private and who I am in social contexts when I am not online.*

Accurate judgments (temporal)

- *The social media platform is able to keep my user data profile up to date with my interests, behaviors, and beliefs as they change over time.*
- *My user data profile from a month ago includes conclusions about me that are still accurate today.*
- *My user data profile from a year ago includes conclusions about me that are still accurate today.*
- *After having been an active user on the social media platform for several years, the platform can conclude whether I have changed as a person since I started using the platform.*
- *My entire user data profile tells an accurate story of the life that I have lived since I started using the platform.*

The normativity of an accurate UDP

- *The social media platform should take precautions to make sure that my user data profile is accurate.*
- *The social media platform should double-check my user data profile for accuracy, even if it takes them time or possibly other resources (e.g., money or additional employees) to do so.*
- *The social media platform should collect as much of my data as possible to ensure my user data profile is as correct as possible.*
- *The social media platform should collect my data to generate my user data profile.*

Transparency of data collection & use

- *The collection of my data should be disclosed to me clearly and transparently.*
- *The social media platform should disclose the way they collect and use my data.*
- *I want to know what data the social media platform has used to show advertisements to me.*

Transparency of SMP conclusions about me

- *I want to know what the social media platform has concluded about me.*
- *I am interested in understanding all the conclusions the social media platform has made about me.*
- *It is valuable to me to understand all the conclusions the social media platform has made about me.*
- *I do not care about the conclusions that the social media platform makes about me.*

Transparency of UDP

- *It is important to me that I am aware and knowledgeable about how my personal data will be used for my user data profile.*
- *The social media platform should allow me to access my user data profile.*

Changing my UDP

- *The social media platform should allow me to correct errors in my user data profile.*
- *I am motivated to change conclusions that I think are wrong in my user data profile.*
- *I am upset if the social media platform concludes something about me that I think is wrong.*
- *If my user data profile was made transparent to me, then correcting and maintaining my user data profile would be too tedious for me.*

If I could view my UDP

- *If I had the ability to view my user data profile, I would compare elements of the user data profile to the person that I think I am.*
- *If I had the ability to view my interests (i.e., movies, music, or books that I like) in my user data profile, I would compare them to my own real interests.*
- *If I had the ability to view what the social media platform claims I have bought in the past, I would compare it to what I have actually bought.*
- *If I had the ability to view where the social media platform claims I went in the past week, I would compare it to where I really went last week.*
- *If I had the ability to view who the social media platform claims I have met in the past week, I would compare it to who I met in the past week.*
- *If I had the ability to view my political stance in my user data profile, I would compare it to my own real political stance.*
- *If I had the ability to view my religious beliefs in my user data profile, I would compare them to my own real religious beliefs.*

If I was shown the content of my UDP

- *If I was shown the content of my user data profile, it would cause me to reevaluate who I am.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my interests (i.e., movies, music, or books that I like).*
- *If I was shown the content of my user data profile, it would cause me to reevaluate what I will buy in the future.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my political stance.*
- *If I was shown the content of my user data profile, it would cause me to reevaluate my religious beliefs.*

Influence of self-concept (unaware elements)

- *If I could view the content of my user data profile, then the conclusions the social media platform has made about me would have meaning to me.*
- *If my user data profile contains conclusions about who I am that I did not know about, then these conclusions don't have meaning to me.*
- *If my user data profile contains conclusions about my political stance that I did not know about, then these conclusions don't have meaning to me.*
- *If my user data profile contains conclusions about my religious beliefs that I did not know about, then these conclusions don't have meaning to me.*
- *If I was looking for a new interest, I would look into my user data profile for a suggestion.*
- *If my user data profile contains conclusions about my interests (e.g., movies, music, or books that I like) that I do not know about, then these conclusions will likely become new interests of mine.*
- *If my user data profile contains conclusions about what I will likely buy in the future, that I didn't know about, then these conclusions will likely influence what I buy in the future.*

A.4 Demographics

54.3% of participants were female, 43.8% male, and 1.9% defined themselves as other. 69% of participants were between 18 and 35 years old. 56.8% of participants had some form of university education, 33.4% had at least a high school diploma. 50.8% of participants were employees, 18.2% were students. Finally, 92.9% of participants listed their current country of residence as the United States.

B APPENDIX FIGURE 7

Figure 7 shows respondents' beliefs on the influence of the UDP on their self-concept. In particular, we wanted to understand whether participants would attribute meaning to identity declarations in their user data profile (UDP) that they were not aware of. Figure 7 is shown on the following page.

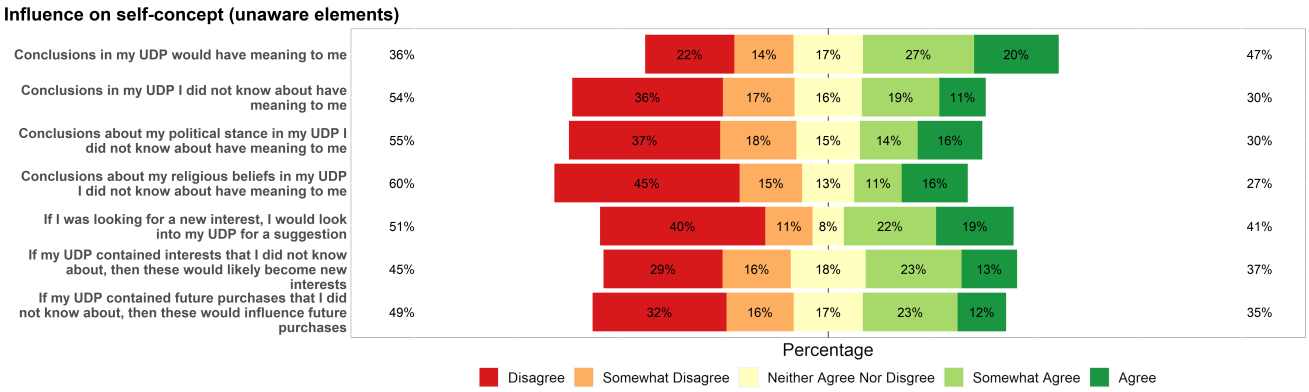


Figure 7: Respondents' ratings were largely divided over the question whether UDP conclusions would be meaningful to them and whether unknown identity declarations would carry meaning for them.

Setting the Stage: Towards Principles for Reasonable Image Inferences

Severin Engelmann
engelmas@in.tum.de
Chair of Cyber Trust
Department of Informatics
Technical University of Munich

Jens Grossklags
jens.grossklags@in.tum.de
Chair of Cyber Trust
Department of Informatics
Technical University of Munich

ABSTRACT

User modeling has become an indispensable feature of a plethora of different digital services such as search engines, social media or e-commerce. Indeed, decision procedures of online algorithmic systems apply various methods including machine learning (ML) to generate virtual models of billions of human beings based on large amounts of personal and other data. Recently, there has been a call for a "Right to Reasonable Inferences" for Europe's General Data Protection Regulation (GDPR). Here, we explore a conceptualization of reasonable inference in the context of image analytics that refers to the notion of evidence in theoretical reasoning. The main goal of this paper is to start defining principles for reasonable image inferences, in particular, portraits of individuals. Based on an image analytics case study, we use the notions of first- and second-order inferences to determine the reasonableness of predicted concepts. Finally, we highlight three key challenges for the future of this research space: first, we argue for the potential value of hidden quasi-semantics. Second, we indicate that automatic inferences can create a fundamental trade-off between privacy preservation and "model fit" and, third, we end with the question whether human reasoning can serve as a normative benchmark for reasonable automatic inferences.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Personalization*; *Clustering and classification*; • **Security and privacy** → *Social aspects of security and privacy*.

KEYWORDS

Image data, Reasonable inferences, Machine learning

ACM Reference Format:
Severin Engelmann and Jens Grossklags. 2019. Setting the Stage: Towards Principles for Reasonable Image Inferences. In *27th Conference on User Modeling, Adaptation and Personalization Adjunct (UMAP'19 Adjunct), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3314183.3323846

UMAP'19 Adjunct, June 9–12, 2019, Larnaca, Cyprus

ACM ISBN 978-1-4503-6711-0/19/06…$15.00
https://doi.org/10.1145/3314183.3323846

1 INTRODUCTION

Recently, user modeling techniques have been used to infer aesthetic (e.g., beauty), mental (e.g., beliefs, intentions), emotional (e.g., happiness, depression), and social (e.g., group affiliation) features about individuals based on their personal data as well as their digital footprints. The possibilities of user modeling techniques go far beyond the mere classification of individuals into types of customers: they create virtual models of individuals at an industrial scale based on personal and other data. This data is commonly associated with implicit mental characteristics and social situational factors often unknown to the corresponding individual. Thereby, many big data companies produce billions of virtual models of people to connect a particular informational resource (e.g., an advertising material) to the individual with the most "appropriate" model.

This signifies what we refer to as a hermeneutic shift: parts of the interpretative potential of the person is realized not by the person itself but by the "quasi-semantic power"[1] of textual extraction, image understanding, emotion and speech analysis, location analysis or even inaction interpretation (among others) [4, 25, 34, 49, 50]. Assigning quasi-semantic values to implicit identity claims stands in contrast to The Enlightenment's core idea that humans have the ability to freely and autonomously assign meaning to what they have experienced. From this perspective, user modeling techniques can create tensions with the autonomy of individuals to form a hermeneutic self-concept.

Moreover, the quasi-semantic power of user modeling techniques can lead to consequential discriminatory biases, for example, when credit decisions are based on the collection and analysis of digital footprints unknown to the corresponding individual. The opacity of user modeling processes makes it generally difficult to detect, understand and correct such biases.

Recently, there has been a call for a "Right to Reasonable Inferences" to set legally-binding standards with the purpose to protect individuals against inferences that are privacy-invasive, reputation-damaging, and difficult to verify [45]. Yet, the decisive question is what *reasonable* ought to mean in the context of an automatic inference about a person based on some published media content.

Here, we wish to set the stage for a productive discussion between the computer and social sciences in determining standards for reasonable inferences in image analytics.[2] Based on an image analytics case study using the Clarifai concept prediction prototype[3], we show that inferences about human portraits can be unreasonable when they predict concepts with underlying beliefs that cannot be

[1] Since humans are the only semantic engines in nature, see, for example, [11].
[2] Specifically, images that depict human beings.
[3] Available at: https://www.clarifai.com/demo.

revised in light of further evidence of the same type. Our claims are based on an empiricist view of reasonableness[4] that considers a knowledge-object's quality of evidence for a particular inference to qualify as reasonable or unreasonable.

We proceed as follows. In Section 2, we discuss why image analytics result in epistemic *and* ethical challenges and review related work in Section 2.1. In Section 3, we introduce an empiricist conceptualization of reasonableness that demands that what one is justified in believing is determined exclusively by evidence. We then upload two portraits to the Clarifai web interface image prediction prototype and analyze the reasonableness of the concepts the engine returns (see Section 4). Finally, in Section 5, we consider the potential autonomy-enabling value of hidden quasi-semantics and discuss a fundamental trade-off between privacy and model fit.

2 BACKGROUND

Social media users engage in both explicit[5] and implicit identity claims. Generally, images are among the most prevalent forms of self-presentation techniques on social media. Given their inherent semantic ambiguity, images are considered implicit identity claims. Implicit identity claims are "given off" in various indirect manners. Typical examples of implicit identity claims are showing one's affiliation to certain individuals, social or institutional groups, or expressing preferences and interests in an indirect manner [7, 48]. Indeed, there is evidence that "showing rather than telling" has become the most common self-presentation strategy on social media platforms [21, 43].

Consequently, marketers value images more than other media content. According to Socialbakers, images posted on Instagram[6] create four times more user engagement than other user content on Facebook[7]. Another reason is that image understanding further closes the gap between organic and commercial media content since objects in an image can be classified as products. Overall, there have been significant efforts made in the advancement of image-understanding technologies to model users based on pictorial identity claims in both academia and industry.[8]

When modeling an individual, image-understanding technologies do not simply draw semantics from the content of images but assign, add, and possibly produce their meaning in the first place. Despite their quasiness, user modeling techniques model features of individuals that are likely inaccessible for the individual herself. Thereby, user modeling techniques presumably attempt to transfer what is radically subjective (and therefore difficult if not impossible to falsify) into the realm of objective evaluation. They, therefore, try to explain something that is essentially first-person in third-person terms.

The majority of contemporary philosophical theories on personal identity support the idea that being free in interpreting one's self is a constitutive element of the conceptual boundaries of personal identity [12, 26, 31, 39]. Importantly, a moral status comprising moral rights and duties presupposes autonomy over one's self-concept. In other words, it is *because* individuals can evaluate what they are, shape whatever they wish to be on this basis, that they can be made responsible for what they become [40]. Moral accountability would, therefore, be impossible if individuals did not have the freedom and autonomy to form and negotiate such a hermeneutic self-concept.

Furthermore, empirical studies in psychology have demonstrated that individuals have the ability to attribute meaning to their experiences as a processes of hermeneutic identity formation [24, 36, 37]. Studies by [38] show that individuals interact with other individuals strategically in order to verify their self-concept: self-concept negotiation denotes the verification attempt of a person's self-concept through the interaction with other individuals. Whether individuals perceive user modeling outcomes as a means of technologically-mediated self-verification or self-discontinuity remains to be studied. Yet, hiding a person's quasi-semantic self-concept, i.e. disallowing user modeling techniques to partake in a self-verification process, could have some benefits (see Section 5).

Taken together, an autonomous self-concept emerges when an individual carries out *the psychological work required to attribute meaning to certain experiences*. Image analytics signify a hermeneutic shift because they transform implicit identity claims into explicit declarations of identity. Image analytics are not solely epistemic tools but quasi-semantic engines that potentially interfere with a person's autonomy to freely form a self-concept.

2.1 Related Work

With the rise of search engines in the early 2000s, automatizing the attribution of semantics to images returned high accuracy on object identification [23]. In the context of search tasks, object identification proved to be an efficient strategy.[9] In social media's people-based marketing mere object identification does not suffice for advertisement delivery based on implicit identity claims. Today, learning from content and structure of social network sites as well as correlating aspects about natural persons and groups to online content is a fast-growing research field. In the following, we briefly discuss main trends as they pertain to image data analyses.

Popularity prediction of image data: Several projects focus on determining the likelihood that certain image postings will achieve high view counts and high positive approval. Using a variety of machine learning approaches the context of a user and posting is taken into consideration to predict the future attention given to a newly posted image (e.g., [15, 27, 46, 47]).

Self-presentation: Various papers explore how (and under what circumstances) individuals strategically manage their social network accounts to aim for more favorable reception by the intended audience (e.g., [32, 41]). In the context of image data, for example, researchers have begun exploring users' management of multiple accounts on Instagram to present themselves to different audiences

[4] The terms "reasonableness" and "rationality" are considered synonymous in this work.

[5] For example, when individuals communicate specific self-relevant information in written form, they usually engage in explicit identity claims: "I am 20 years of age and I like reading biographies of great scientists".

[6] Advertising campaigns on Instagram are run via the Facebook advertising platform including the choice of custom audiences and lookalike audiences: see https://business.instagram.com/advertising/.

[7] https://www.socialbakers.com/blog/instagram-engagement

[8] For example, Amazon: https://aws.amazon.com/de/rekognition/, Microsoft: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/, Facebook: https://code.fb.com/ai-research/fair-fifth-anniversary/, Google: https://cloud.google.com/vision/.

[9] Object inferences can be semantically ambiguous. For example, while distinct colors and shapes can be mapped to mathematical vectors with relative ease, the same is more difficult with objects containing continuous features [44].

in strategically altered ways. On a "Rinsta" (Real Instagram) account, a curated self is presented to a wider audience; whereas on a "Finsta" (Fake Instagram) account, less perfect material is presented to a hand-selected group of individuals for feedback and banter [21]. Interestingly, research has shown that users perceive their carefully styled images on the Finsta accounts to capture their real self more accurately in comparison to their Rinsta accounts with presumably more "genuine" material [21].

Inferring personality traits and user characteristics from image data: Partly triggered by the Gaydar research study [19] in 2009, significant attention has been given by the research community to finding associations between aspects of user profiles, user relationships, and posts, on the one hand, and traits/characteristics of the user or groups of users, on the other hand. In the context of image data, recent research suggests a relationship between personality traits and style aspects of posted pictures (e.g., hue, brightness and saturation); likewise, the content of pictures can be associated with personality characteristics [8–10].

Previous work also aims to find image characteristics that match specific user groups [17]. Likewise, analyses focus on automatically detecting gender and age from posted image content [16, 33].

Behavioral research has also explored how different personality characteristics (e.g., narcissistic tendencies [20]) impact the perception of image data.

Relationship of mental health and image data: Numerous research projects have focused on uncovering correlations between the usage of social network sites and mental health aspects such as addiction, anxiety, depression or body image (see, for example, a recent review [13]). Similar work can be found that is focused on image data. For example, perusal of attractive pictures of celebrities and peers has been found to be associated with a more negative body image by women [3, 18]. Likewise, uploaded image data can also be revealing of mental health indicators such as related to depression [30].

While there is a plethora of technical research and behavioral studies to understand social network site usage and its impact on users, also in the context of image data, we are unaware of any work that explores principles to develop reasonable standards for image inferences made by automated systems.

3 FIRST STEPS TOWARDS PRINCIPLES FOR REASONABLE IMAGE INFERENCES

3.1 An empiricist view of reasonable inferences

Fundamentally, there are two types of reasoning: practical and theoretical reasoning also sometimes referred to as instrumental and epistemic reasoning, respectively (see for example [35]). Practical reasoning is concerned with the question "What to do?". Theoretical reasoning asks "What to believe?". Practical and theoretical reasoning are not mutually exclusive. When choosing a reasonable action for a desirable outcome an individual relies on a theoretically reasonable belief. Thus, practical or instrumental reasoning usually follows theoretical reasoning.

In this work, we assume an empiricist view that considers a knowledge-object's quality of evidence to decide whether a particular inference qualifies as reasonable or unreasonable. The empiricist view of a reasonable inference considers whether the belief about a proposition is *proportional* to the evidence available. Generally, the empiricist view on being reasonable in the theoretical sense considers the "goodness" or "fitness" of reasons provided that favors the truth of a proposition. While this conceptualization of reasonableness perhaps seems simple or even trivial, empirical research has demonstrated that individuals exhibit many information-processing biases pursuant to this empiricist account of reasonableness [2, 42].[10]

The goal of this work is to start developing principles for *portrait* image inferences that are eligible to be called reasonable. To do this, we need an example output from an image analytics engine. Here, we use the Clarifai web interface image prediction demo, which is based on deep convolutional neural networks (CNNs). We upload two portraits (see Figure 1 and Figure 2) to this image prediction demo and analyze the reasonableness of the concepts the engine returns. Corresponding to the literature reviewed in Section 2.1, we view a single image as a stand-alone knowledge-object whereby a predicted concept (i.e., the predicted outcome) is based only on the content of that single image.

3.2 Case study: Reasonableness and correctness of predicted concepts for two portraits

Reasonable and correct inferences

Consider the two images in Figure 1 and Figure 2. Is the content of these two images eligible to serve as evidence for the inferences made (see "predicted concepts" top right corner on both images)?

Figure 1 displays the face of a woman. The first three predicted concepts "woman", "portrait", and "facial expression" cannot be argued against, just like the first five predicted concepts in Figure 2. Here, the given beliefs about these propositions are *proportional* to the evidence available and therefore these inferences can be said to be reasonable. All of these features can be reasonably inferred from the evidence given. Note that we do not evaluate the potential discriminatory or unfair *consequences* of specific labels, rather we are first and foremost interested in their epistemic justification. For example, returning the label "gender" may lead to consequential discrimination independent from whether it is a (epistemically) reasonable inference. Additionally, considering our two portraits, the features "woman", "portrait" and "facial expression" (Figure 1) and "portrait", "eye", "face", "guy", "man" (Figure 2) have been classified correctly.[11] Overall, these inferences are – to a large enough degree – reasonable and correct.

Reasonable inferences with incorrect predictions

Other predicted concepts can in principle be reasonable but seem to have been classified incorrectly for the specific portraits given. In Figure 2, for example, the CNNs predict the concept "smile", which is incorrect since the person depicted does not seem to smile. Note that this would not be an unreasonable inference since a face can potentially bear a smile. Rather, the accuracy of the training set's classification (i.e., the ground truth) is insufficient in returning an otherwise reasonable inference correctly. In this specific case, the

[10] For example, category mistakes, anchoring, representative bias, ignoring the context, framing effects etc.

[11] For Figure 2, the predicted concepts "hair", "model", "skin" seem to be reasonable and correct as well.

(a) Female portrait

PREDICTED CONCEPT	PROBABILITY
woman	0.980
portrait	0.965
facial expression	0.964
fashion	0.930
pretty	0.930
multicultural	0.925
one	0.911
wear	0.908
promotion	0.899
contemporary	0.858
indoors	0.856
friendly	0.847
arrival	0.846
people	0.842
elegant	0.822
intelligence	0.786
dentition	0.742
casual	0.728
business	0.720
charming	0.714

(b) Predicted concepts

Figure 1: Concept results using the Clarifai image prediction demo for a female portrait. The engine returns predictions on gender "woman", ethnicity-related features "multicultural", cognitive skills "intelligence", and presumably aesthetic features "pretty", "elegant", "friendly", "charming" (among others). For copyright purposes, we artistically rendered the original picture. Original picture ©https://thispersondoesnotexist.com/.

prediction seems to be incorrect but only in relation to an otherwise reasonable assumption made when annotating the training set.

Unreasonable inferences due to non-falsifiability

There seem to be inferences that are unreasonable due to their non-falsifiability. For example, both images contain predicted concepts of aesthetic evaluations or judgments. For a judgment to be an aesthetic judgment it necessarily needs to be subjective, making it the exact opposite of an empirical judgment. More generally, judgments on beauty and ugliness are commonly taken to be core examples of aesthetic judgments. In Figure 1, an example of an aesthetic judgment is "pretty" and in Figure 2 "fine-looking". Other, perhaps more indirect, aesthetic evaluations seem to be "elegant", "friendly", and "charming" (Figure 1) as well as "serious" (Figure 2). Overall, such aesthetic judgments of taste are unreasonable since they cannot be falsified by additional evidence of the same type. For such inferences, additional image evidence cannot *in principle* verify or falsify, in other words, change the proposition.[12]

Similarly to aesthetic inferences, another class of inferences are unreasonable due to their non-falsifiability. These inferences contain category mistakes because they take a physical or anatomical property to be evidence for a mental feature. In Figure 1, the facial proportions of the woman are taken to be evidence for her "intelligence" while the face in Figure 2 is taken to be evidence for the person to be "crazy". Portraits seem to be inadequate evidence for a person's mental capabilities or, generally, their mental characteristics. This inference cannot be made more reasonable by providing more portraits of the two people shown in Figure 1 and Figure 2. In other words, the proposition that the person in Figure 2 is actually crazy does not become more likely the more pictures of that person are analyzed. Again, the prediction for such labels can be correct

[12]There are, however, reasonable physical or anatomical inferences, for example, "freckle" in Figure 2.

(a) Male portrait

PREDICTED CONCEPT	PROBABILITY
portrait	0.993
eye	0.986
face	0.974
guy	0.974
man	0.971
fine-looking	0.968
young	0.961
hair	0.938
boy	0.933
people	0.932
blood	0.930
dark	0.916
freckle	0.909
serious	0.909
model	0.909
crazy	0.909
fashion	0.895
funny	0.893
smile	0.890
skin	0.888

(b) Predicted concepts

Figure 2: Concept results using the Clarifai image prediction demo for a male portrait. The engine returns predictions on gender "man", age "young"/"boy", mental "crazy"/"funny", and presumably aesthetic features "fine-looking", "serious" (among others). For copyright purposes, we artistically rendered the original picture. Original picture ©Bruce Gilden.

but only in relation to the unreasonable assumptions made when annotating the training set.

4 ANALYSIS OF THE CASE STUDY

There is an epistemic difference between descriptively identifying the objects "basketball" and "person" and conclusively inferring "Interest person x = basketball", merely because these objects have been identified. In a similar vein, there is a difference between measuring the physical property "wide space between eyes"and the object "glasses" and inferring some measure of intelligence based on these features. In our case study, we generally judged inferences that could be "directly" read off the portrait as reasonable. Such first-order inferences, as one might want to call them, seem epistemically valid and are henceforth difficult to object morally. They are reasonable independent of the predictive strength of the model.

Unreasonable inferences, on the other hand, seem to be predominantly constructed inferences. In our case study, they included claims about the person that could not be observed or accessed through the evidence given. Such second-order inferences presuppose a selection (and naturally a disregard) of specific first-order inferences that – combined – produce a new proposition. Second-order inferences must not necessarily be unreasonable. Consider, for example, the predicted concept "indoors" for the portrait in Figure 1. Predicting whether a depicted scenery is indoors or outdoors is a second-order inference because a single object is unlikely to produce a definite conclusion. The difference is that this second-order inference is responsive to additional evidence of the same

type resulting in belief revision. Thereby, an inference is unreasonable in the case that novel or additional evidence becomes available that defeats the previous justification to believe in a proposition. In case of better evidence one ought to change the previously held belief in light of this new evidence. For example, another image of this scenery could in principle provide what Pollock refers to as "rebutting evidence" [29]. The new image is the same type or source of evidence. But because it is a reasonable second-order inference it is responsive to belief revision, which in this case is equivalent to the principles of Bayesian inference.

This claim does not hold for unreasonable second-order inferences. Bayesian inference (or belief revision) cannot convert an unreasonable second-order inference into a reasonable inference (e.g., predicted concept "intelligence" in Figure 1). Such category mistakes can only be reverted by changing the underlying assumption or by gathering different types of evidence but not by considering more evidence with the same category mistake.

5 DISCUSSION & CONCLUDING REMARKS

In this discussion paper, we applied an empiricist account of reasoning to determine the reasonableness of predicted concepts in the context of an image analytics case study. This is only one of many possible accounts of reasoning each of which comes with specific trade-offs. Arguably, an empiricist account is autonomy-preserving but limited to first-order inferences about individuals. Regardless of the account of reasonableness, an inference may be reasonable and correct but still be rejected by the individual. Here, one could argue that an inference becomes reasonable only when the data subject agrees with its proposition.

The recent call for a "Right to Reasonable Inferences" proposes a "Right to know about Inferences" and a "Right to rectify Inferences" (among others) [45]. However, hiding the quasi-semantic power of user modeling techniques does have its benefits. By revealing the logic involved in making hermeneutic inferences, the system directly recommends these hermeneutics to the user. It remains to be explored how individuals would perceive information on inferences as given in our two image examples. Revealing at least in part the manner and content of user modeling processes and outcomes enables internalization and conformation to the proposed inferences. Perhaps individuals would welcome such a degree of transparency as a mechanism to "offload" the psychological work necessary to attribute meaning to certain life events. Revealing such inferences to the individual means recognizing their quasi-semantic power in shaping who we are and who we can become – we accept that they have their own narrative capacity. Thus, transparency of user modeling inferences could even exacerbate the polarization effect observed in social media personalization.

Another key challenge is privacy. Image inferences tend to become more reasonable the more personal data is collected and analyzed. This creates a privacy trade-off. The trade-off consists in the observation that a representative model of an individual is possible only at the expense of privacy. For example, ML classifiers must be able to respond to concept drift without "neglecting" the outdated data when learning a model of personal identity [51]. For example, sliding windows of fixed and variable sizes of training data are used to build an updated model [14]. Since both fixed and variable windows are definite in their size, some old data will necessarily be forgotten. What criteria determine which data are to be forgotten and which ones are to be considered in creating an updated representative model of a person? Model fit requires a potentially uninterrupted flow of data possibly resulting in significant privacy challenges [5].

Finally, a key question is whether we should take human reasoning as a benchmark for reasonable automatic inferences. In the empirical literature on human reasoning *..."the ordinary person is claimed to be prone to serious and systematic error in deductive reasoning, in judging probabilities, in correcting his biases, and in many other activities"* [6]. For example, humans make judgments about cognitive capabilities based on physical properties [1, 28]. Following our image analytics case study, we conclude that inferences about individuals' cognitive and mental features are unreasonable since an image does not provide the kind of evidence needed to justify such claims. This also counts for inferences made about individuals' intentions or goals based on image evidence (see [22]).

Overall, it will remain a pressing ethical challenge to define normative standards of reasonableness that automatic image inferences should comply with.

Acknowledgments: We thank the reviewers for their insightful comments that helped to improve our work. The paper is based on research conducted as part of a Volkswagen Foundation planning grant project.

REFERENCES

[1] Michael Argyle and Robert McHenry. Do spectacles really affect judgements of intelligence? *British Journal of Social and Clinical Psychology*, 10(1):27–29, 1971.
[2] Dan Ariely, George Loewenstein, and Drazen Prelec. "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1):73–106, 2003.
[3] Zoe Brown and Marika Tiggemann. Attractive celebrity and peer images on Instagram: Effect on women's mood and body image. *Body Image*, 19:37–43, 2016.
[4] Buru Chang, Yonggyu Park, Donghyeon Park, Seongsoon Kim, and Jaewoo Kang. Content-aware hierarchical point-of-interest embedding model for successive POI recommendation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3301–3307, 2018.
[5] Michela Chessa, Jens Grossklags, and Patrick Loiseau. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 90–104, 2015.
[6] Jonathan Cohen. Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4(3):317–331, 1981.
[7] Nicole Ellison, Charles Steinfield, and Cliff Lampe. Connection strategies: Social capital implications of Facebook-enabled communication practices. *New Media & Society*, 13(6):873–892, 2011.
[8] Bruce Ferwerda and Marko Tkalcic. Predicting users' personality from Instagram pictures: Using visual and/or content features? In *Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 157–161. ACM, 2018.
[9] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Predicting personality traits with Instagram pictures. In *Workshop on Emotions and Personality in Personalized Systems*, pages 7–10. ACM, 2015.
[10] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Using Instagram picture features to predict users' personality. In *International Conference on Multimedia Modeling (MMM)*, pages 850–861, 2016.
[11] Luciano Floridi. Web 2.0 vs. the semantic web: A philosophical assessment. *Episteme*, 6(1):25–37, 2009.
[12] Harry Frankfurt. Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1):5–20, 1971.
[13] Rachel Frost and Debra Rickwood. A systematic review of the mental health outcomes associated with Facebook use. *Computers in Human Behavior*, 76:576–600, 2017.
[14] João Gama, Raquel Sebastião, and Pedro Pereira Rodrigues. On evaluating stream learning algorithms. *Machine Learning*, 90(3):317–346, 2013.
[15] Steve Göring, Konstantin Brand, and Alexander Raake. Extended features using machine learning techniques for photo liking prediction. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.

[16] Kyungsik Han, Sanghack Lee, Jin Yea Jang, Yong Jung, and Dongwon Lee. Teens are from mars, adults are from venus: Analyzing and predicting age groups with behavioral characteristics in Instagram. In *Conference on Web Science*, pages 35–44. ACM, 2016.

[17] Kyungsik Han, Yonggeol Jo, Youngseung Jeon, Bogoan Kim, Junho Song, and Sang-Wook Kim. Photos don't have me, but how do you know me? Analyzing and predicting users on Instagram. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, pages 251–256. ACM, 2018.

[18] Joshua Hendrickse, Laura Arpan, Russell Clayton, and Jessica Ridgway. Instagram and college women's body image: Investigating the roles of appearance-related comparisons and intrasexual competition. *Computers in Human Behavior*, 74:92–100, 2017.

[19] Carter Jernigan and Behram Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[20] Seunga Venus Jin and Aziz Muqaddam. "Narcissism 2.0! Would narcissists follow fellow narcissists on Instagram?" The mediating effects of narcissists personality similarity and envy, and the moderating effects of popularity. *Computers in Human Behavior*, 81:31–41, 2018.

[21] Jin Kang and Lewen Wei. Let me be at my funniest: Instagram users' motivations for using Finsta (aka, fake Instagram). *The Social Science Journal*, 2019.

[22] Owen King. Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems. In Don Berkich and Matteo Vincenzo d'Alfonso, editors, *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, pages 265–282. Springer, 2019.

[23] Victor Lavrenko, Raghavan Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems*, pages 553–560, 2004.

[24] Mark Leary and June Price Tangney. *Handbook of Self and Identity*. Guilford Press, 2011.

[25] Weizhi Ma, Min Zhang, Chenyang Wang, Cheng Luo, Yiqun Liu, and Shaoping Ma. Your tweets reveal what you like: Introducing cross-media content information into multi-domain recommendation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3484–3490, 2018.

[26] Alasdair MacIntyre. *After Virtue: A Study in Moral Theology*. University of Notre Dame Press, 1981.

[27] Eric Massip, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Kai-Lung Hua. Exploiting category-specific information for image popularity prediction in social media. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 45–46, 2018.

[28] Fhionna Moore, Dimitra Filippou, and David Ian Perrett. Intelligence and attractiveness in the face: Beyond the attractiveness halo effect. *Journal of Evolutionary Psychology*, 9(3):205–217, 2011.

[29] John Pollock and Joseph Cruz. *Contemporary Theories of Knowledge*. Rowman & Littlefield, 1999.

[30] Andrew Reece and Christopher Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):15, 2017.

[31] Marya Schechtman. The narrative self. In Shaun Gallagher, editor, *The Oxford Handbook of the Self*. Oxford University Press, 2011.

[32] Gwendolyn Seidman. Expressing the "true self" on Facebook. *Computers in Human Behavior*, 31:367–372, 2014.

[33] Junho Song, Kyungsik Han, Dongwon Lee, and Sang-Wook Kim. "Is a picture really worth a thousand words?": A case study on classifying user attributes on Instagram. *PloS One*, 13(10):e0204938, 2018.

[34] Dusan Sovilj, Scott Sanner, Harold Soh, and Hanze Li. Collaborative filtering with behavioral models. In *Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 91–99, 2018.

[35] Keith Stanovich. *Decision Making and Rationality in the Modern World (Fundamentals in Cognition)*. Oxford University Press, 2009.

[36] William Swann, Alan Stein-Seroussi, and Brian Giesler. Why people self-verify. *Journal of Personality and Social Psychology*, 62(3):392–401, 1992.

[37] William Swann, Peter Rentfrow, and Jennifer Guinn. Self-verification: The search for coherence. In Mark Leary and June Price Tangney, editors, *Handbook of Self and Identity*, pages 367–383. 2003.

[38] William Swann. Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, 53(6):1038, 1987.

[39] Charles Taylor. Responsibility for self. In Amélie Rorty, editor, *The Identities of Persons*. University of California Press, 1976.

[40] Charles Taylor. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, 1989.

[41] Leman Pınar Tosun. Motives for Facebook use and expressing "true self" on the Internet. *Computers in Human Behavior*, 28(4):1510–1517, 2012.

[42] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[43] José Van Dijck. 'You have one identity': Performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, 2013.

[44] Jan Van Gemert, Cor Veenman, Arnold Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[45] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: Rethinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019.

[46] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. Sequential prediction of social media popularity with deep temporal context networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 3062–3068, 2017.

[47] Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, and Jiebo Luo. How to become Instagram famous: Post popularity prediction with dual-attention. In *IEEE International Conference on Big Data (Big Data)*, pages 2383–2392, 2018.

[48] Shanyang Zhao, Sherri Grasmuck, and Jason Martin. Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior*, 24(5):1816–1836, 2008.

[49] Qian Zhao, Martijn Willemsen, Gediminas Adomavicius, Maxwell Harper, and Joseph Konstan. Interpreting user inaction in recommender systems. In *ACM Conference on Recommender Systems*, pages 40–48, 2018.

[50] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(1s):Article No. 14, 2019.

[51] Indrė Žliobaitė. Learning under concept drift: An overview. *arXiv preprint arXiv:1010.4784*, 2010.

What People Think AI Should Infer From Faces

Severin Engelmann*
severin.engelmann@tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

Chiara Ullstein*
chiara.ullstein@tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

Orestis Papakyriakopoulos
orestis@princeton.edu
Princeton University, Center for Information Technology Policy
Princeton, USA

Jens Grossklags
jens.grossklags@in.tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

ABSTRACT

Faces play an indispensable role in human social life. At present, computer vision artificial intelligence (AI) captures and interprets human faces for a variety of digital applications and services. The ambiguity of facial information has recently led to a debate among scholars in different fields about the types of inferences AI should make about people based on their facial looks. AI research often justifies facial AI inference-making by referring to how people form impressions in first-encounter scenarios. Critics raise concerns about bias and discrimination and warn that facial analysis AI resembles an automated version of physiognomy. What has been missing from this debate, however, is an understanding of how "non-experts" in AI ethically evaluate facial AI inference-making. In a two-scenario vignette study with 24 treatment groups, we show that non-experts (N = 3745) reject facial AI inferences such as trustworthiness and likability from portrait images in a low-stake advertising and a high-stake hiring context. In contrast, non-experts agree with facial AI inferences such as skin color or gender in the advertising but not the hiring decision context. For each AI inference, we ask non-experts to justify their evaluation in a written response. Analyzing 29,760 written justifications, we find that non-experts are either "evidentialists" or "pragmatists": they assess the ethical status of a facial AI inference based on whether they think faces warrant sufficient or insufficient evidence for an inference (evidentialist justification) or whether making the inference results in beneficial or detrimental outcomes (pragmatist justification). Non-experts' justifications underscore the normative complexity behind facial AI inference-making. AI inferences with insufficient evidence can be rationalized by considerations of relevance while irrelevant inferences can be justified by reference to sufficient evidence. We argue that participatory approaches contribute valuable insights for the development of ethical AI in an increasingly *visual* data culture.

*Denotes equal contribution.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533080

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; • **Social and professional topics** → **User characteristics**; *Computing / technology policy*; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

artificial intelligence, computer vision, human faces, participatory AI ethics

ACM Reference Format:
Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What People Think AI Should Infer From Faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533080

1 INTRODUCTION

Human faces and the information they convey are essential in human interaction. When seeing a person for the first time, humans rapidly and automatically make a variety of judgments, such as whether a person looks trustworthy or likable [75, 76, 78, 99]. People's faces can play a significant role in some of society's most important decision-making scenarios: first facial impressions can determine hiring choices [76, 84], election outcomes [6, 59, 77], or jail sentences [26, 105, 109]. Yet, we are often told not to judge a book by its cover, an imperative that it is morally wrong to form beliefs about a person based on insufficient evidence. Indeed, inferring inner character traits based on looks had been foundational for once lauded physiognomic and phrenological practices in organizations and institutions [22, 35, 83, 92, 93].

Today, research in psychology and evolutionary anthropology shows that first facial impressions have an "irresistible" force, but are nonetheless largely inaccurate [13, 27, 78, 99, 100]. This line of research provides ample evidence that there is no relationship between how we look and how trustworthy or intelligent we actually are. Surprisingly, another body of research studies continues to suggest that first facial impressions are accurate or, at least, not completely invalid [45, 51, 56, 62, 72, 80]. Commonly recognizing this latter body of literature, computer vision artificial intelligence (AI) – the computerization of visual perception – has recently developed datasets, algorithms, and models to automate social perception tasks in fields such as affective computing (e.g., [19]) and social

robotics [15, 97]. Using computer vision AI, studies have claimed to successfully infer emotion expression and intensity [10, 25], sexual [60, 103] and political orientation [54, 107], as well as a variety of latent traits in personality assessments based on people's faces in images [4, 16, 30–32, 81, 88, 89, 108]. AI research has established tools for feature extractions from faces (e.g., Face++ [1], EmoVu[2]) as well as for open training datasets (ImageNet[3], First Impression V2[4], PsychoFlickr dataset[5]) and models [1, 11] for facial analysis AI.

Computer vision AI drives software that helps "make sense" of user images on social media for advertising purposes, video interviews in hiring software, or mood detection in car systems. The AI emotion recognition industry alone is said to be worth US37$ billion by 2026 [20]. AI systems play an increasingly important role in the semantic interpretation of our world, and because faces have an indispensable social signaling function, they are taken to be particularly revealing of who we are. But how should AI interpret people's faces? All imagery is semantically ambiguous and computer vision AI inference-making necessarily follows from the semantic annotation of visual data by humans, in most cases, by crowd-sourced platform workers [67, 74, 95]. This complicated ethical question has led to debates between policymakers, researchers in computational and social sciences, and companies that develop or use such AI. A number of research papers, including from the FAccT research community, have pointed out ethical challenges with regard to computer vision AI inferences [21, 22, 29, 34, 35, 68, 69, 82, 83, 87, 92–94]. However, we believe that such an effort must at least be cognizant of how "ordinary" people, i.e., non-experts in AI, evaluate the normativity of computer vision inferences.

In this work, we follow calls for more empirically-informed AI ethics [55, 85] and investigate what non-experts (N = 3745) think AI should and should not infer from portrait images – images that only show a person's face. Using a two-scenario vignette study with 24 treatment groups, we show that non-experts find AI latent trait inferences (e.g., intelligence) morally impermissible regardless of the decision context for which the inference is used for (advertising & hiring). A majority of subjects evaluates inferences such as gender, skin color, and emotion expression as morally permissible in the low-stake decision context (advertising) but impermissible in the high-stake decision context (hiring). None of our framing effects influenced subjects' evaluations indicating a strong value disposition toward AI facial analysis. We use the transformer-based model RoBERTa [63] to analyze subjects' 29,760 written justifications for each AI inference. We find that subjects raise ethical concerns about all AI inferences in both contexts. When justifying the normativity of an AI inference, subjects use one of two meta-principles: an AI facial inference is permissible when facial information warrants sufficient evidence or when making the inference results in beneficial outcomes. Our analysis illustrates the normative complexity behind facial AI inferences, and provides guidance for forthcoming technology policy debates.

[1] https://www.faceplusplus.com/
[2] https://www.programmableweb.com/api/emovu
[3] https://www.image-net.org/
[4] http://chalearnlap.cvc.uab.es/dataset/24/description/
[5] https://figshare.com/articles/dataset/zahra_plos_data_zip/6469577

2 RELATED WORK: THE IMPOSITION OF MEANING IN A VISUAL DATA CULTURE

2.1 Power dynamics between requesters and data annotators

Recently, several authors have raised ethical questions regarding the creation, management, and application of computer vision datasets. Computer vision companies (also known as "requesters") hire data processing companies, most often located in "less developed" countries, to perform efficient and cost-effective dataset creation, including data annotation. The emergence of a visual data culture – across Facebook's services alone, 2 billion images are shared every day[6] – together with the need for manual, human semantic labeling has led to the establishment of a data annotation industry[7] [67, 95]. Critical data science (broadly speaking) highlights challenges related to accountability and transparency gaps resulting from the near-unbounded power of computer vision AI companies and AI research institutes (i.e, requesters) to determine the interpretative potential of visual content [33, 68–70, 85, 87].

Studies find that requesters face little pressure to justify data labeling projects when hiring data processing companies for dataset labeling [67–69, 85]. In a field study on two data processing companies, Miceli et al. concluded that the work of image annotators is largely guided by the interests of the requester organization [68]. The authors report that this power dynamic does not allow image annotators to voice ethical concerns during the data labeling process. The hierarchical managerial structure at data processing companies restricts the possibility for the deliberative input by annotators [69]. In [68], the authors assert that "the one who is paying has the right to the imposition of meaning". To increase transparency and accountability of dataset creation, researchers have developed proposals to standardize documentation. For example, Gebru et al. suggest that each dataset should have a corresponding datasheet, explaining, among others, the purpose for which the dataset was created, the description of the images (or other data types), procedural aspects such as data cleaning and labeling, as well as the tasks and their unique contexts that the dataset is intended to be used for [33]. Holland et al. propose a "Dataset Nutrition Label" that specifies different modules, including the data origin, dataset variables, and ground truth correlations [41]. These and other standardized documentation practices [e.g., 70] can help AI developers to select more suitable datasets for their model development. However, such documentation practices are currently voluntary and rely entirely on the initiative and implementation of dataset creators.

2.2 Faces as sources of meaning and means for classification?

Authors have raised critical questions regarding a second key ethical challenge that is the subject of this work: What kind of inferences should a computer vision AI make about people based on visual data? Moreover, how do we justify what differentiates

[6] *Using Artificial Intelligence to Help Blind People 'See' Facebook*: https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/
[7] For a contribution by investigative journalists on the data annotation industry, see: *A.I. Is Learning From Humans. Many Humans.* https://www.nytimes.com/2019/08/16/technology/ai-humans.html

permissible from impermissible facial inferences when the context application changes? Given the inherently semantic ambiguity of visual data, fixing the large space of interpretive possibilities to a selection of target variables is an act of classification that inevitably demands an ethical justification [22, 40, 47, 82, 87, 93, 93]. This particularly applies to inferences about people based on their facial looks. Human faces are among the most frequently used "objects of interpretation" in computer vision AI. A recent review of nearly 500 prominent computer vision AI datasets found that 205 were "face-based": no other object was represented more often in computer vision datasets than human faces [85]. Social psychologists assert that humans are "obsessed" with faces and that they "cannot help but form impressions based on facial appearances" [99–101]. On first encounter, faces influence first impressions and shape whether we think someone *appears* trustworthy, intelligent, assertive, or attractive (among other traits) [76, 78, 101]. In many ancient cultures, and still today, there are persistent beliefs that faces are "a window to a person's true nature" [101], the idea that there is a reliable relationship between facial appearance and character[8]. The "irresistible influence" of faces can be consequential: first impressions can determine to whom we speak at a social gathering, whether we perceive a politician to be trustworthy, or whether we judge a job applicant as intelligent [100, 101, 110].

Recently, computer vision AI has purportedly inferred such first facial impressions for a variety of different contexts, for example in social media and for automatic hiring software [5, 11, 30–32, 39, 88, 89, 98, 108]. In the United States alone, millions of job applicants have participated in automatic hiring procedures that assess, among others, candidates' faces to produce an employability score [82, 94]. Sensitive categories such as gender and race are often treated as "commonsense categories" in computer vision datasets [22, 69, 82, 87]. However, a recent comparison between computer vision datasets presents findings that some racial categories show more variance than others across datasets despite nominally equivalent categorization [47]. Buolamwini and Gebru show that facial analysis AI produces the highest error rate for darker-skinned women and the lowest error rate for lighter-skinned males [14]. Critical perspectives warn that gender and skin color classification by facial analysis AI echoes colonial acts of "reading race onto the body" [86]. Facial analysis AI tends to rely on binary, cis-normative gender classifications [46, 86], thereby neglecting a trans-inclusive view of gender. Emotion recognition and sentiment analysis based on facial expressions have been the subject of multiple AI research projects and a plethora of digital companies – from large corporations to startups – use AI to infer facial emotion expression for social media, hiring, education, health, or security [93]. Other studies present facial analysis AI that is "better" at inferring sexual and political orientation from facial features than people [54, 103]. Others have organized yearly "first impression challenges" – competitions to create benchmark vision models for automatic first impression inferences in job candidate screening[9]. Computer vision AI studies often embrace research studies that underscore the apparent validity of first impressions or that, at least, assert that the invalidity of first impressions is inconclusive [45, 51, 56, 62, 72, 80, 102]. However, there is strong evidence that first facial impressions do not go beyond a "kernel of truth" [13, 78, 79, 99–101].

The conviction that facial configurations are indicative of a person's character inevitably rests on the pseudoscientific ideas of physiognomy and phrenology. Once celebrated scientific theories, prominent figures in the field of physiognomy such as Caspar Lavatar, Ceseare Lombroso, and Francis Galton developed entire taxonomies of facial configurations with what they believed to be corresponding character interpretations (for a historic account on physiognomy, see [99]). Critical data science research points to several ethical concerns resulting from the AI classification of people based on their facial appearance. Hanley et al. criticize that inferences about people based on visual data necessarily represent only those factors of an inference concept that are visibly discernible [40]. Similarly, Stark & Hoey underscore a "fixation on the visible" in their conceptual analysis on the ethics of emotion recognition AI [93]. Computer vision AI inference-making can be presumptuous when designed to predict aims or intentions of people in images [49]. Such systems are morally objectionable because they treat individuals as objects of categorization [40, 50]. Studying the influential ImageNet dataset, Crawford & Paglen find "highly questionable semiotic assumptions [that] echoe(s) of nineteenth-century phrenology" [22]. Other authors call for a ban on "Physiognomic AI" altogether [94].

Research in fairness, accountability, and transparency has successfully produced different formalizations of fairness metrics and approaches for de-biased datasets. However, when it comes to fair visual data inferences it is the selection of target variables that requires careful ethical consideration. If such ethical evaluations are "subjective" and "inescapably political", then how can we make progress in justifying a line between permissible and impermissible inferences? Contributing to this metaethical challenge, we analyze non-experts' ethical evaluations of specific computer vision AI inferences in a low-stake advertising and a high-stake hiring context. We argue that the input of non-experts (i.e., their moral intuitions) can help us critically advance the debate concerning fair computer vision inferences. We consider a participatory approach to be *at least* complementary to conceptual ethical analyses. For example, much of AI ethics in companies and research institutes is guided by "principlism": efforts of expert groups defining often vague ethical principles for algorithmic systems such as transparency, justice or responsibility [44]. Principlism has recently received criticism (e.g., [71]) arguing that abstract ethical principles too often leave room for interpretation and are therefore particularly susceptible to forms of "ethics washing" [12]. Relying on ethical principles alone critically fails to account for the influence of unique contextual factors on the ethical status of AI inference-making. Moreover, by democratic principle, whenever power hierarchies lead to an accountability vacuum, non-expert "users" should have – minimally – a voice in

[8] In evolutionary psychology, current research debates whether facial attributes (first impressions) are solely innate, evolutionary adaptive heuristics [99] or whether they also have a learned, cultural dimension [78, 79].

[9] ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results: https://hal.archives-ouvertes.fr/hal-01381149, 2017 Looking at People CVPR/IJCNN Competition: https://chalearnlap.cvc.uab.cat/challenge/23/description/

formulating values for the interpretative potential of visual data, including their own. We see this as one element of a holistic approach to advance computer vision AI ethics. For the purpose of the current study, we developed a factorial vignette study that we describe in more detail in the next section. Experimental vignette studies have been extensively used in different fields (including human computer interaction, psychology, experimental philosophy, business ethics) to elicit participants' explicit ethical judgments in a variety of hypothetical scenarios [2, 3, 18, 36, 42, 52, 53, 66, 73]. Our study follows calls for more survey-based AI computer vision ethics [85] and more experimentally-informed AI ethics in general [55]. For a review on the value of studying the "moral intuitions" of non-experts in ethics and philosophy more generally, see [53].

3 METHODS AND EXPERIMENTAL PROCEDURE

3.1 Data Collection

3745 subjects (male = 50.7%, female = 48.9%, other = 0.4%) participated in our study. Subjects were recruited via Amazon Mechanical Turk. Only "Turkers" with an approval rating above 95% were selected for the study. We deliberately chose to conduct our study via this platform because Turkers have been indispensable for the labeling of some of the most important datasets in computer vision [91, 106]. Besides the large subject pool required for our study, we were interested to understand how a community involved in the labeling of computer vision datasets would ethically evaluate AI facial inference-making.

Our home institution does not require an ethics approval for questionnaire-based online studies. When conducting the study and analyzing the data, we followed standard practices for ethical research: presenting detailed study procedures, obtaining consent, not collecting identifiable information or device data, and using a survey service[10] that guaranteed compliance with the European Union's General Data Protection Regulation. The study did not include any deceptive practices. Subjects could drop out of the study at any point. All data were fully anonymized, the privacy of all subjects was maintained at all times during the study. Following recommended principles of ethical crowdsourced research [104], we first ran a pre-study with 120 Turkers to determine the average time it would take to complete the survey and used this reference time to determine a payout above the US minimum wage (*mean*= 8.03 min). In our study (N = 3745), the *mean* was 10.4 min (min = 3.35 min, max = 31.55 min).

3.2 Vignette Study

The experiment was a between-subject design; each participant was randomly assigned to one of 24 groups. The 24 groups were composed of three experimentally altered variables: two decision contexts (advertising vs. hiring), six evaluative adjective terms (reasonable, fair, justifiable, acceptable, responsible, appropriate), and the presentation or absence of a dictionary definition of the evaluative adjective term. The use of different evaluative adjective terms with or without a dictionary definition accounted for framing effects and tested the robustness of subjects' conception of a normative AI inference [53, 55, 64].

First, subjects were randomly assigned to one of two hypothetical decision contexts: either a low-stake advertisement scenario (n = 1869; mean per group = 155) or a high-stake hiring scenario (n = 1876; mean per group = 156). In the hypothetical advertisement scenario, participants were told that an advertising company deployed computer vision AI to make a variety of judgments about social media users based on their portrait image. Participants were told that the inferences were used to show users more suitable *product advertisements*. We explicitly referred to product advertisements to avoid associations with political advertisements that could have raised the stakes of the decision context. In the hypothetical hiring scenario, a declared high-stake decision context by other studies on algorithmic perception [48, 90], participants were told that a company used computer vision AI to make a variety of judgments about applicants based on their application photo. Subjects were told that portrait inferences were used, together with other assessment metrics, to determine whether or not a candidate is suitable for a job. These scenarios presented curated, *hypothetical* decision contexts typical in vignette research on moral phenomena [3, 52, 53] and fulfilled one of our study's main purposes: to understand whether non-experts evaluate the same set of AI facial inferences differently across low-stake and high-stake contexts. The vignettes can be found in the Appendix in Figs. 1 and 2.

Second, past research has shown that vignettes can be prone to framing effects and that such effects can indicate weak value dispositions in morally-laden scenarios [17, 53]. In our vignettes, the *evaluative adjective term* that prompted subjects' normative deliberation prior to the primary rating task could have exerted a framing effect. To control for this potential framing effect, each participant was assigned *one* of six evaluative adjective terms – reasonable, fair, justifiable, acceptable, responsible, or appropriate – when performing the rating task: "Do you agree or disagree that this sort of inference made by a software using artificial intelligence is [evaluative adjective term]?". This increased the external validity of our vignette. Using only the evaluative term "fair" could have biased subjects' ratings and justifications. Some people (and in fact cultures) associate the term "reasonable" more descriptively with logical thinking and deliberation while other cultures associate it more prescriptively, such as being honest and responsible [37]. The same was found for people's intuitions about perceptions of normality (also part descriptive, part prescriptive) [9].

Third, studies in experimental philosophy have used "definition vs. no definition" conditions to understand whether subjects use their own intuitive concept when they evaluate essentially contested concepts (such as: what is a reasonable inference?) [52, 53, 55, 64]. Accordingly, half of subjects were presented with a generic dictionary definition of the evaluative adjective term assigned to them, the other half was not. For example: "What do we mean by fair? Something is fair if it's based on equality without favoritism or discrimination." All definitions were taken from the Cambridge Dictionary and were slightly adjusted for our context (see Appendix Table 1). The "definition vs. no definition" treatment allowed us to further test the robustness of subjects' normative evaluations for specific AI inferences: If non-experts' normative judgments were arbitrary to the extent that they could be manipulated by

[10] SoSci Survey: https://www.soscisurvey.de/

the presentation of a different evaluative adjective term (fair vs. reasonable, for example) or absence of a generic definition of that term, then this would indicate subjects' concept of a normative AI inference to lack robustness. Subjects would then have a low value disposition toward AI facial analysis inferences (studies in experimental philosophy typically use such and similar framing conditions see, for example, [17, 23, 53, 64]).

3.3 Facial inferences

To allow for comparison across contexts, inferences needed to have an acceptable degree of appropriateness for two very different decision contexts: advertising and hiring. To keep the cognitive load of our subjects at an acceptable level, we restricted the number of inferences rated and justified by each subject. We decided to present subjects with a total of eight inferences, first asking them to rate their agreement/disagreement and then to provide a short, written justification for each inference rating. We selected the inference "emotion expression" due to its prevalence in emotion detection AI [20, 93]. Similarly, the two inferences "skin color" and "gender" are common attributes in AI inference-making [14, 46]. Four inferences – "trustworthiness", "assertiveness", "intelligence", "likability" – were selected for their importance in studies on *human* first impression-making [76, 78, 79, 99–101]. Finally, we wanted to understand how subjects would evaluate a facial accessory. We chose "glasses" instead of piercings or tattoos, for example, because the latter two objects exist in more diverse forms. We constructed an 8-item scale to measure agreement with these eight facial inferences made by an AI on a 7-point Likert scale (1 = "strongly agree" to 7 = "strongly disagree", "can't answer"). We did not present subjects with sample portraits, since the impression they would have formed based on the face in the portrait would have likely influenced their normative judgments [99, 100]. The goal of the study was to explore non-experts' ethical evaluations of facial AI inferences *in principle*.

3.4 Classification of subjects' justifications

After rating each inference, subjects were asked to justify their evaluation in a written statement. This allowed us to understand the rationale behind subjects' inference ratings and increased data quality (e.g., understanding the plausibility and validity of evaluations, see, [57, 58]). While there is an entire research field dedicated to studying first impressions (e.g., [99–101]), we could not identify studies investigating people's *ethical evaluations* of such first impressions. This meant that we could not draw from an existing coding scheme for the classification of the 29,760 written justifications. Therefore, we derived the codes directly from the textual corpus. The manual coding process consisted of two iterative cycles. First, one researcher labelled 500 comments to discover major recurring types of reasoning. Another researcher labelled 250 of these comments with the same intent. The researchers then met to discuss and refine the set of identified "justification labels". In a second coding cycle, we randomly sampled 1,250 comments. Two researchers independently added a justification label to each comment. The intercoder reliability was high (Krippendorff's alpha = 0.953). In case of disagreement between the two coders, the comment was discussed with and reviewed by a third researcher. The final set of justification types consisted of the following: 1. "AI can tell", 2. "AI cannot tell", 3. "Inference relevant for decision", 4. "Inference not relevant for decision", 5. "Inference creates harm", 6. "AI has human biases", and 7. "Incomprehensible responses".

Based on this developed coding scheme, we used the language model RoBERTa [63] to analyze the remaining comments. RoBERTa is a more efficiently trained version of BERT [24], an NLP architecture designed for general-purpose language understanding. This required collecting 100 example comments for each justification type (i.e., code). One researcher collected 100 example comments for each justification type. A second researcher then verified classifications. Disagreement was resolved by a third researcher. We split our labeled dataset in 1,001 training and 250 test samples, and performed over-sampling of the smaller classes to create a balanced training dataset. The final optimized model had an overall accuracy on the test set of 95% and each label's F-1 score was higher than 0.94. For the optimization process, we used a learning rate of 3e-5, a maximum sequence length of 32 tokens, and warm-up initialization. We then predicted the labels of the remaining justifications based on the trained model. For the class overview with F-1 scores, see Appendix Table 7.

Our analysis strategy comprised statistical testing of subjects' inference ratings, an exploratory factor analysis, automated text classifications, and a multivariate analysis of variance with follow-up tests. Given the large number of subjects in our sample, we calculated the effect sizes for all significant ($p<0.01$) test results on subjects' ratings.

4 RESULTS

4.1 The consequentiality of the scenario influences non-experts' ethical evaluations of AI facial inferences

We first compared mean aggregate ratings of all inferences between the advertisement and the hiring scenario. A two-sided Welch two-sample t-test found subjects showed greater preference for the same set of inferences in the advertisement scenario (*mean*=3.85; *SE*=1.06) than in the hiring scenario (*mean*=4.41; *SE*=1.2). The difference was significant (t(3687.3)=-15.30; $P<0.001$; 95% *CI*: (-0.64, -0.49)) and represented a small to medium effect ($d=0.50$) (Fig. 1a).

We then compared mean ratings for each inference in the advertisement and hiring scenarios using a two-sided Welch two-sample t-test with Bonferroni corrections for eight tests (Fig. 1b). Subjects rated the inferences gender, emotion expression, wearing glasses, and skin color (e.g., skin color, *mean* AD=2.88, mean HR=4.19; $d=0.60$; $P<0.001$; 95% *CI*: (-1.44, -1.17)) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. In contrast, the inference ratings for intelligence, trustworthiness, and likability (e.g., likability, *mean* AD=5.04, *mean* HR=5.16; $d=0.06$; $P=0.31$; 95% *CI*: (-0.24, -0.006)) did not show a significant difference between the two scenarios. Ratings for the assertiveness inference were significantly different between the two scenarios, but the effect size was negligible (*mean* AD=4.69, *mean* HR=4.89; $d=0.10$; $P=0.01$; 95% *CI*: (-0.32, -0.078)).

To summarize, comparing the inference ratings solely based on the grouping variable *context*, the consequentiality of the decision context influenced subjects' ratings: in the hiring context, subjects showed significantly more disagreement with the AI inferences

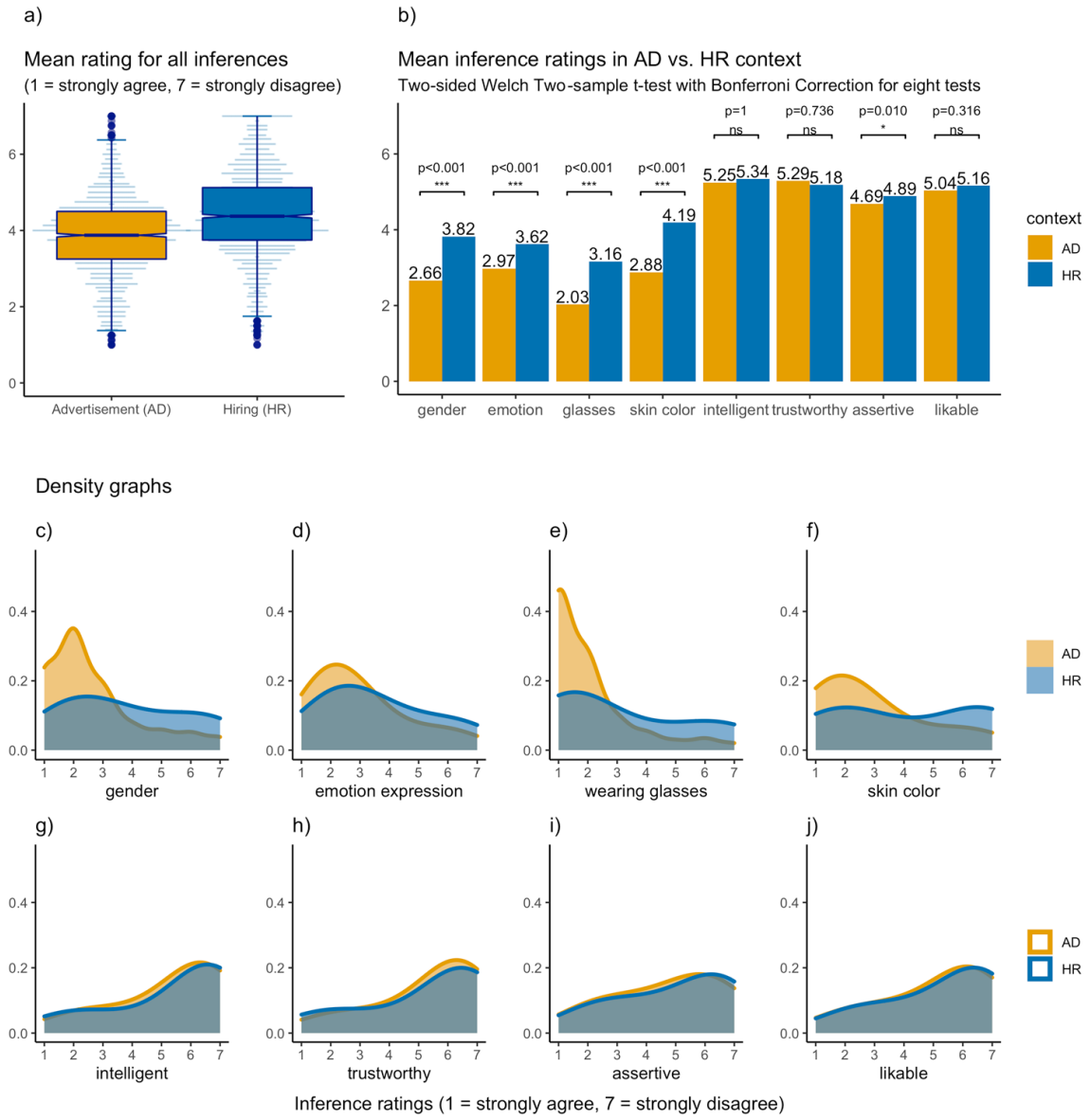


Figure 1: (a) Mean aggregate ratings for inferences were more positive in the advertising context than in the hiring context. (b) Participants rated the inferences gender, skin color, emotion expression, and wearing glasses significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent, trustworthy, and likable did not show a significant difference between the two scenarios. Only ratings for the inference assertive were significantly different between the two scenarios, but the effect was negligible (see Appendix 5 for statistics). (c-j) Density plots of inference ratings. 1 = strongly agree; 7 = strongly disagree; 4 = neutral.

gender, skin color, emotion expression, and glasses than in the advertising context. Cohen's *d* was particularly large for ratings on gender, skin color, and wearing glasses between the two contexts. This difference did not replicate to ratings for the inferences trustworthiness, intelligence, assertiveness, and likability (Fig. 1).

4.2 Subjects differentiate between "first-order" and "second-order" inferences

To explore underlying constructs in our set of eight inferences, we conducted an exploratory factor analysis (EFA) (Appendix 6). Parallel analysis, scree plot, and the MAP criterion all suggested

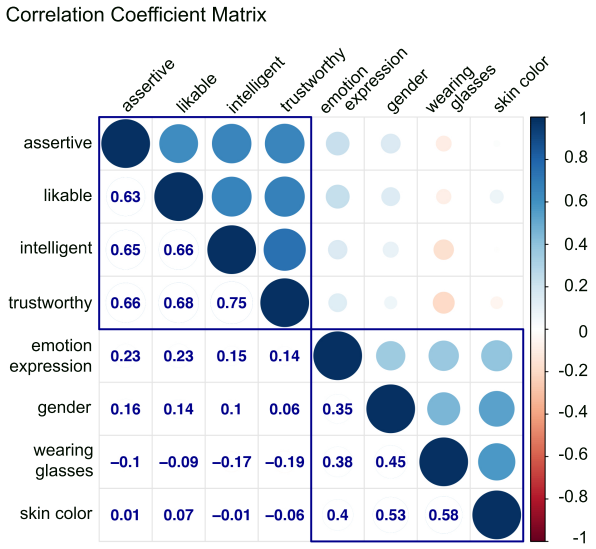


Figure 2: Exploratory factor analysis (EFA) resulted in two underlying constructs for subjects' ratings. One factor included the emotion expression, gender, wearing glasses, and skin color inferences. We termed this set of inferences *first-order inferences*. The other factor included the latent trait inferences assertive, likable, intelligent, and trustworthy. We termed this set of inferences *second-order inferences*.

two factors. One factor included the inferences gender, skin color, wearing glasses, and emotion expression. To use this group of inferences for further statistical comparison, we termed this construct *first-order inferences*. The other factor included the four latent trait inferences intelligence, trustworthiness, assertiveness, and likability. We termed this construct *second-order inferences*. We used these terms (first-order/second-order) as linguistic categories to reflect the statistical reality of subjects' ratings and less as an initial semantic interpretation of subjects' ethical evaluations. Both sub-scales had high reliability, the overall α was 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences* (Fig. 2; see Appendix 6.6 for distribution of EFA factor scores).

4.3 Decision context only influences agreement with first-order inferences

We then extended our analysis to the entire set of treatment conditions. To test significant group differences among the 24 treatment groups on a combination of *first-order* and *second-order* factor scores from the EFA as a dependent variable, we computed a 2 (context: advertisement, hiring) x 6 (evaluative adjective terms) x 2 (definition, no definition) multivariate analysis of variance (MANOVA; Appendix 7). We controlled for main first-order justification theme, main second-order justification theme, AI knowledge, age, gender, occupation and education. Using Pillai's trace, there were significant main effects at an α-level of 0.01 for first-order justification (V=0.50, F(12, 6892)=190.76, P <.001, partial η^2 = 0.249), second-order justification (V=0.45, F(12, 6892)=164.60, P <.001, partial η^2 = 0.223), AI knowledge (V=0.03, F(8, 6892)=13.43, P <.001, partial η^2 = 0.015), and context (V=0.04, F(2, 3445)=73.68, P <.001, partial η^2 = 0.041) (Appendix Table 5).

Finally, univariate analysis with two separate ANOVAs on the *first-order* factor scores and on the *second-order* factor scores from the EFA revealed varying effect structures (Table 1; Appendix 7.2). With respect to the experimentally altered variables, *context* was the only significant treatment effect found, but only had an effect on ratings of first-order inferences (F(1, 3446) = 146.08, P <0.001, partial η^2 = 0.04). This finding supported the results from the two-sided Welch two-sample t-test. The experimental treatments *evaluative terms* and *definition vs. no definition* had no significant effect on subjects' ratings. This indicated that the subjects in our sample had a robust concept of a normative facial AI inference. AI knowledge had a small but significant effect on both inference ratings, whereas age had only a small effect on first-order ratings. Gender, occupation, and education did not have a statistically significant effect on subjects' ratings. Pairwise comparisons confirmed the results by identifying significant group differences between the advertisement and hiring context (Appendix 7.3).

4.4 Subjects find AI cannot tell second-order inferences in both contexts. Gender, skin color, and emotion expression produce more complex justifications.

4.4.1 Subjects evaluate the normativity of an AI inference according to two meta-principles. In their written evaluations, subjects considered whether or not an inference was proportional to the evidence (i.e., an epistemic justification) or whether making the inference resulted in positive or negative outcomes (i.e., a pragmatic justification). Representing epistemic principles, we introduced two codes: "AI can tell" and its opposite "AI cannot tell". For example, the comment *"I believe that someone's facial expressions can easily tell if they are assertive. I feel like facial expressions are easy to read and a computer could do that even better."* (assertiveness, HR) was classified as "AI can tell". The comment *"A person's intelligence is internal and based on learning, education, and other experiences. This can't be reflected in someone's looks."* was classified as "AI cannot tell" (intelligence, HR).

With the second meta-principle, subjects considered pragmatic reasons: we identified two contrary justification types "Inference relevant for decision" and "Inference not relevant for decision". The justification *"The reason I believe it is appropriate...is because this will help to select the potential candidate that possesses the assertiveness that could be useful for the job."* was classified as "Inference relevant for decision". The comment "*I don't think assertiveness makes or breaks a job applicant*" was classified as "Inference not relevant for decision" (both assertiveness, HR). A third justification type "Inference creates harm" classified comments stating AI inference-making could be harmful if used as part of the decision-making process (e.g., discrimination due to racism or sexism). For example, the justifications "*Seems like phrenology where intelligence and other traits were determined by the shape of someones head.*" (intelligence,

Table 1: Follow-up ANOVAs for factor scores from exploratory factor analysis (EFA)

	ANOVA for first-order					ANOVA for second-order				
	SS	df	F	Bonferroni	part. η^2	SS	df	F	Bonferroni	part. η^2
(Intercept)	7.32	1	22.22	**0.000**	0.006	5.135	1	15.399	**0.001**	0.004
Justifications										
first-order justifications	946.163	6	478.774	**0.000**	**0.455**	46.331	6	23.157	**0.000**	**0.039**
second-order justifications	18.785	6	9.506	**0.000**	**0.016**	844.717	6	422.212	**0.000**	**0.424**
Control Variables										
AI knowledge	14.069	4	10.679	**0.000**	**0.012**	26.058	4	19.537	**0.000**	**0.022**
age	9.939	5	6.035	**0.000**	0.009	5.648	5	3.387	0.052	0.005
gender	0.272	2	0.414	1.000	0.000	2.463	2	3.693	0.275	0.002
occupation	7.834	8	2.973	0.028	0.007	5.720	8	2.144	0.317	0.005
education	1.553	7	0.674	1.000	0.001	2.749	7	1.178	1.000	0.002
Experimental Variables										
context	48.115	1	146.081	**0.000**	**0.041**	2.325	1	6.972	0.092	0.002
terms	6.502	5	3.948	0.016	0.006	5.140	5	3.083	0.097	0.004
definition	0.161	1	0.487	1.000	0.000	0.293	1	0.880	1.000	0.000
Residuals	1135.010	3446				1149.065	3446			

Note:
All Bonferroni-corrected P-values are compared to a Bonferroni-corrected $\alpha = 0.005$ for the computation of two ANOVAs.
Significant P-values and partial η^2 values of relevant size are marked in **bold**.
Partial $\eta^2 = 0.01$ small effect; partial $\eta^2 = 0.06$ medium effect; partial $\eta^2 = 0.14$ large effect.

AD) or *"Color should not matter in job hiring. This would be discrimination."* (skin color, HR) were classified as "Inference creates harm". Finally, a justification type that we called "AI has human biases" classified comments stating AI inference-making was flawed by biased human inference-making. Justifications in "AI has human biases" contained epistemic reasons (e.g., *"The software could be implanted with the bias of its creator"*; trustworthy, HR) or pragmatic reasons (e.g., *"The inference is unfair as the AI may be programmed to favor one sex over the other without context."*; gender, HR).

The classification results of subjects' written responses underline the semantic ambiguity of facial portraits: for each inference, we found a corpus of diverse explanations that fell back on epistemic and pragmatic accounts (the two meta-principles). We show the general line of subjects' justifications in Fig.3, where we map ratings (agreement/disagreement) to justification types. We complement subjects' general line of justifications with example comments. More example comments can be found in our "code book" in Appendix Table 8.

4.4.2 Subjects believe AI second-order inferences are invalid inferences regardless of the decision-making context. The majority of subjects believed that faces do not provide sufficient evidence ("AI cannot tell") for inferences intelligence, trustworthiness, likability, and assertiveness (i.e., all second-order inferences) – regardless of the decision context. *"If you're just looking at a person and trying to determine if they're assertive, you're going to score no better than a random guess, I don't care how sophisticated this AI is."* (assertiveness, HR). Some subjects believed second-order inferences to be epistemically valid. *"Assertive people tend to have a set in their jaw, and eyes that is a bit more severe in the angles at the corners than those who are more passive...It might be possible to quantify those angles and measurements to have an AI program analyze the likelihood that they match those of assertive people...If you can come up with a mathematical formula to determine this, then the AI would be capable of measuring it."* (assertiveness, HR). The largest group of subjects agreeing with second-order inferences argued for their *relevance* in the hiring context (54.8%, "Inference relevant for decision"). Here, subjects did not express any epistemic reasoning, but asserted that such inferences were desirable qualities for employers. *"Almost always when you are working, you will work in teams and have to get along with others. You have to be likable to be successful on these teams - I would want the AI to try and assess this as best they could."* (likability, HR).

4.4.3 Subjects believe first-order inferences are epistemically valid, but irrelevant and harmful in hiring. For the inferences emotion expression, wearing glasses, skin color, and gender, subjects' justification profile was more complex (Fig.3 e-j). The majority of subjects that agreed with these inferences believed in their epistemic validity in both contexts ("AI can tell"; AD: 62.6%, HR: 68.1%). However, in comparison to second-order inferences, the justification patterns differed between the advertising and hiring context: in the hiring

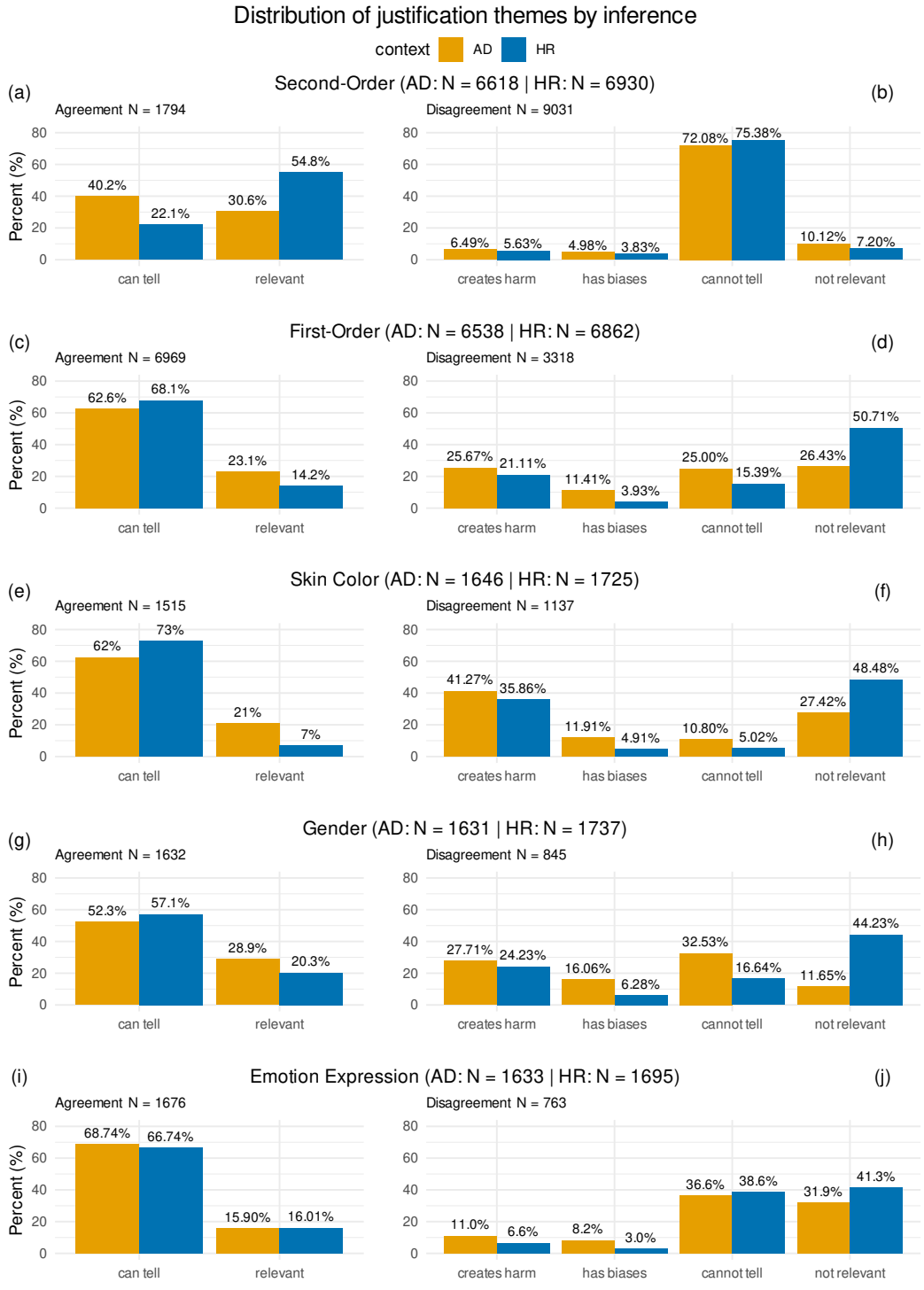


Figure 3: Distribution of justification types. Plots a) to o) present the proportions of the justification types used per context. E.g., for first-order ratings, 62.6% of participants in the AD context justified their agreement with an explanation allocated to the justification type "AI can tell" and 50.71% of respondents in the HR context justified their disagreement with an explanation related to the justification type "not relevant". The sum of N for AD and HR for an inference does not amount to the total N because the plot does not include individuals who neither agreed or disagreed. Percentages by context and agreement/disagreement do not sum up to 100%, since the visualization does not include a minority of individuals who provided a counter-intuitive justification based on their score.

context, considerations of relevance became more important reasons to reject an inference in comparison to the advertising context (Fig.3 c). The majority of subjects agreeing with skin color and gender in both contexts believed an "AI can tell" such inferences from facial information (Fig.3 e-h): *"Photos reveal this pretty easily assuming the photo is reasonably high rez. I would probably trust a computer to get this right more than some people."* (skin color, HR) or *"This is something that we, as humans can perceive with our sights, so an AI is definitely capable of inferring this.*" (gender, AD). However, subjects that believed "AI can tell" skin color and gender still raised concerns in their written responses even when agreeing with these inferences. For example, subjects noted that accurately inferring skin color may be constrained by photo quality and lighting and may not be an indication of race or ethnicity as the following two comments illustrate: *"I believe a properly calibrated AI could estimate a person's skin color, but lighting, photo quality etc., would have to be accounted for. Also, skin color doesn't necessarily inform us about race."* (skin color, HR). *"Mixed feelings about this one – although skin color is something that can be visually seen in a photo, there is lots of room for error here depending on lighting in photo. Also, whether it's morally right is a whole different subject."* (skin color, AD). Likewise, for gender, subjects pointed to classification problems of non-binary gender identities: *"For the most part, male/female is an easy question, but there are many people that defy these binary categories that would be excluded."* (gender, HR).

Among the subjects rejecting skin color and gender in hiring, the most common justifications were "Inference not relevant for decision" (skin color: 48.48%; gender: 44.23%) and "Inference creates harm" (skin color: 35.86%; gender: 24.23%). With regard to skin color, most comments stated that skin color does not matter in hiring, while a few added that the inference was justifiable if it resulted in a more diverse workplace: *"This does not matter unless this information is being used to ensure a diverse workplace."* (skin color, HR). Subjects generally agreed that gender does not matter in hiring, however, some subjects asserted that some jobs may be more suitable for certain genders: *"Gender has nothing to do with how capable a person is to do a job unless the job itself requires a specific gender (which is very rare)."* (gender, HR). In contrast, subjects believed that both skin color (21%) and gender (28.9%) are a relevant AI inference in advertising: *"People with different skin colors need different products, and tend to shop for different styles, colors, and patterns."* (skin color, AD) or *"I think this is a 50/50 subject, but I believe personally that this is fair...Perhaps men wouldn't like to see advertisements for bras which would be avoided with this scan."* (gender, AD).

4.4.4 A majority of subjects believe emotion expression indicates emotion sensation. For emotion expression (Fig.3 i-j), subjects' agreement or disagreement mainly depended on whether or not they believed facial expressions to be a valid indicator for emotion sensation. Comments classified as "AI can tell" (agreement, AD: 68.74%, HR: 66.74%) claimed internal emotional states could be expressed via the face: *"It is reasonable to judge emotions by looking at a person's face, humans do it all the time. Though some faces can be more expressive than others."* (emotion expression, HR). Given that many Turkers have engaged in portrait image labelling tasks, we also found comments that highlighted the possibility of AI emotion expression inference based on previously conducted labelling tasks: *"A person's emotion can be seen pretty well by looking at a picture as I have done surveys in the past deciding emotion through facial expressions"* (emotion expression, AD). Comments classified as "AI cannot tell" (disagreement, HR: 38.6%, AD: 36.6%) stated the opposite. *"An emotion could be expressed, but the person may not actually be expressing it. In other words, the emotion viewed externally could be one of joy, but, inside the actual person, they may have a different emotion from what is outwardly being expressed."* (emotion expression, HR). The difficult relationship between emotion expression and emotion inference was also evident in comments with the justification types "Inference relevant for decision" (agreement, AD: 15.9%, HR: 16.01%) and "Inference not relevant for decision" (disagreement, AD: 31.9%, HR: 41.3%). To give one example, in comments classified as "Inference relevant for decision" in hiring, subjects claimed that employers may seek employees that need to be friendly, particularly in jobs involving customer interaction: *"Depending on the job emotional expressiveness may be a requirement, you don't want a person in a customer service position who's monotonous and robotic."* (emotion expression, HR).

5 KEY OBSERVATIONS & FINAL DISCUSSION

The vast abundance of digital imagery together with recent advances in computer vision analysis have raised concerns about the kinds of conclusions AI should make about people based on their face. How do we design computer vision AI in such a way that it will incorporate those preferences and values that are ethically desirable? We explored non-experts' normative preferences of AI portrait inferences in a two-scenario vignette study with 24 treatment groups. One MANOVA and two ANOVAs found that none of our framing effects influenced subjects' ratings, indicating that subjects have a robust, intuitive concept of a normative AI inference for both contexts. Future studies need to further explore how strong this normative concept is in light of other trade-offs such as cost-efficiency, narratives of bias-free technology, or success of the decision outcome, for example.

Conducting an exploratory factor analysis on subjects' evaluations of eight AI facial inferences, two inference categories emerge: we term one category of inferences first-order inferences and the other second-order inferences. Factor loadings of emotion expression as a first-order inference together with subjects' justifications suggest that a majority of the subjects in our sample subscribe to the so-called "Basic View" of emotions [28], which proposes that facial expressions (or "facial action units") are reliable indicators of emotion. Note that this perspective has recently been challenged by emotion researchers arguing that contextual and social factors lead to variability in facial emotion expression that make such inferences unreliable and unspecific [7, 93]. Nonetheless, subjects are aware of the volatility of AI emotion inference from facial expression. They assert that emotion expression as social signaling can be different from the internal phenomenological experience.

Finally, independent of the decision context, subjects believe AI should not draw inferences common in human first facial impression-making due to their epistemic invalidity, i.e., intelligence, likability, assertiveness, and trustworthiness [99–101]. Subjects raised concerns about all AI inferences in both contexts, even for the – perhaps

intuitively – non-problematic "glasses" inference in the low-stake advertising context (Appendix Fig. 7). This leads us to assume that other facial AI inferences, such as beauty, sexual orientation, or political stance, that all have been inferred from faces using AI will likely draw their own justification profiles.

Our analysis highlights the normative complexity behind facial AI inferences. We find that some subjects use a *pragmatic* rationalization of AI facial inferences when they believe that an AI inference is relevant for (i.e., has a supposedly positive effect on) a decision's outcome. However, why should the normativity of a *vision*-based inference be evaluated by criteria other than evidence? The decision context does not have any bearing on the relationship between evidence and inference and therefore should not lead to a different normative evaluation. Thus, our results show that epistemically invalid AI vision inferences can be rationalized by considerations of relevance. The fact that AI research organizations, academic and commercial, commission data annotation companies to label visual data relevant for a specific application purpose necessarily creates a conflicting negotiation between epistemic and pragmatic considerations. Taken together, over-reliance on AI capabilities, narratives of bias-free technological decision-making, and beliefs in the relevance of an inference for the decision context may form a line of reasoning that supports justification of epistemically invalid AI inference-making. The ongoing publication of research studies that purportedly find a significant correlation between second-order inferences and facial information produces a quasi-epistemic legitimization of first-impression AI. Our study provides evidence that a vast majority of non-expert subjects do not form a justification of AI inference-making along these lines of reasoning.

Finally, how would experts differ in their justification of AI inference-making in comparison to non-experts? Indeed, critical data scientists argue that facial inferences are not reasonable because of their lack of scientific validity (evidentialists) [20, 92], while some AI experts deploying computer vision AI point to positive outcomes in terms of efficiency, cost-reduction, and flexibility that AI inference-making will facilitate [8, 43, 61, 65, 96]. Future studies will need to provide evidence for a unique ethical justification profile of AI vision inferences among AI expert groups. Other future studies should explore to what extent cultural factors play a role in evaluating the normativity of AI inferences based on visual data. We also believe it would be valuable to understand whether subjects evaluate AI video analysis inferences differently than AI image inferences. In fact, AI video analysis interprets visual content at the level of individual frames (i.e., decomposed as a collection of single images) [38].

We hope that the present study underlines the importance of including non-experts in the process of arguing for and against ethically permissible and non-permissible computer vision inferences. We expect norms regarding AI inference-making to shift over time. Allowing non-experts to engage in the formulation of goals and values for AI helps identify such shifts in sociocultural norms. Our study lays an important foundation for determining what types of inferences machines should and should not make about one of the most significant characteristics of us and our place in the social world: our faces.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments that improved the paper. For their valuable feedback we thank the participants of the 2021 CEPE/International Association of Computing and Philosophy Conference, the participants of the 2021 Ethics and Technology Lecture Series of the Munich Center for Technology in Society, and the participants of the Venice 2019 Metaethics of AI & Self-learning Robots Workshop.

FUNDING & SUPPORT

This research was conducted with the help of a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

REFERENCES

[1] Noura Al Moubayed, Yolanda Vazquez-Alvarez, Alex McKay, and Alessandro Vinciarelli. 2014. Face-based automatic personality perception. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 1153–1156. https://doi.org/10.1145/2647868.2655014

[2] Kwame Anthony Appiah. 2008. *Experiments in Ethics*. Harvard University Press.

[3] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138. https://doi.org/10.1027/1614-2241/a000014

[4] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences* 124 (2018), 150–159. https://doi.org/10.1016/j.paid.2017.12.018

[5] Mitja Back, Juliane Stopfer, Simine Vazire, Sam Gaddis, Stefan Schmukle, Boris Egloff, and Samuel Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science* 21, 3 (2010), 372–374. https://doi.org/10.1177/0956797609360756

[6] Charles C. Ballew and Alexander Todorov. 2007. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17948–17953. https://doi.org/10.1073/pnas.0705435104

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[8] Johannes M. Basch and Klaus G. Melchers. 2019. Fair and flexible?! Explanations can improve applicant reactions toward asynchronous video interviews. *Personnel Assessment and Decisions* 5, 3 (2019), Article 2. https://doi.org/10.25035/pad.2019.03.002

[9] Adam Bear and Joshua Knobe. 2017. Normality: Part descriptive, part prescriptive. *Cognition* 167 (2017), 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

[10] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. 2016. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5562–5570. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Benitez-Quiroz_EmotioNet_An_Accurate_CVPR_2016_paper.html

[11] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. 53–56. https://doi.org/10.1145/2388676.2388689

[12] Elettra Bietti. 2020. From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 210–219. https://dl.acm.org/doi/abs/10.1145/3351095.3372860

[13] Jean-François Bonnefon, Astrid Hopfensitz, Wim De Neys, et al. 2015. Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences* 19, 8 (2015), 421–422. https://doi.org/10.1016/j.tics.2015.05.002

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[15] Filippo Cavallo, Francesco Semeraro, Laura Fiorini, Gergely Magyar, Peter Sinčák, and Paolo Dario. 2018. Emotion modelling for social robotics applications: A review. *Journal of Bionic Engineering* 15, 2 (2018), 185–203. https://doi.org/10.1007/s42235-018-0015-y

[16] Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 1101–1104. https://doi.org/10.1145/2647868.2654977

[17] Dennis Chong and James N. Druckman. 2007. Framing Theory. *Annual Review of Political Science* 10, 1 (2007), 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

[18] Cory J. Clark, Jamie B. Luguri, Peter H. Ditto, Joshua Knobe, Azim F. Shariff, and Roy F. Baumeister. 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106, 4 (2014), 501–513. https://doi.org/10.1037/a0035880

[19] Jeff F. Cohn and Fernando De la Torre. 2015. Automated face analysis for affective computing. In *The Oxford Handbook of Affective Computing*, Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas (Eds.). Oxford University Press, 131–150. https://doi.org/10.1093/oxfordhb/9780199942237.013.020

[20] Kate Crawford. 2021. Time to regulate AI that interprets human emotions. *Nature* 592, 7853 (2021), 167–167. https://doi.org/10.1038/d41586-021-00868-5

[21] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, A. Sánchez, et al. 2019. *AI Now 2019 report.* Research Report. The AI Now Institute, NYU. https://ainowinstitute.org/AI_Now_2019_Report.pdf

[22] Kate Crawford and Trevor Paglen. 2019. *Excavating AI: The politics of images in machine learning training sets.* Research Report. The AI Now Institute, NYU. https://excavating.ai/

[23] Joanna Demaree-Cotton. 2016. Do framing effects make moral intuitions unreliable? *Philosophical Psychology* 29, 1 (2016), 1–22. https://doi.org/10.1080/09515089.2014.989967

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* (2018). arXiv:1810.04805

[25] Abhinav Dhall, Oruganti V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 423–426. https://doi.org/10.1145/2818346.2829994

[26] Rafaële Dumas and Benoît Testé. 2006. The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie* 65, 4 (2006), 237–244. https://doi.org/10.1024/1421-0185.65.4.237

[27] Charles Efferson and Sonja Vogt. 2013. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports* 3 (2013), Article 1047. https://doi.org/10.1038/srep01047

[28] Paul Ekman and Wallace V. Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues.* Malor Books.

[29] Severin Engelmann and Jens Grossklags. 2019. Setting the stage: Towards principles for reasonable image inferences. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 301–307. https://doi.org/10.1145/3314183.3323846

[30] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with Instagram pictures. In *Proceedings of the Workshop on Emotions and Personality in Personalized Systems*. 7–10. https://doi.org/10.1145/2809643.2809644

[31] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2016. Using Instagram picture features to predict users' personality. In *Proceedings of the 22nd International Conference on Multimedia Modeling*. 850–861. https://doi.org/10.1007/978-3-319-27671-7_71

[32] Bruce Ferwerda and Marko Tkalcic. 2018. Predicting users' personality from Instagram pictures: Using visual and/or content features?. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 157–161. https://doi.org/10.1145/3209219.3209248

[33] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[34] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336. https://doi.org/10.1145/3351095.3372862

[35] Jake Goldenfein. 2019. The profiling potential of computer vision and the challenge of computational empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 110–119. https://doi.org/10.1145/3287560.3287568

[36] Armin Granulo, Christoph Fuchs, and Stefano Puntoni. 2019. Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour* 3, 10 (2019), 1062–1069. https://doi.org/10.1038/s41562-019-0670-y

[37] Igor Grossmann, Richard P. Eibach, Jacklyn Koyama, and Qaisar B. Sahi. 2020. Folk standards of sound judgment: Rationality versus reasonableness. *Science Advances* 6, 2 (2020), Article eaaz0289. https://doi.org/10.1126/sciadv.aaz0289

[38] Yağmur Güçlütürk, Umut Güçlü, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel A.J. Van Gerven, and Rob Van Lier. 2017. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing* 9, 3 (2017), 316–329. https://doi.org/10.1109/TAFFC.2017.2751469

[39] Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H. Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying personality through the content of posted and liked images on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. 223–227. https://doi.org/10.1145/3091478.3091522

[40] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 543–554. https://doi.org/10.48550/arXiv.2105.12754

[41] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint* (2018). https://doi.org/10.48550/arXiv.1805.03677

[42] Michael R. Hyman and Susan D. Steiner. 1996. The vignette method in business ethics research: Current uses, limitations, and recommendations. In *Proceedings of the Annual Meeting of the Southern Marketing Association*. 261–265.

[43] Srirang K. Jha, Shweta Jha, and Manoj Kumar Gupta. 2020. Leveraging artificial intelligence for effective recruitment and selection processes. In *Proceedings of the International Conference on Communication, Computing and Electronics Systems*. 287–293. https://doi.org/10.1007/978-981-15-2612-1_27

[44] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[45] Alexander Kachur, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. 2020. Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports* 10 (2020), Article 8487. https://doi.org/10.1038/s41598-020-65358-6

[46] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22. https://doi.org/10.1145/3274357

[47] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 587–597. https://doi.org/10.48550/arXiv.2102.02320

[48] Kimon Kieslich, Marco Lünich, and Frank Marcinkowski. 2021. The threats of artificial intelligence scale (TAI). *International Journal of Social Robotics* 13 (2021), 1563–1577. https://doi.org/10.1007/s12369-020-00734-w

[49] Owen C. King. 2019. Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Don Berkich and Matteo V. d'Alfonso (Eds.). Springer, 265–282. https://doi.org/10.1007/978-3-030-01800-9_14

[50] Owen C. King. 2020. Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. *Ethics and Information Technology* 22, 1 (2020), 25–37. https://doi.org/10.1007/s10676-019-09512-3

[51] Karel Kleisner, Veronika Chvátalová, and Jaroslav Flegr. 2014. Perceived intelligence is associated with measured intelligence in men but not women. *PloS ONE* 9, 3 (2014), Article e81237. https://doi.org/10.1371/journal.pone.0081237

[52] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194. https://www.jstor.org/stable/3329308

[53] Joshua Knobe and Shaun Nichols. 2017. Experimental Philosophy. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[54] Michal Kosinski. 2021. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11, 100 (2021). https://doi.org/10.1038/s41598-020-79310-1

[55] Steven R. Kraaijeveld. 2021. Experimental philosophy of technology. *Philosophy & Technology* 34 (2021), 993–1012. https://doi.org/10.1007/s13347-021-00447-6

[56] Robin S. Kramer and Robert Ward. 2010. Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology* 63, 11 (2010), 2273–2287. https://doi.org/10.1080/17470211003770912

[57] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement?. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 805–814. https://doi.org/10.1145/3209978.3210033

[58] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial*

Intelligence Research 69 (2020), 143–189. https://doi.org/10.1613/jair.1.12012
[59] Gabriel S. Lenz and Chappell Lawson. 2011. Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science* 55, 3 (2011), 574–589. https://doi.org/10.1111/j.1540-5907.2011.00511.x
[60] John Leuner. 2019. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. *arXiv preprint arXiv:1902.10739* (2019). https://doi.org/10.48550/arXiv.1902.10739
[61] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 166–176. https://doi.org/10.1145/3461702.3462531
[62] Anthony C. Little and David I. Perrett. 2007. Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology* 98, 1 (2007), 111–126. https://doi.org/10.1348/000712606X109648
[63] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* (2019). arXiv:1907.11692
[64] Bertram F. Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33, 2 (1997), 101–121. https://doi.org/10.1006/jesp.1996.1314
[65] Julie M. McCarthy, Talya N. Bauer, Donald M. Truxillo, Neil R. Anderson, Ana Cristina Costa, and Sara M. Ahmed. 2017. Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?". *Journal of Management* 43, 6 (2017), 1693–1725. https://doi.org/10.1177/0149206316681846
[66] David E. Melnikoff and Nina Strohminger. 2020. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour* 4 (2020), 1258–1264. https://doi.org/10.1038/s41562-020-00943-3
[67] Milagros Miceli and Julian Posada. 2021. Wisdom for the crowd: Discoursive power in annotation instructions for computer vision. *arXiv preprint* (2021). https://doi.org/10.48550/arXiv.2105.10990
[68] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25. https://doi.org/10.1145/3415186
[69] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 161–172. https://doi.org/10.1145/3442188.3445880
[70] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229. https://doi.org/10.1145/3287560.3287596
[71] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4
[72] Laura Naumann, Simine Vazire, Peter Rentfrow, and Samuel Gosling. 2009. Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin* 35, 12 (2009), 1661–1671. https://doi.org/10.1177/0146167209346309
[73] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41, 4 (2007), 663–685. https://www.jstor.org/stable/4494554
[74] Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*. 557–566. https://doi.org/10.1145/1743384.1743478
[75] Christopher Y. Olivola, Dawn L. Eubanks, and Jeffrey B. Lovelace. 2014. The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly* 25, 5 (2014), 817–834. https://doi.org/10.1016/j.leaqua.2014.06.002
[76] Christopher Y. Olivola, Friederike Funk, and Alexander Todorov. 2014. Social attributions from faces bias human choices. *Trends in Cognitive Sciences* 18, 11 (2014), 566–570. https://doi.org/10.1016/j.tics.2014.09.007
[77] Christopher Y. Olivola, Abigail B. Sussman, Konstantinos Tsetsos, Olivia E. Kang, and Alexander Todorov. 2012. Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science* 3, 5 (2012), 605–613. https://doi.org/10.1177/1948550611432770
[78] Harriet Over and Richard Cook. 2018. Where do spontaneous first impressions of faces come from? *Cognition* 170 (2018), 190–200. https://doi.org/10.1016/j.cognition.2017.10.002
[79] Harriet Over, Adam Eggleston, and Richard Cook. 2020. Ritual and the origins of first impressions. *Philosophical Transactions of the Royal Society B* 375, 1805 (2020), Article 20190435. https://doi.org/10.1098/rstb.2019.0435
[80] Ian S. Penton-Voak, Nicholas Pound, Anthony C. Little, and David I. Perrett. 2006. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition* 24, 5 (2006), 607–640. https://doi.org/10.1521/soco.2006.24.5.607
[81] Lin Qiu, Jiahui Lu, Shanshan Yang, Weina Qu, and Tingshao Zhu. 2015. What does your selfie say about you? *Computers in Human Behavior* 52 (2015), 443–449. https://doi.org/10.1016/j.chb.2015.06.032
[82] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481. https://doi.org/10.1145/3351095.3372828
[83] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151. https://doi.org/10.1145/3375627.3375820
[84] Nicholas O. Rule and Nalini Ambady. 2008. The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science* 19, 2 (2008), 109–111. https://doi.org/10.1111/j.1467-9280.2008.02054.x
[85] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37. https://doi.org/10.1145/3476058
[86] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (2021), Article 20539517211053712. https://doi.org/10.1177/20539517211053712
[87] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–35. https://doi.org/10.1145/3392866
[88] Cristina Segalin, Dong Seon Cheng, and Marco Cristani. 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding* 156 (2017), 34–50. https://doi.org/10.1016/j.cviu.2016.10.013
[89] Crisitina Segalin, Alessandro Perina, Marco Cristani, and Alessandro Vinciarelli. 2016. The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing* 8, 2 (2016), 268–285. https://doi.org/10.1109/TAFFC.2016.2516994
[90] Aaron Smith, Lee Rainie, Kenneth Olmstead, Jingjing Jiang, Andrew Perrin, Paul Hitlin, and Meg Hefferon. 2018. Public attitudes toward computer algorithms. *Pew Research Center* 16 (2018). https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/11/PI_2018.11.19_algorithms_FINAL.pdf
[91] Alexander Sorokin and David Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 1–8. https://doi.org/10.1109/CVPRW.2008.4562953
[92] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129
[93] Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 782–793. https://doi.org/10.1145/3442188.3445939
[94] Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Available at SSRN* (2021). https://doi.org/10.2139/ssrn.3927300
[95] Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *Proceedings of the Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 40–46. https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/download/5350/5599
[96] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61, 4 (2019), 15–42. https://doi.org/10.1177/0008125619867910
[97] Adriana Tapus, Antonio Bandera, Ricardo Vazquez-Martin, and Luis V. Calderita. 2019. Perceiving the person and their interactions with the others for social robotics – A review. *Pattern Recognition Letters* 118 (2019), 3–13. https://doi.org/10.1016/j.patrec.2018.03.006
[98] Thales Teixeira, Michel Wedel, and Rik Pieters. 2012. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research* 49, 2 (2012), 144–159. https://doi.org/10.1509/jmr.10.0207
[99] Alexander Todorov. 2017. *Face Value: The Irresistible Influence of First Impressions*. Princeton University Press.
[100] Alexander Todorov, Sean G. Baron, and Nikolaas N. Oosterhof. 2008. Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience* 3, 2 (2008), 119–127. https://doi.org/10.1093/scan/nsn009
[101] Alexander Todorov, Christopher Y. Olivola, Ron Dotsch, and Peter Mende-Siedlecki. 2015. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology* 66 (2015), 519–545. https://doi.org/10.1016/j.tics.2014.09.007

[102] Richard J.W. Vernon, Clare A.M. Sutherland, Andrew W. Young, and Tom Hartley. 2014. Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences* 111, 32 (2014), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

[103] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114, 2 (2018), 246–257. https://doi.org/10.1037/pspa0000098

[104] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81. https://doi.org/10.1017/S104909651500116X

[105] John Paul Wilson and Nicholas O. Rule. 2015. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science* 26, 8 (2015), 1325–1331. https://doi.org/10.1177/0956797615590992

[106] Gerhard Wohlgenannt. 2016. A comparison of domain experts and crowdsourcing regarding concept relevance evaluation in ontology learning. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*. Springer, 243–254. https://doi.org/10.1007/978-3-319-49397-821

[107] Nan Xi, Di Ma, Marcus Liou, Zachary C. Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*. 726–737. https://ojs.aaai.org/index.php/ICWSM/article/view/7338

[108] Yan Yan, Jie Nie, Lei Huang, Zhen Li, Qinglei Cao, and Zhiqiang Wei. 2015. Is your first impression reliable? Trustworthy analysis using facial traits in portraits. In *Proceedings of the 21st International Conference on Multimedia Modeling*. 148–158. https://doi.org/10.1007/978-3-319-14442-9_13

[109] Leslie A. Zebrowitz and Susan M. McDonald. 1991. The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior* 15, 6 (1991), 603–623. https://doi.org/10.1007/BF01065855

[110] Leslie A. Zebrowitz and Joann M. Montepare. 2008. Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass* 2, 3 (2008), 1497–1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Appendix: What People Think AI Should Infer From Faces

SEVERIN ENGELMANN*, Technical University of Munich, Chair of Cyber Trust, Germany
CHIARA ULLSTEIN*, Technical University of Munich, Chair of Cyber Trust, Germany
ORESTIS PAPAKYRIAKOPOULOS, Princeton University, Center for Information Technology Policy, USA
JENS GROSSKLAGS, Technical University of Munich, Chair of Cyber Trust, Germany

ACM Reference Format:
Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. Appendix: What People Think AI Should Infer From Faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3531146.3533080

*Denotes equal contribution.

Manuscript submitted to ACM

1 VIGNETTE SCENARIOS

a) Advertisement Scenario

A company developed a software that uses **artificial intelligence** to analyze images.

The software analyzes portraits of **users** uploaded to a social media platform in order to show these users suitable advertisements for products. How does that work? The artificial intelligence is presented with a portrait of a user showing only the user's face but nothing else. The software scans the user's face and makes a variety of inferences about the user.

Based on these and other inferences a user will be shown a particular advertising material on the social media platform.

Which statement best describes the scenario presented above?

- ○ Product advertisements will be recommended to a user based on inferences by an artificial intelligence on his or her profile picture.
- ○ Recommended product advertisements are based on inferences by a company's employees, who assess the portraits of users.

b) Hiring Scenario

A company developed a software that uses **artificial intelligence** to analyze images.

The software will analyze portraits of **applicants** in order to select suitable candidates during hiring procedures. How does that work? The artificial intelligence is presented with a portrait of an applicant showing only the applicant's face but nothing else. The software scans the applicant's face and makes a variety of inferences about the applicant.

Based on these and other inferences an applicant will be selected or rejected for a job position.

Which statement best describes the scenario presented above?

- ○ The selection of candidates is based on inferences by a company's employees, who assess the portraits of applicants.
- ○ Candidates will be selected based on inferences by an artificial intelligence on the applicant's profile picture.

Fig. 1. Vignette description of the hypothetical advertising scenario a) and hiring scenario b).

2 PRIMARY TASK

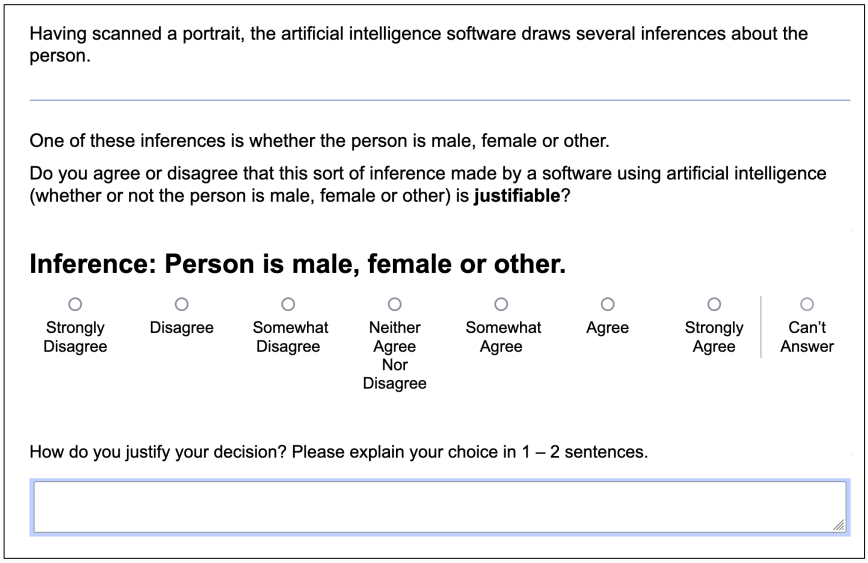


Fig. 2. Example interface of the primary rating task and the prompt to provide a written response. Example does not show treatment with the presentation of a definition of the evaluative term.

3 GENERIC DEFINITIONS OF EVALUATIVE TERMS

Table 1. Generic definitions of the six evaluative adjectives presented to half of the participants. All definitions were based on the Cambridge Dictionary, some formulations were slightly adapted to fit our context.

inference	definition
reasonable	What do we mean by **reasonable**? Something is reasonable if it's based on good sense and/or in accordance with reason.
fair	What do we mean by **fair**? Something is fair if it's based on equality without favoritism or discrimination.
justifiable	What do we mean by **justifiable**? Something is justifiable if it can be marked by a good or legitimate reason.
responsible	What do we mean by **responsible**? Something is responsible if it can answer for its conduct and obligations.
appropriate	What do we mean by **appropriate**? Something is appropriate if it's suitable or compatible in the circumstances.
acceptable	What do we mean by **acceptable**? Something is acceptable if it can be agreed on and is worthy of being accepted.

4 DATA CLEANING

The data was cleaned based on the criteria presented in Table 2, which gives an overview on the measures taken and a count of identified cases per measure. The SoSci Survey online survey tool provides a relative speed index (RSI) that identifies fast responding participants. This index indicates how much faster a participant has completed the experiment than the typical participant (median). As recommended by SoSci, all respondents with an RSI >= 2 (n = 418) are removed. All samples with duration time between 2 minutes and 4 minutes, cases that rated all inferences with the same rating, and cases with a RSI value above 1.75 were manually checked. Cases identified as problematical were discussed with a second researcher and removed in case of agreement.

Table 2. Summary of measures to clean data and number of removed cases

description	removed cases	N
Original N		4752
Time_RSI > 2	418	4334
< 18 years old	1	4333
Attention Check AD	245	4088
Attention Check HR	208	3880
Duration < 120	0	3880
Duration > 120 & < 240	9	3871
Straightliners	52	3819
TIME_RSI > 1.75 & < 2	67	3752
Double Turkers	4	3748
Nonsense Samples	3	3745

5 TWO-SIDED WELCH TWO-SAMPLE T-TEST

Participants rated the inferences gender (*mean* AD=2.66, *mean* HR=3.82; t(3513.1)=-18.536; P<0.001; 95% *CI*: (-1.28, -1.04); d=0.62), skin color (*mean* AD=2.88, mean HR=4.19; t(3513.1)=-18.536; P<0.001; 95% *CI*: (-1.44, -1.17); d=0.61), emotion expression (*mean* AD=2.97, *mean* HR=3.62; t(3654.7)=-11.079; P<0.001; 95% *CI*: (-0.75, -0.52); d=0.36), and wearing glasses (*mean* AD=2.03, *mean* HR=3.16; t(3147.2)=-18.082; P<0.001; 95% *CI*: (-1.26, -1.01); d=0.59) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario.

Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent (*mean* AD=5.25, *mean* HR=5.34; t(3662.2)=-1.425; P=1; 95% *CI*: (-0.21, 0.03); d=0.05), trustworthy (*mean* AD=5.29, *mean* HR=5.18; t(3637.5) = 1.685; P=0.74; 95% *CI*: (-0.02, 0.23); d=0.06), and likable (*mean* AD=5.04, *mean* HR=5.16; t(3695.7)=-2.059; P=0.32; 95% *CI*: (-0.24, -0.006); d=0.06) did not show a significant difference between the two scenarios. Only ratings for the inference assertive (*mean* AD=4.69, *mean* HR=4.89; t(3668.3) = -3.219; P=0.01; 95% *CI*: (-0.32, -0.078); d=0.11) were significantly different between the two scenarios, but the effect was negligible.

6 EXPLORATORY FACTOR ANALYSIS (EFA)

Prior to the computation of the exploratory factor analysis (EFA), several assumptions were tested.

6.1 Assumptions

Missing Data for Inference Ratings. Missing values appeared to be random and were less than 2% per variable (max. n=71 for the variable *assertive*, accounting for 1.9%; min n=31 for the variable *wearing glasses*, accounting for 0.83%). For EFA, all samples with missing values for the inference ratings were removed (in total 208). The sample size was reduced to 3537.

Normality and Linearity. Table 3 lists statistics for each of the dependent inference variables, including skewness and kurtosis. The deviations from normal skewness and kurtosis are within an acceptable range. Additionally, given the large sample size, the impact of departures from normal skewness and kurtosis is negligible.

Table 3. Statistics for each dependent variable

	mean	sd	median	trimmed	skew	kurtosis	se
gender	3.26	1.96	3.00	3.07	0.68	-0.80	0.03
emotion expression	3.30	1.80	3.00	3.16	0.67	-0.64	0.03
wearing glasses	2.59	2.00	2.00	2.26	1.13	-0.12	0.03
skin color	3.53	2.25	3.00	3.41	0.46	-1.36	0.04
intelligent	5.32	1.92	6.00	5.58	-0.95	-0.46	0.03
trustworthy	5.25	1.93	6.00	5.52	-0.95	-0.44	0.03
assertive	4.80	1.88	5.00	4.94	-0.46	-1.06	0.03
likable	5.12	1.85	6.00	5.33	-0.73	-0.72	0.03

Absence of Multicollinearity and Singularity. None of the correlation coefficients displayed in Fig. 2 of the main article are greater than .8. This suggested there is no multicollinearity or singularity. Additionally, the determinant of the R-matrix was 0.031 and greater than the heuristic of 0.00001. [2, p. 771]

Factorability of the Correlation Matrix. The correlation coefficient matrix in Fig. 2 of the main article displayed several correlations above .3. An alternative measure is the Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy [6]. A factor analysis is said to yield reliable and distinct factors, if values are close to 1, which suggests that correlation patterns are relatively compact [2, p. 769]. We used the KMO criteria based on [5]. The KMO values for all inference ratings were above .71 and fell within the range of middling values. The overall MSA value was .82, falling in the range of meritorious values [4, 6].

6.2 Number of Factors

Given the result from the parallel analysis and scree plot in Fig. 3 and other criteria such as the Velicer's MAP test, Very Simple Structure test of complexity 1, and Kaiser's criterion, first a two-factor solution was computed and compared to the results of a three-factor solution and a four-factor solution.

6.3 Test Specifications

It was reasonable to assume that the constructs underlying the measured dependent variables correlated, because we measured the agreement to inferences made from the facial region. Therefore, we first applied oblimin as oblique

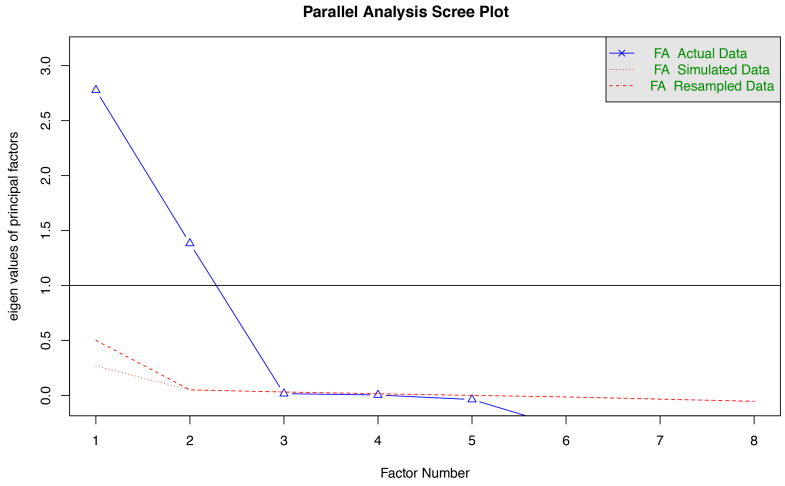


Fig. 3. Graphical analysis for the number of factors using parallel analysis scree plot.

rotation and estimated factor scores using tenBerge for preserving correlations. Supporting this decision, [1, 2] points out that in practice there are many reasons to believe that orthogonal rotation is not appropriate for data involving people, because any construct of psychological nature is correlated in some way with another psychological construct. However, for two factors, oblique rotation resulted in two factors with no correlation. This indicates that the two factors were independent. For correlations of factors below 0.32, [7] suggest orthogonal rotation. Therefore, we applied varimax for orthogonal rotation. Minimum residual (minres) was retained as factoring method, because multivariate normality does not have to be assumed [8]. Factor scores were estimated using regression. To compute the exploratory factor analysis, the R psych package and the GPArotation package were used.

6.4 Factor analysis model with 2 factors

Fig. 4 a) displays the structure of the factor analysis with two factors and indicates the rounded loadings. MR1 represents the first factor labeled *second-order inferences* and MR2 the second factor labeled *first-order inferences*. Fig. 4 b) is a graphical representation of the item's grouping based on their loadings on both of the factors.

There were no residuals > 0.05. The root-mean-square residual was 0.014. The residuals appeared to be approximately normally distributed. Regarding the factor scores, no outliers were identified.

We validated the results by randomly splitting the data in half and running the factor analysis on both subsets. This procedure was repeated three times. For each validation procedure, both factor analyses on the two subsets of the data set resulted in the variables having the same patterns of the factor loadings as with the complete sample. Additionally, the communalities were similar. This validated the factor solution previously obtained on the full dataset.

Both sub-scales had high reliability, the overall α is 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences*.

Table 4 displays all solutions with two, three and four factors.

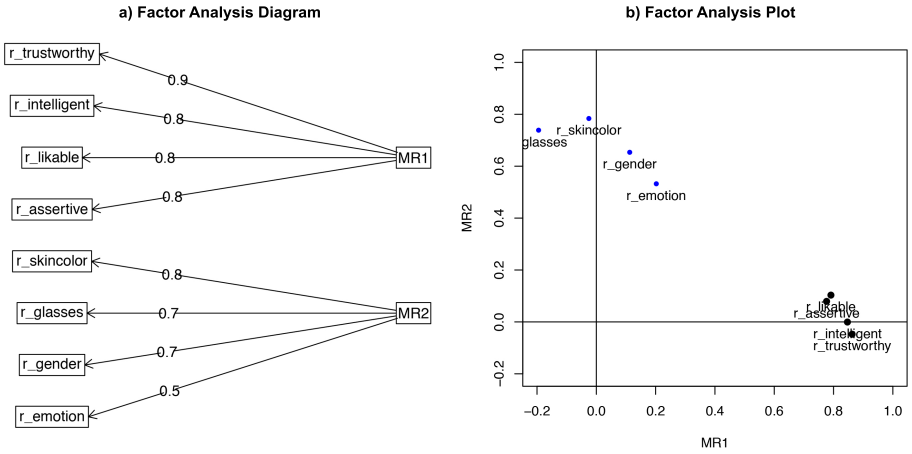


Fig. 4. Summary of two-factor solution with factor diagram and factor plots.

Table 4. Overview of Exploratory Factor Analysis Solutions with 2, 3 and 4 Factors.

	Two Factors		Three Factors			Four Factors			
	MR1	MR2	MR1	MR2	MR3	MR1	MR2	MR3	MR4
gender	0.11	**0.65**	0.14	**0.65**	0.01	0.07	**0.66**	-0.01	0.09
emotion expression	0.20	**0.53**	0.08	0.09	**0.62**	0.01	-0.00	**1.00**	-0.00
wearing glasses	-0.19	**0.74**	-0.21	**0.60**	0.17	-0.19	**0.67**	0.07	0.01
skin color	-0.03	**0.78**	0.01	**0.83**	-0.03	0.06	**0.82**	-0.01	-0.05
intelligent	**0.85**	-0.00	**0.87**	0.05	-0.08	**0.86**	0.01	-0.02	0.00
trustworthy	**0.86**	-0.05	**0.87**	-0.04	-0.03	**0.87**	-0.05	0.00	-0.00
assertive	**0.78**	0.08	**0.75**	-0.04	0.14	0.01	-0.00	-0.00	**0.99**
likable	**0.79**	0.10	**0.77**	0.03	0.08	**0.73**	0.06	0.05	0.06
eigenvalues	2.78	1.89	2.73	1.48	0.45	2.07	1.57	1.00	1.00
proportion variance	0.35	0.24	0.34	0.17	0.06	0.26	0.20	0.13	0.13
cumulative variance	0.35	0.58	0.34	0.53	0.58	0.26	0.46	0.58	0.71
α	0.89	0.77	0.89	–	0.76	0.87	0.76	–	–

6.5 Factor analysis for 3 and 4 factor solutions

The factor analyses with three and four factors resulted in one and two factors with only one indicator variable respectively (see Table 4). This is opposed to the general idea of a factor analysis identifying latent constructs by forming factors out of a combination of at least two variables [3]. Additionally, for the three-factor solution, the cumulative variance was equal to the cumulative variance for a two-factor solution. The third factor had an eigenvalue of < 1. The composition of the three factors was not robust when computing the factor analysis on randomly sampled subsets of the complete data. While the cumulative variance explained by a factor analysis for four factors was the greatest among all tested factor analysis models, this solution was also not robust. Running the factor analysis on two randomly sampled subsets resulted in different patterns of the loadings on the factors. Altering the random sampling produced different patterns of loadings once again.

Although the fit based upon off diagonal values equaled 1 in each of the models, the solutions with three and four factors were neither appropriate in terms of variables per factor nor robust across subsets of the data. Hence, exploratory factor analysis of the eight items measured in this study revealed that two factors were sufficient to explain the underlying structure of common inferences from faces.

6.6 Distribution of EFA factor scores and original ratings

The global means for all variables that load on the first factor and all variables that load on the second factor are highlighted by the horizontal lines in Fig. 5 a) and b). The bold lines in panels a) and b) indicate the means for the individual groups. By using the factor scores as dependent variables for further analysis, the interpretation of the dependent variables depicted in panels c) and d) changes compared to the original inference ratings. A factor score of approximately 0 indicates that a participant's mean rating of all variables that load on this factor is close to the global mean of these variables (horizontal lines in panels a) and b)). A negative factor score indicates this subject gave lower than average ratings. A factor score close to 1 indicates that the subject's ratings for the variables loading on this specific factor are about one standard deviation above the average rating.

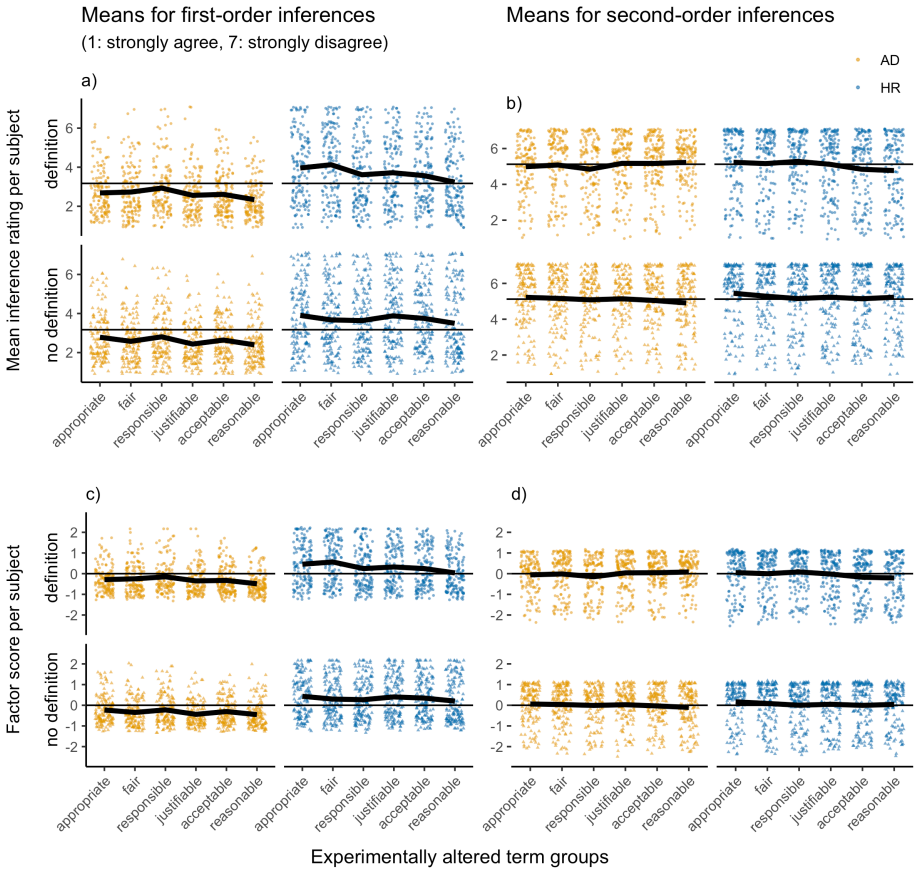


Fig. 5. Distribution of participants' ratings and distribution of the factor scores extracted from the exploratory factor analysis.

7 MANOVA

We performed a multi-factorial MANOVA to statistically test the differences in group means. The two factors identified by performing exploratory factor analysis served as dependent variables. We included three experimentally altered independent variables (context, adjective terms, definition), all measured control variables (AI knowledge, gender, age, education and occupation) and the main justification types for first-order and second-order inferences from the classification. All predictors were included as categorical variables. For the MANOVA and ANOVA analysis, the R car package was used.

7.1 Assumption tests and fitting the model

Assumption tests prior to fitting the model

Although the exploratory factor analysis produced uncorrelated factor scores, we first computed a MANOVA to obtain an overview of patterns between first-order ratings and second-order ratings as dependent variables. Given the lack of correlation and thus no further information from the correlation structure of the dependent variables, we expected a diffused structure of results. Running the MANOVA based on factor scores from the factor analysis with oblique rotation did not change the results. Nine further cases with missing data, i.e., no justification provided for their ratings, were additionally removed.

The following assumptions were tested prior to computing the MANOVA. **Adequate Sample Size**. We applied the one-in-ten-rule for adequate sample size. Our sample size of 3,528 with at least 133 subjects per group based on the experimentally altered independent variables exceeded the threshold of 100 subjects (ten times the number of independent variables: Context, Adjective Terms, Definition, AI Knowledge, Age, Gender, Education, Occupation, Main Justification First-Order, Main Justification Second-Order). **Independent Observations**. Given the randomization, all observations were independent. **Outliers Based on Raw Data**. Neither univariate extreme outliers based on the boxplot method with observations being three interquartile ranges far from the first or third quartile nor multivariate outliers based on Mahalanobis distance were identified. **No Multicollinearity**. There was no multicollinearity.

Model Fitting 1: Testing for Interaction Effects

To test the other assumptions based on residual analysis, we fitted a model with interaction terms first. There were no significant interaction effects. All partial η^2 were calculated using the etasq function from the R heplots package.

Model Fitting 2: Residual Analyses

Because none of the interaction effects were significant at α =0.01, they were removed and a new model without interaction effects was fitted. Residual analyses were conducted on the linear model of this MANOVA.

The following assumptions were tested after fitting the MANOVA. **Linearity of Data**. The residuals vs. fitted values plot indicates that the linearity assumption is met. The line is approximately horizontal at zero. **Homogeneity of Variances of Residuals**. The spread-location plot shows that the residuals have an equal variance above and below the line, which is approximately horizontal across the plot. This indicates that the spread of the residuals is approximately equal at all fitted values and that the assumption of homoscedasticity is satisfied. **Normality of Residuals**. The histogram of residuals indicates that the residuals are approximately normally distributed. However, in the Q-Q plot of residuals, the points in the lower left and upper right corner of the plot deviate somewhat from the reference line. A further analysis of outliers and influential cases could help identify cases that might cause the deviations.

Observations having extreme residuals (> 3.5, < -3.5), extreme Cook's Distance values (> 0.0056), extreme hat values (> 0.062, < -0.062), or extreme dffits values (> 0.5, < -0.5) were identified and inspected. These thresholds are based on graphical analysis and are all less strict than common thresholds such as the > 2(p+1)/n for hat values (with p being the number of predictors and n the sample size). Model results for the removal of varying sets of outliers and influential cases were compared. Finally, 36 cases having either extreme residuals (> 3.5, < -3.5) or extreme Cook's Distance values (> 0.0057) were removed. Removing more of the previously identified cases did not improve the results.

Model Fitting 3: Final Multivariate Assumption Check

Table 5 presents the output for the model after removing the identified 36 cases. Significant effects are highlighted in bold. The panels in Fig. 6 indicate that linearity of data, homogeneity of variances of residuals as well as normality of residuals are now met.

Table 5. Final MANOVA without interaction effects and with outliers and influential cases removed

	Df	test stat	approx F	num Df	den Df	Pr(>F)	Bonferroni	partial η^2
(Intercept)	1	0.01	21.43	2	3445	0.000	0.000	0.012
first-order justification	6	0.50	190.76	12	6892	0.000	**0.000**	**0.249**
second-order justification	6	0.45	164.60	12	6892	0.000	**0.000**	**0.223**
AI knowledge	4	0.03	13.43	8	6892	0.000	**0.000**	**0.015**
age	5	0.01	4.50	10	6892	0.000	0.000	0.006
gender	2	0.00	1.97	4	6892	0.097	1.000	0.001
occupation	8	0.01	2.38	16	6892	0.001	0.016	0.006
education	7	0.00	0.94	14	6892	0.519	1.000	0.002
context	1	0.04	73.68	2	3445	0.000	**0.000**	**0.041**
terms	5	0.01	3.58	10	6892	0.000	0.001	0.005
definition	1	0.00	0.61	2	3445	0.543	1.000	0.000

Comparison of final model with model based on an equalized dataset

The results of the final model from Table 5 were compared to the results of a model for an equalized dataset based on the three experimentally altered independent variables (context, adjective terms, definition). The same outliers and influential cases as in the previous model were removed. After equalization, this dataset contained 3,168 subjects. Because the assumptions based on the graphical analysis did not differ and the results were similar to the previous results of Table 5, this model was discarded in favor of retaining more observations in a sample without equalized groups.

7.2 Follow-up analysis

To identify which individual predictors had a significant effect on which dependent variable, we conducted univariate analyses.

Univariate Analysis: ANOVA for First-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of 13 identified extreme

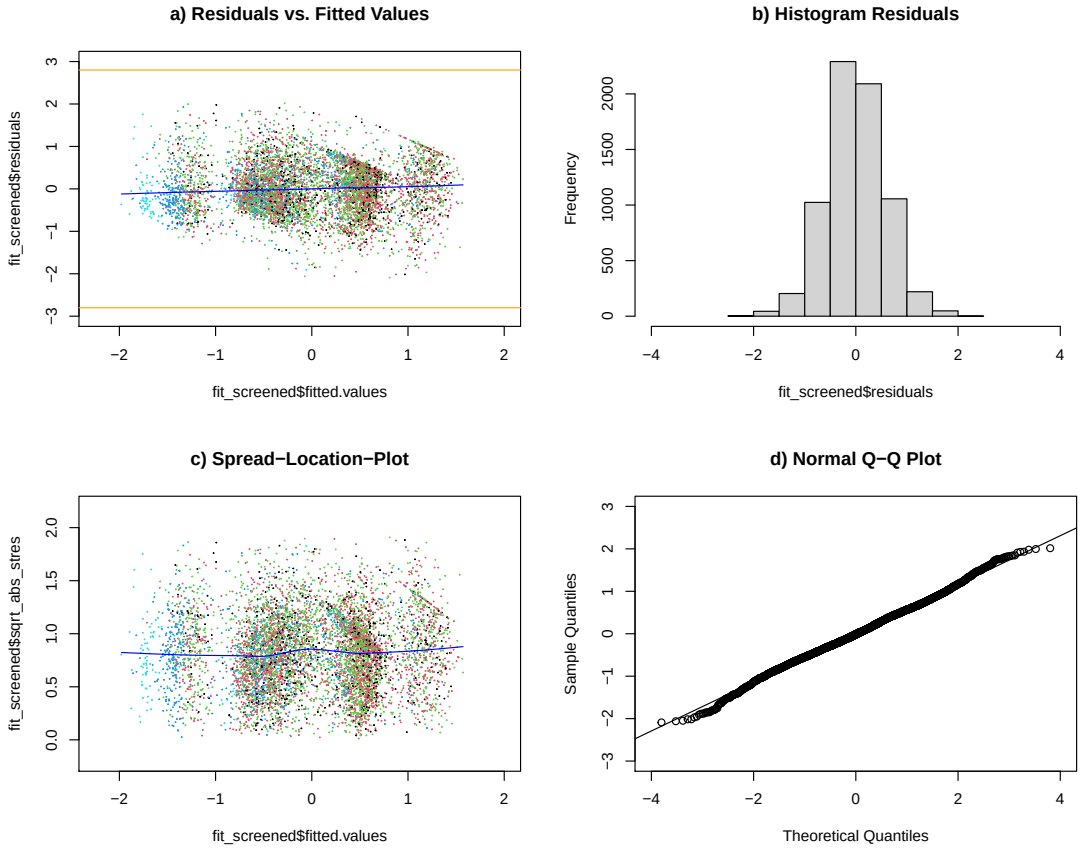


Fig. 6. Graphical analysis of MANOVA test assumptions after removing 36 identified cases.

outliers and influential cases did not improve the homogeneity of variances. To control for the family-wise error rate, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. Additionally, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

Univariate Analysis: ANOVA for Second-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of twelve extreme outliers and influential cases did not improve homogeneity of variances. As we did for the ANOVA for the first-order dependent variable, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. In addition, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

7.3 Pairwise comparisons

For first-order inferences, pairwise comparisons for the variable *adjective terms* and the significant experimental variable *context* based on estimated marginal means revealed significant group differences between the advertisement and the hiring context at each level of the variable *adjective terms* (see Table 6, rows 1-6). These differences could not be observed for second-order inferences. All groups differed significantly between first-order and second-order inferences (see Table 6, rows 7-18). These results are in line with the rating behavior depicted in Fig. 5 and the ANOVA results (see Appendix 7.2 and for ANOVA outputs Table 1 of the main text), i.e., the assignment to a context, either advertisement or hiring, had a significant effect on the rating behaviors of participants for first-order inferences. Also, the rating behaviors on first- and second-order inferences within one context differed significantly.

Table 6. All significant pairwise tests for context and adjective terms based on estimated marginal means for the complete model

terms	variety	context	contrast	estimate	SE	df	t.ratio	p.value
acceptable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
appropriate	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
fair	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
justifiable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
reasonable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
responsible	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
acceptable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.08	0.00
acceptable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.89	0.00
appropriate	.	AD	factor2nd - factor1st	-0.54	0.11	3454.00	-5.06	0.00
appropriate	.	HR	factor2nd - factor1st	-0.74	0.11	3454.00	-6.88	0.00
fair	.	AD	factor2nd - factor1st	-0.64	0.11	3454.00	-5.87	0.00
fair	.	HR	factor2nd - factor1st	-0.84	0.11	3454.00	-7.67	0.00
justifiable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.01	0.00
justifiable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.81	0.00
reasonable	.	AD	factor2nd - factor1st	-0.58	0.11	3454.00	-5.30	0.00
reasonable	.	HR	factor2nd - factor1st	-0.78	0.11	3454.00	-7.12	0.00
responsible	.	AD	factor2nd - factor1st	-0.71	0.11	3454.00	-6.55	0.00
responsible	.	HR	factor2nd - factor1st	-0.91	0.11	3454.00	-8.36	0.00

The influence of the *justification* variables becomes apparent when computing estimated marginal means for a model without the *justification* variables. When controlling for the *justifications*, the effect of the variable *context* decreases. Nevertheless, the same significant differences of main interest are identified between the AD and HR context.

8 SUBJECTS' JUSTIFICATIONS

8.1 Documentation of category classes and F1 scores

Table 7. Generated category classes for participants' justifications, together with example comments of classified observations per class and test set F-1 score for each class.

	Category classes	Examples	F1 score
1	AI can tell	"You should be able to determine the race of a person with a picture of their face."	0.94
2	AI cannot tell	"You can not tell if a person is likable or not in a photo."	0.96
3	Inference is relevant for the decision making	"Some positions require emotion, or at least sympathy or empathy."	0.96
4	Inference is not relevant for the decision making	"it does not matter if a person is black or white when the AI is recommending products and services"	0.95
5	Inference creates harm (e.g., illegal, discrimination).	"This is unacceptable, as it may be discriminatory against the transgender population."	0.97
6	AI has human biases	"Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.."	0.97
7	Incomprehensible & nonsensical responses	"this person is not fully trustworthy", "Not very like"	0.95

8.2 Categories

Table 8 defines all categories, provides application descriptions, and differentiates the category to related ones. More examples comments are provided.

Table 8. Definition of categories and examples (Code book).

Category	Description	Example
AI can tell (e.g. "easy to tell")	**Definition:** The AI/software is able to/can make an inference because the portrait image provides sufficient evidence for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are suitable/good/sufficient for the AI to make the inference.	*Very easy to tell. All you need is a picture and a database.* (P635/2575)
		Can always tell this from a color pic. (P1329/4565)
	Application: The category is assigned when someone *agrees* that an AI is able to make the inference based on sufficient evidence. Sometimes a *specific reference* to the photograph, portrait, image, picture, or visual data type is made. The word "obvious" can be an indicator to use this category.	*AI can determine this easily. It can see if you wear glasses or not.* (P557/2327)
		Also extremely obvious and superficial. (P1257/4338)
AI cannot tell (e.g. "not easy to tell")	**Definition:** The AI/software is not able to/cannot make an inference because the evidence in the portrait image is insufficient for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are not suitable/good/sufficient for the AI to make the inference.	*AI cannot determine whether a person is trustworthy or not.* (P333/1605)
		Intelligence is not a physical trait and cannot be determined from a photograph by an AI. (P220/1207)
	Application: The category is assigned when someone *disagrees* that an AI is able to make the inference. In some cases, it is *specifically highlighted* that a facial image or visual data type is not correct/insufficient to make a certain inference.	*You cannot determine whether someone is intelligent based on the way that they look.* (P1362/4610)
Inference is relevant for the decision making	**Definition:** The inference is relevant/important and/or useful for the purpose of application.	*[...] this piece of information is needed for better predictions.* (P260/1339)
	Application: This category is assigned if someone explains why/that a certain inference is relevant for making a decision for a specific application.	*[...] I think having emotions is a crucial part of an interview.* (P3515/5661)
Inference is not relevant for the decision making	**Definition:** The inference is not relevant/important/appropriate and/or not useful for the purpose of the application.	*It does not matter whether a person is assertive or not.* (P46/550)
	Application: This category is assigned if someone explains why/that a certain inference is not relevant for making a decision for a specific application.	*A sex does not define a person.* (P1109/3856)

Inference creates harm (e.g. illegal, discrimination)	**Definition:** An AI inference is considered discriminatory and/or violates personal rights. **Application:** This category is assigned when drawing an inference would lead to a discriminatory outcome or harm a person in any other way.	*this form of racism should be unacceptable. you cannot infer such a thing on skin color alone.* (610/2491) *Trying to determine a user's personality and trustworthiness is a pretty massive breach of privacy.* (P133/894)
AI has human bias	**Definition:** Inference is affected by human bias; the inference cannot be made without human bias. **Application:** This category is assigned if someone highlights the dependency of AI on humans and hence the implicit integration of human bias, for example, into the data and ultimately into the decision made by an AI.	*I do not see how an AI could make such a determination without relying on human biases to be programmed into it. [...]* (P1862/1966) *Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.* (P1708/1272)
Incomprehensible responses	**Definition:** The comment is unrelated to the task and/or contains text copied from the instructions or nonsensical text. **Application:** This category is assigned if the comment is not a justification for the rating. Additionally, this category is applied if it becomes apparent from the comments that a participant did not understand the task. If one comment of a respondent can clearly be assigned to this category, all comments by this same respondent have to be assigned to this category, because it cannot be assumed that the person trustfully filled out the questionnaire.	*ok a so like in* (P1419/4830) *they are intelligent* (P607/2486) *I agree that person is or is not wearing glasses. because it is useful to portrait a person.* (P928/3352)

8.3 Justifications results for the "Glasses" inference

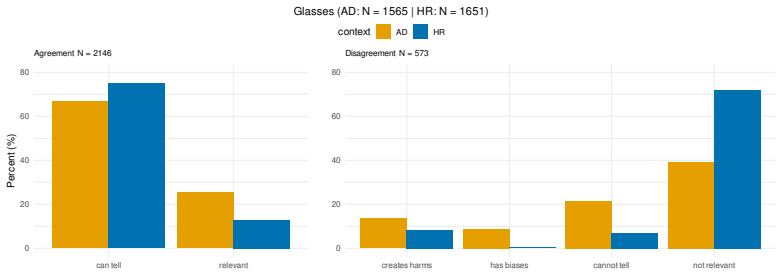


Fig. 7. Justifications results for the "Glasses" inference.

REFERENCES

[1] Anna B. Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation* 10, 1 (2005), 7. https://doi.org/10.7275/jyj1-4868

[2] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R.* Sage.

[3] Robin K. Henson and J. Kyle Roberts. 2006. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement* 66, 3 (2006), 393–416. https://doi.org/10.1177/0013164405282485

[4] Matt C. Howard. 2016. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction* 32, 1 (2016), 51–62. https://doi.org/10.1080/10447318.2015.1087664

[5] Graeme D. Hutcheson and Nick Sofroniou. 1999. *The multivariate social scientist: Introductory statistics using generalized linear models.* Sage.

[6] Henry F. Kaiser. 1970. A second generation little jiffy. *Psychometrika* 35, 4 (1970), 401–415. https://doi.org/10.1007/BF02291817

[7] Barbara G. Tabachnick and Linda S. Fidell. 2013. *Using multivariate statistics.* Pearson Education.

[8] Conrad Zygmont and Mario R. Smith. 2014. Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology* 10, 1 (2014), 40–55. https://doi.org/10.20982/tqmp.10.1.p040

AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI

Chiara Ullstein*
chiara.ullstein@tum.de
Technical University of Munich,
Chair of Cyber Trust
Munich, Germany

Severin Engelmann*
severin.engelmann@tum.de
Technical University of Munich,
Chair of Cyber Trust
Munich, Germany

Orestis Papakyriakopoulos†
orestis@princeton.edu
Princeton University, Center for
Information Technology Policy
Princeton, USA

Michel Hohendanner‡
michel.hohendanner@hm.edu
University of Applied Sciences
Munich, Center for Digital Sciences
and AI & Faculty of Design
Munich, Germany

Jens Grossklags
jens.grossklags@in.tum.de
Technical University of Munich,
Chair of Cyber Trust
Munich, Germany

ABSTRACT

Recent advances in computer vision analysis have led to a debate about the kinds of conclusions artificial intelligence (AI) should make about people based on their faces. Some scholars have argued for supposedly "common sense" facial inferences that can be reliably drawn from faces using AI. Other scholars have raised concerns about an automated version of "physiognomic practices" that facial analysis AI could entail. We contribute to this multidisciplinary discussion by exploring how individuals with AI competence and laypeople evaluate facial analysis AI inference-making. Ethical considerations of both groups should inform the design of ethical computer vision AI. In a two-scenario vignette study, we explore how ethical evaluations of both groups differ across a low-stake advertisement and a high-stake hiring context. Next to a statistical analysis of AI inference ratings, we apply a mixed methods approach to evaluate the justification themes identified by a qualitative content analysis of participants' 2768 justifications. We find that people with AI competence (N=122) and laypeople (N=122; validation N=102) share many ethical perceptions about facial analysis AI. The application context has an effect on how AI inference-making from faces is perceived. While differences in AI competence did not have an effect on inference ratings, specific differences were observable for the ethical justifications. A validation laypeople dataset confirms these results. Our work offers a participatory AI ethics approach to the ongoing policy discussions on the normative dimensions and implications of computer vision AI. Our research seeks to inform, challenge, and complement conceptual and theoretical perspectives on computer vision AI ethics.

*Denotes equal contribution.
†Currently affiliated with Sony AI, Switzerland.
‡Also affiliated with University of Wuppertal, Industrial Design, Germany.

EAAMO '22, October 6–9, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10.
https://doi.org/10.1145/3551624.3555294

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Applied computing** → *Law, social and behavioral sciences*; • **Security and privacy** → **Human and societal aspects of security and privacy**; **Social aspects of security and privacy**.

KEYWORDS

artificial intelligence, computer vision, human faces, ethics, public participation

ACM Reference Format:
Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Michel Hohendanner, and Jens Grossklags. 2022. AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551624.3555294

1 INTRODUCTION

Companies and research institutes increasingly produce and release artificial intelligence (AI) applications that draw conclusions about individuals from human faces [22, 33, 34]. One task of such facial processing technologies is facial analysis (hereafter called *facial analysis AI*), which classifies facial characteristics as demographic or physical traits [82] and even personality traits from portrait images. Driven by scientific advances in the areas of face-based inferences on intelligence, trustworthiness, likability and other personality traits [1, 5, 98], as well as sexual orientation [63, 96], such AI products find application in various domains including human resources and advertising. In response, a community of critical data scientists has raised ethical concerns regarding the development of such facial analysis AI [e.g., 24, 69, 70, 81, 85].

In policy-making, researchers from various disciplines have argued that the veracity of inferences from faces is not significant enough to counterbalance negative consequences [10], and have pointed out the unreliability of human inferences from faces, such as trustworthiness or intelligence [94, 95]. Others have highlighted the variability and context-dependency of emotions depicted in pictures and videos showing faces [6]. Members of the European

Parliament recently called "for a ban on the use of private facial recognition databases" [32]. Moreover, serious misclassifications have been uncovered in commercial gender detection tools [12] and job candidate selection software [80, 90]. Nonetheless, many industry actors see an enormous market potential – the AI emotion recognition industry alone is predicted to become worth multiple billion dollars in the coming years [23].

Fundamental questions are how to draw a line between ethically permissible and impermissible AI facial inferences as well as who should be involved in making these decisions. These two questions are central to understand how AI systems and their regulatory frameworks can be developed in a socially-sustainable manner. We contribute to this research debate by exploring how laypeople and individuals with AI competence evaluate facial analysis AI inference-making. We believe that both groups, potential future designers of AI systems and subjects of facial analysis AI, should play a more critical role in the development of ethical computer vision AI.

Prior work has illustrated that the general population (i.e., laypeople) may be aware that facial analysis AI applications exist but that it has little knowledge of their technological characteristics [14]. Mainstream media and science fiction contribute to the propagation of AI narratives that create unrealistic expectations of AI capabilities [13–15, 17, 20, 35, 43], and pay little attention to their feasibility [66]. Hopes and fears are part of AI narratives [17] and although some argue that current perceptions are skewed or extreme [15] such perceptions can influence the acceptance and adoption of AI systems by the general public [13, 15, 17, 35, 43, 66]. How popular narratives on technology, including the role of AI, can influence the imagination of future societies has, for instance, been explored using research through design and narrative analysis [e.g., 16, 44].

It has become increasingly clear that challenges arising from AI systems do not have purely technical solutions. For example, the decision to use one fairness metric over another requires value judgments that cannot be solved by formalistic approaches. Normative decisions *always* attract support, skepticism or rejection by different groups in society. Achieving consensus on topics such as "algorithmic fairness will be difficult unless we understand why people disagree in the first place" [77, p.1]. In the context of facial analysis AI, we believe it is important to understand how individuals with AI competence perceive AI inference-making and how their perception differs from the perception of AI inference-making by laypeople. Overall, we ask the following research question:

How do ethical justifications of AI inference-making from faces differ between individuals with AI competence and laypeople?

We build this research on our prior work in which we explored a conceptualization of reasonable inference [30] and asked laypeople how they evaluate such inferences [31]. In this study, we extend this work and compare evaluations of AI inference-making of laypeople with those of individuals with AI competence. We first survey researchers and students studying AI or computer vision AI (N=122) for our sample of "individuals with AI competence". We then compare their ratings and open-text justifications to a laypeople dataset (N=122). Furthermore, we analyze whether a range of demographic factors correlates with differences in the ethical evaluation of AI inference-making from portrait pictures. We confirm the results using a validation laypeople dataset.

2 RELATED WORK

2.1 Research on AI inferences of social constructs and character traits from faces

Many companies have developed facial analysis products used for market research, customer targeting, health care or education. For instance, Face++ sells services that infer "face related attributes including age, gender, smile intensity, [...] emotion, beauty" [33]. EmoVu [29] and FaceReader by Noldus perform facial expression analysis and infer, amongst others, personal characteristics and the six basic emotions [28] "happy, sad, angry, surprised, scared, and disgusted" [73]. Betaface and SkyBiometry classify glasses, beard, mustache, mood, or ethnicity [8, 9]. Faception claims to be able to identify people with high IQ [34].

The foundation for these analyses stems from research on inferences from human faces by humans. Research in evolutionary anthropology and psychology presents findings that humans "cannot help" but form first facial impressions despite their proven inaccuracy [10, 27, 74, 92, 93]. In the past, organizational and institutional physiognomic practices relied on making inferences about character traits from visual appearance [25, 39, 82, 88, 89]. Well-known for their contributions to physiognomy, Francis Galton, Caspar Lavatar or Cesare Lombroso, amongst others, developed taxonomies of character interpretations and corresponding facial configurations (see [92] for physiognomy's history). Today, a line of research persists that advocates the accuracy of first facial impressions [47, 54, 71, 76]. Research in computer vision datasets, algorithms, and models is clearly aware of this line of research. Projects in computer vision AI have asserted to successfully infer sexual [63, 96] and political orientation [57, 97] or emotion intensity and emotion expression [7, 26] based on people's faces in images. Others claim to be able to infer a variety of latent traits in personality assessment, such as trustworthiness [98] or the big 5 personality traits [18, 36–38, 64, 78, 86, 87] from profile images. However, considerable evidence suggests that first facial impressions do not surpass a "kernel of truth" [10, 74, 75, 92, 93, 95].

Researchers in the field of critical data science highlight ethical concerns arising from classifying individuals with AI on the basis of their facial appearance. Image-based inferences about people can only represent visibly apparent factors of an inferred concept [42]. However, as such inferences are used today, they may be based on bold or questionable semiotic assumptions when predicting intentions, aims, and capabilities or characters of individuals based on their facial characteristics found in portrait images [25, 52]. Judgments of this kind are epistemologically unreliable [30, 90]. Some researchers have argued that such systems are morally objectionable because they treat individuals as categorized objects [42, 53], and others have proposed to abolish physiognomic AI [90].

2.2 Does knowledge of AI correlate with ethical perceptions of AI?

While prior research has investigated users' perceptions of AI-based systems, only a handful of research studies exist that investigate experts' ethical perceptions of AI systems [49, 77, 100]. Here, measuring AI knowledge has proven to be difficult. Approaches vary

from attempts to identify actual AI knowledge over the recruitment of specific subject pools to measures involving programming and numeracy skills (see Appendix A.1 for an overview). Another difficulty in comparing the studies arises from the diversity of application contexts and the diversity of AI systems, e.g., "automated decision-making by AI" [3], "expert systems" [50], "algorithms" [65], "artificial intelligence" [99] , or "algorithmic decision-making" [49].

Some positive associations were observed: Araujo et al. [3] found that both higher levels of education and technical knowledge, including AI knowledge, have a positive association with perceived usefulness, but no significant association with perceived risk of AI decision-making. Higher technical knowledge levels show a positive association with AI fairness perceptions. Similarly, Kaufmann [50] reported that teachers with knowledge on expert systems perceive higher utility of advice from these systems compared to teachers lacking such knowledge; there was no relation between numeracy and acceptance of algorithmic advice. Logg et al. [65] found that less numerate people appreciate advice from algorithms less in the context of forecasting and estimation tasks.

In contrast, Zerfass et al. [99] found that AI expertise and perceptions on AI adoption were not related. Lee and Baykal [62] found that greater levels of computer programming knowledge decreased the perceived fairness of algorithmic decisions in the context of dividing household chores. The authors assumed that participants with higher levels of knowledge were either confronted with unexpected algorithmic decision-making results and/or had greater knowledge about the limitations of such systems. Generally, discussion-based decision outcomes were perceived as fairer than outcomes produced by algorithms. Audio-recorded interviews highlighted the importance of participation in decision-making – i.e., the ability to choose and to agree or disagree – as well as enhanced social transparency of decision outcomes via discussion of the perceptions of whether an outcome was fair or not. Logg et al. [65] observed that greater familiarity with algorithms led to less acceptance of advice from automated forecasting tasks.

Zhang et al. [100] found AI researchers to favor a prioritization of research on AI safety, to support pre-publication reviews to evaluate potential harms, to strongly disagree with AI research on lethal autonomous weapons, and, finally, to highly trust scientific and international organizations in shaping the development of AI applications for the public interest. Across three different scenarios (dynamically-priced premium of car insurance, re-routing of flight passengers, automatic loan allocation), Kasinidou et al. [49] did not find students' AI knowledge to influence ethical perceptions of AI. Instead, individual differences were observed between undergraduate and postgraduate participants. For the context of criminal justice, undergraduate computer science students changed their perceptions of algorithmic fairness after one discussion-intensive class [77]: After the intervention, students preferred adding the gender feature to the algorithms, which may be explained by weaknesses of the concept "fairness through blindness". They also preferred algorithms, as opposed to human judges, and favored algorithmic transparency as a general principle. However, consensus did not increase. Rather, opinions were more varied regarding some topics.

The literature reviewed above reveals mixed results regarding the influence of AI knowledge on AI perception. The present study contributes to this line of research by comparing how ethical perceptions of facial analysis in two different contexts vary between laypeople and individuals with AI competence.

3 STUDY PROCEDURE AND METHODS

3.1 Recruitment process and participants

We recruited 346 survey participants across three samples, one of which served validation purposes. We sampled AI-competent individuals at the end of 2021 and beginning of 2022 (N=122, female=27.05%, male=69.67%, other=3.28%). We targeted graduate and PhD students focusing on AI at two large European universities and one large European research institute via social media and news channels of computer science and data science study programs. We describe the exact filtering criteria to determine AI competence in Section 3.3 (and provide further data such as course experience in Appendix A.3.4). Each participant was compensated with a fixed payment of 5€. The mean duration was 16.31 minutes (min: 6.50, max: 32.25). The age distribution was: 46.72% with age 18-24, 49.18% with age 25-34, 2.46% with age 35-44, 0.82% with age 45-54, and 0.82% with age 55 or above (see Appendix A.4 for data cleaning).

We collected a laypeople sample at the end of 2019 and at the beginning of 2020 via Amazon Mechanical Turk (MT) in the course of another study [31]. Participation was limited to those registered in the United States. We produced a final sample of 3102 participants. For the present study, we randomly selected 122 laypeople (female=46.09%, male=48.36%, other=0%) from all participants who indicated to have either very little or novice AI knowledge (46.09% of the entire dataset). The mean duration was 9.98 minutes (min: 3.87, max: 25.08). The age distribution was: 8.20% with age 18-24, 36.07% with age 25-34, 23.77% with age 35-44, 13.93% with age 45-54, 9.02% with age 55-65, and 9.02% with age 65 or above.

We collected a validation laypeople sample in June of 2022 in a second semester undergraduate lecture at a large European university (N=102, female=18.63%, male=81.37%, other=0%). We excluded respondents with high AI competence from the sample. The mean duration was 21.88 minutes (min: 5.16, max: 37.4). We assume that the higher average duration was due to the perceived complexity of the AI knowledge quiz by participants who were not competent in AI. 99.02% were aged between 18-24, 0.98% were aged between 25-34. Survey completion was incentivized by being part of a number of voluntary tasks to become eligible for a grade bonus on the final exam. The validation dataset also allowed for a useful complementary comparison with the sample of AI-competent individuals due to their shared similarities in demographic features (gender balance, age and country of origin).

Our home institution does not require an ethics approval for questionnaire-based online studies. All participants in the dataset were informed about the procedure, the length and the basic premise of the study, and gave consent to the use of the data for research purposes. Participants could drop out at any point in the survey, or could exit the survey if they did not agree with the use of their data for research purposes. All analysis data was fully de-identified and the privacy of all subjects was preserved at all times during the study. The service used to collect the data guaranteed compliance with the European Union's General Data Protection Regulation

(GDPR). The compensation offered in the two paid studies was above minimum wage.

3.2 Vignette study

Experimental vignette studies are a common instrument to study people's perceptions and judgments in a variety of hypothetical scenarios [2, 4, 21, 40, 46, 55, 56, 68, 72]. The design of our factorial vignette study is based on our prior work [31]. It consists of two hypothetical decision scenarios: participants were either drawn into a low-stake advertisement (AD) or a high-stake hiring (HR) scenario. In both scenarios an AI system scans a portrait picture and makes a variety of inferences about an individual. Based on these and other inferences, in the AD context, a social media user will be shown a particular advertisement. In the HR context, an applicant will either be selected or rejected for a job position (see Figure 1 in Appendix A.2). Participants then rated on a 7-point Likert scale their level of agreement or disagreement (1 = "strongly agree", 7 = "strongly disagree") with eight distinct AI-made inferences from a portrait picture, drawn for the above described purpose of the application context: *gender*, *emotion expression*, *wearing glasses* and *skin color*, *intelligent*, *trustworthy*, *assertive*, and *likable*. These ratings are hereafter called *inference ratings*. After each inference rating and before proceeding to the next inference, participants were asked to justify their rating in one to two sentences.

3.3 Measuring AI competence

We developed an AI knowledge test with a total of nine questions. Four of them were directed at computer vision, out of which three were based on the computer vision textbook by Chollet [19]. The other six questions were based on an instrument designed to assess student's AI and machine learning knowledge by Rodríguez-García et al. [83]. Here, we adjusted questions for the purpose of this study and removed some items (see Appendix A.3). The AI knowledge test was first discussed with three researchers and the resulting feedback was implemented. The scale was evaluated via a pre-study with three participants, who had varying AI knowledge levels. The pre-study additionally included one question on the difficulty of each item. The pre-study illustrated that the AI knowledge test has easy, moderate and difficult questions, and was able to map out a variety of AI knowledge levels.

3.4 Mixed method analysis strategy

All analyses were performed in R and Python.

3.4.1 Content-structuring qualitative content analysis. The design of our research study followed an embedded design, which we analyzed using mixed methods by integrating qualitative and quantitative data [61, 79]. To analyze the application of justification themes, we applied content-structuring qualitative content analysis and developed a detailed category scheme to map justification patterns within the responses by participants [60, 61, 67, 79, 101]. First, one researcher labeled 15% of the two main datasets and formulated 57 detailed categories, which were discussed with a second researcher and grouped into 21 super-ordinate categories. Second, both researchers independently applied this category scheme to 10% [79] of both datasets using the instructions documented in the code book in Appendix C. The inter-coder reliability was above Krippendorff's $\alpha \geq 0.8$ for each of the inferences [59]. Differences were discussed with a third researcher. No further categories were included. Finally, one researcher labeled the entire dataset using the final category scheme. The coding occurred at the word level. This meant that as little as one word up to the entire answer could be assigned a code. Three researchers labeled the validation dataset applying the previously developed category scheme. They achieved Krippendorff's $\alpha \geq 0.7$ for each of the inferences. Differences were discussed and resolved among the three researchers.

3.4.2 Frequency and co-occurrence analysis of justification themes. We analyzed the justification themes using co-occurrence and frequency analysis. We compared the results for subgroups of the sample, e.g., AI-competent vs. laypeople, AD vs. HR context. First, the frequencies of the individual themes were analyzed independently of the co-occurrence with other themes. Second, the frequencies of all unique theme pairs, e.g., the likelihood of two themes being mentioned in combination with each other, were explored.

3.4.3 Factor analysis, Welch two-sample t-test and analysis of variances. To analyze subjects' ratings, we performed an exploratory factor analysis with orthogonal rotation (varimax), minres factor extraction and regression factor estimation for all three samples. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis [45, 48] and Barlett's test of sphericity indicated that the correlations between items were sufficiently large. For all samples, parallel analysis, BIC, the Velicer MAP and the Kaiser criterion, amongst other tests, suggested retaining two factors (see Appendix B.2 for details). Furthermore, Welch two-sample t-tests and analysis of variances (ANOVA) were computed to directly compare the inference ratings.

4 RESULTS

4.1 Inference ratings show no significant differences between AI-competent and laypeople.

4.1.1 Welch two-sample t-test results. Comparing the inference ratings of the two main samples, none of the Bonferroni-corrected Welch two-sample t-tests shows significant group differences (see Figure 1 and Appendix B.1). A robustness check of the results using Yuen's test for trimmed means confirms that there are no significant group differences. The validation laypeople dataset validates the absence of group differences for all inference ratings except for the inference *wearing glasses* ($p_{\text{Bonf.}}$=.04) in the AD context.

4.1.2 Exploratory factor analyses suggest all samples perceive the same two constructs underlying the eight inferences. Exploratory factor analyses produced the same structure of factor loadings, i.e. two factors, for all three samples. The first factor included the inferences *intelligent*, *trustworthy*, *assertive* and *likable*, which will be referred to as *character and personality traits* in the following. The second factor included the inferences *gender*, *emotion expression*, *wearing glasses* and *skin color*, which will be referred to as *social constructs and features*. Although prior tests (see Appendix B.2) proved the data to be appropriate, some factor loadings did not exceed 0.6 [41], and some of the items (e.g., *gender*) loaded on two factors [91]. We assume that this is due to our rather small sample sizes [41].

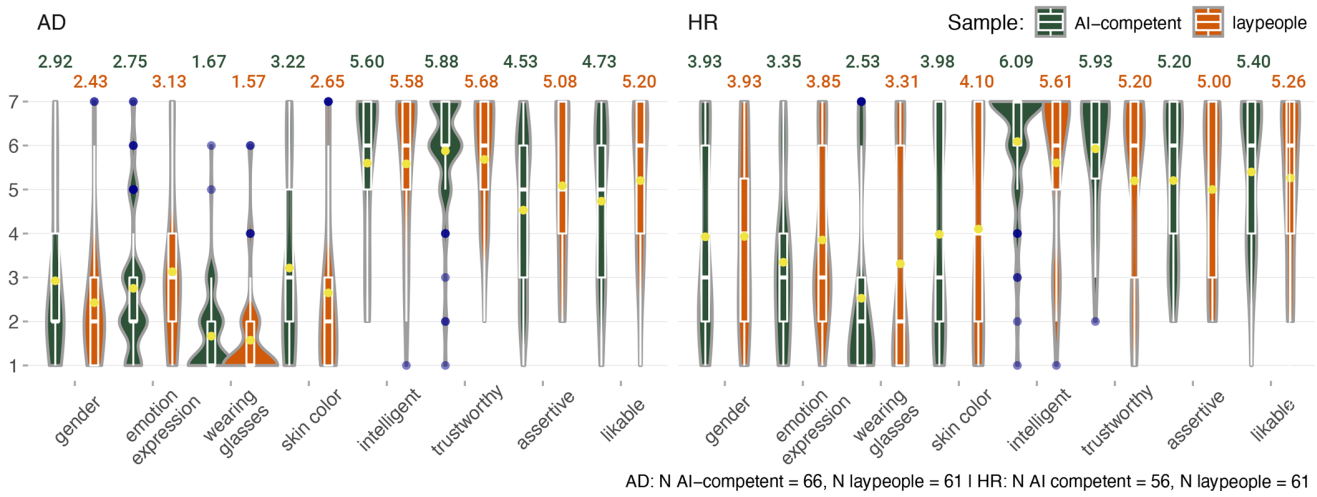


Figure 1: Mean inference ratings in AD vs. HR context by sample. Means of inference ratings for each inference by context and sample show that the AI-competent and laypeople (MT) largely agree in their ratings of facial AI inferences. Rating score 1: "strongly agree", rating score 7: "strongly disagree".

Next, we performed robustness checks by repeating the analysis on random sub-samples of 85% of the datasets. The robustness checks validated the findings. These results replicated findings with a large sample in [31]. The observations also confirmed the results from the Welch two-sample t-test: participants in both samples gave similar agreement-disagreement ratings to each of the inferences.

4.2 AI-competent and laypeople apply similar levels of complexity to their justifications.

To understand how AI-competent and laypeople justified their inference ratings, we first performed a complexity analysis of the open-text justifications. The analyzed justifications consisted of as little as one word up to a few sentences. Depending on the number of arguments embedded in the justification, we assigned a varying amount of themes during the labeling process. For instance, one participant gave the inference *likable* the rating "strongly disagree" and explained that one "absolutely can't tell if someone is likable because of the way they look. It's actually insulting and misleading and unfair to do that." This justification was labeled with the two themes "not sufficient/ good evidence (data) for task", and "bias/ stereotypes/ discrimination". We refer to justifications of this type as two-theme justifications. The use of fewer arguments could indicate that participants have a clear opinion regarding an inference. The use of more themes could indicate a more diverse and complex spectrum of viewpoints regarding an inference.

The analysis (Table 1) shows slight differences in the complexity of justifications by context and inference type. Subjects in the HR context and additionally laypeople in the AD context, provided somewhat more one-theme and less two-theme justifications when justifying their ratings on *character and personality trait* inferences than when justifying their ratings on *construct and feature* inferences. This suggests that evaluations were somewhat clearer for inferences on *character or personality traits*. In contrast, participants discussed inferences on *constructs and features* more diversely.

Table 1: Complexity of subject's justifications (in %)

	AI-competent		laypeople		validation+	
Type	AD	HR	AD	HR	AD	HR
Inferences on constructs and features						
One theme	66.7	64.3	70.9	74.6	62.3	56.4
Two themes	29.2	31.2	27	23.8	30.9	29.4
Three themes	3.8	4.5	2	1.6	6.9	14.2
Four themes	0.4	-	-	-	-	-
# open text answers	*264	224	244	244	204	204
Inferences on character and personality traits						
One theme	66.3	76.8	79.5	80.7	58.8	64.7
Two themes	28.8	19.2	19.3	18.4	32.4	25
Three themes	4.9	2.7	0.8	0.8	8.8	10.3
Four themes	-	-	-	-	-	-
# open text answers	*264	224	244	244	204	204

* After cleaning of the data, more participants from the AI competent sample happened to be in the AD than HR context.
+ More multi-theme justifications by the validation sample may be explained by the longer survey duration.

4.3 Context matters: People agree more with AI inferences in the AD than in the HR context.

We then turned our attention to the experimental variable *context* (AD context vs. HR context) to understand whether and how it influences ratings and justifications of participants.

4.3.1 People agree more with AI inference-making in the low-stake AD context and less in the high-stake HR context. In all three samples, subjects in the HR context showed significantly less agreement with AI facial inferences than subjects in the AD context (AI-competent ($mean_{AD}$ =3.90, $mean_{HR}$ =4.54): $t_{Welch}(99.08)$ =-3.35, p<.01, $\hat{g}_{Hedges}$

=-0.62, $CI_{95\%}$ [-0.99,-0.25]; laypeople ($mean_{AD}$ =3.88, $mean_{HR}$ =4.54): t_{Welch} (118.09) =-3.91, p<.01, $\hat{g}_{Hedges}$ =-0.71, $CI_{95\%}$ [-1.07,-0.34]; validation ($mean_{AD}$ =4.06, $mean_{HR}$ =4.71): t_{Welch} (98.86) =-3.35, p<.01, $\hat{g}_{Hedges}$ =-0.66, $CI_{95\%}$ [-1.06,-0.26]). These results indicate that the application context has an impact on participants' evaluations.

4.3.2 The decision context is the most influential factor in participants' ratings. We performed one six-way ANOVA for each of the eight inferences to analyze the effect of context on the inference rating while controlling for gender, age, education, country, and sample. The variable sample included the AI-competent and laypeople (MT) sample. Using Pillai's trace, ANOVAs with Bonferroni corrections for the eight tests showed that only the variable *context* had a statistically significant effect on inference ratings of *gender* (p<.001), *emotion expression* (p=.015), *wearing glasses* (p<.001) and *skin color* (p=.001). Bonferroni-corrected ANOVAs including the AI-competent and validation laypeople dataset confirmed these results, except for the inference *emotion expression*. We found no other significant effect for any other variable (see Appendix B.3).

4.3.3 Perceptions on the relevance of 'construct and feature' inferences are mixed; in the HR context, laypeople perceive inferences on 'character and personality traits' as relevant. The influence of the decision context was particularly evident when participants emphasized the "irrelevance" or "relevance" of *construct and feature* inferences (see Figure 2, light and dark orange). Participants evaluated these inferences as more "relevant" in the AD context and more "irrelevant" in the HR context. Similarly, participants used the theme "inference (only) sometimes relevant" more frequently in the HR context. This tendency was observed in all samples.

Both laypeople samples applied themes of "(ir)relevance" more frequently than participants with AI competence. Surprisingly, this was particularly the case for MTurk laypeople in the HR context for inferences on *character and personality traits* ("relevant": 15.7%, see Figure 2 light orange). For instance, participants from this sample justified that inferring *intelligence* "would give a hint as to how [...] [applicants] would perform on the job" or that inferring *trustworthiness* "in the workplace can be important and it's not wise to have a dishonest person around". For inferences on *constructs and features*, laypeople underlined the "irrelevance" of the inferences *wearing glasses* (26.2% of laypeople; 29.4% of validation laypeople) and *skin color* (27.9%; 39.2%) in the HR context and the "relevance" of the inferences *wearing glasses* (26.2%; 33.3%) and *gender* (26.2%; 29.4%) in the AD context. Some AI-competent subjects drawn into the AD context agreed that the inferences *wearing glasses* (21.2%) and *gender* (18.2%) are relevant to be inferred (see Appendix D.1).

4.4 Participants justify ratings on *construct and feature* inferences with a wide variety of themes; ratings on *character and personality* inferences with "insufficient data" themes.

Next, we analyzed whether specific themes were of special importance when justifying inference ratings on *constructs and features* or *character and personality traits*.

4.4.1 Ratings on 'construct and feature' inferences are explained by a variety of justification themes. As depicted in Figure 2, all subjects frequently applied themes highlighting "AI ability", "sufficiency" of the data, and – depending on the AD or HR context – the "relevance" or "irrelevance" of an inference. AI-competent participants raised somewhat more "ethical and discriminatory concerns". Overall, justifications included a substantial variety of justification themes.

4.4.2 Ratings on 'character and personality trait' inferences are predominately explained by the "insufficiency" of a profile picture as evidence. The use of the "insufficiency" theme was particularly prevalent for laypeople in the HR context (AI-competent: 37.5%, laypeople: 56.7%; validation: 39.3%). Again, individuals with AI competence raised "ethical and discriminatory concerns" more often than participants in both laypeople samples. Furthermore, participants made references to the "subjectivity" of the inference task.

4.4.3 Participants believe "AI can infer" whether a person is wearing glasses on a portrait picture; they are skeptical about AI's ability to infer emotional expression. All three samples used the themes "technical ability of AI", "accurate and well working" models, and "easy to infer" most frequently to justify ratings on the inference *wearing glasses*. They applied the theme "can infer sometimes/ difficult in some situation" most often to justify ratings on *emotion expression* and *gender*. For instance, one participant explained that while "the majority of people can have a gender revealed through just a picture, not everyone fits that mold."

Some participants from both main samples believed that a "profile picture is good evidence" for the inferences *wearing glasses* and *emotion expression*. At the same time, there were critical voices stating that a profile picture is not sufficient evidence to infer *emotion expression*, e.g., "Emotion changes by the hour or minute. Can't make an inference based on that." The validation dataset supported these latter results.

4.5 Co-occurrence analysis: "AI (in)ability" and data-related themes co-occur most often with other themes.

We then analyzed the co-occurrence of themes with each other to identify patterns in the use of multiple justification themes (see Appendix D.2). We found that for inferences on *constructs and features*, the AI-competent raised concerns but acknowledged AI to be able to make certain inferences. Referring to inferences on *constructs and features*, people with AI competence raised "ethical and discriminatory concerns" in combination with almost all other justification themes, however, most frequently in combinations with themes on "AI ability" or the "sufficiency" of the profile picture as evidence (see Figure 5a and 5b-1 in the Appendix). This relationship reversed for justifications of ratings on *character and personality trait* inferences. Here, "ethical and discriminatory concerns" were most frequently brought forward in combination with themes on the "insufficiency" of a profile picture as evidence (see Figure 5a and 5b-3 in the Appendix).

For inferences on *character and personality traits*, laypeople often paired comments on the "(in)sufficiency" or "(in)adequacy" of the data with another theme. For *constructs and features*, a greater variety of theme combinations was observed.

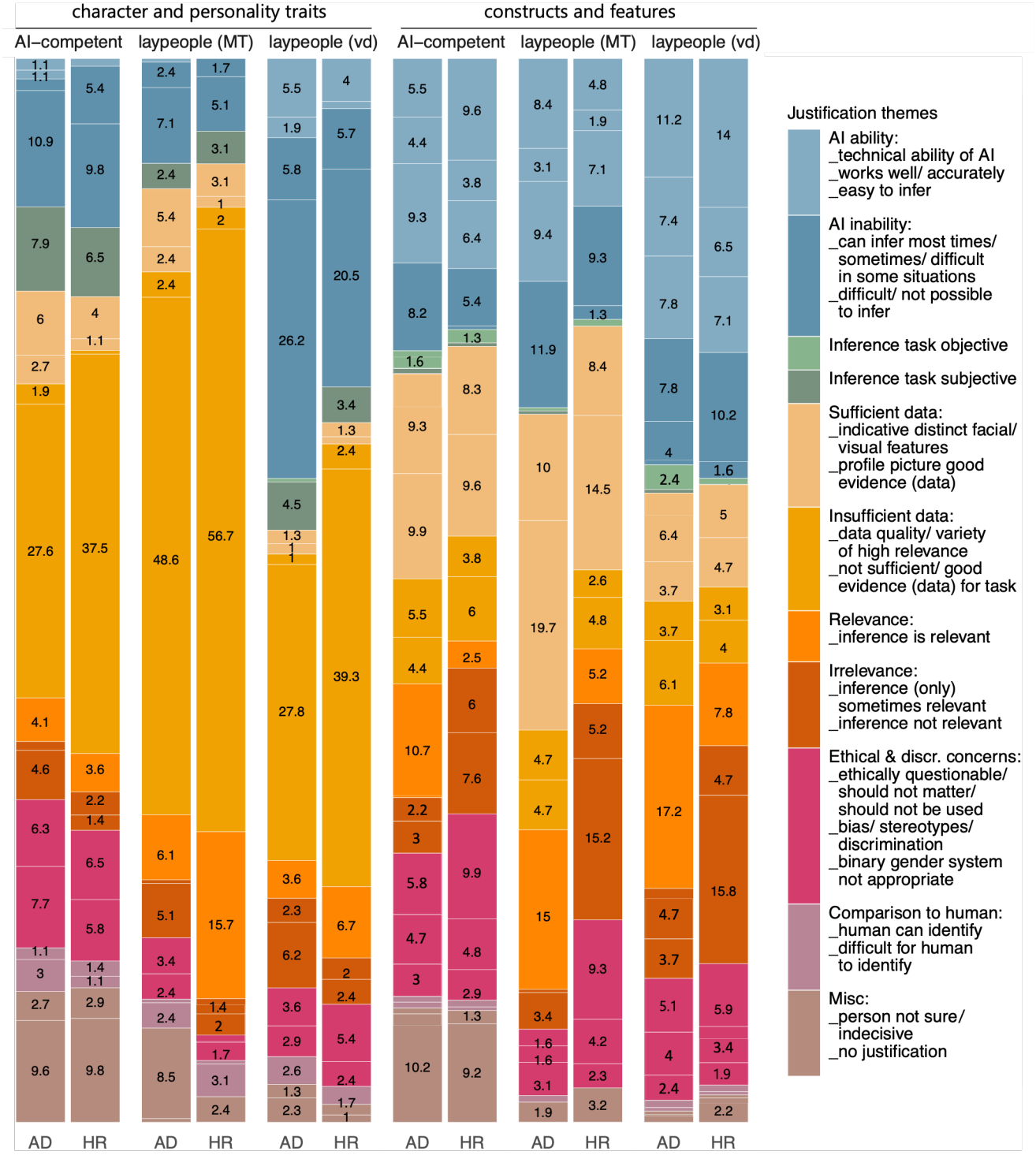


Figure 2: Percentages of individual themes grouped by super-ordinate topic, by context, and by sample. Stacked bars add up to 100% and represent the total of individual themes used by the specific sample. Only percentages > 1% are labeled on the graph.

4.6 Many inferences are based on questionable norms or resemble social constructs and societal stereotypes.

To understand participants' most critical concerns, we finally focused on themes related to "ethical and discriminatory concerns" and "AI inability" (see Figure 4 in the Appendix).

4.6.1 Individuals with AI competence perceive the inference likable as subjective. More than laypeople, individuals with AI competence and subjects from the validation sample described the inference *likable* as "relative", "based on sympathy", and "subjective", e.g., "Likability is a matter of perspective" and "depends on the observer." Comments also referred to other justification themes such as ethical concerns, e.g., "Likability is a highly subjective measure and inherently biased. In addition, it is highly unethical to have such type of decisions made by systems that are not capable of understanding the impact of this decisions" [sic] or "Likeability itself is an ill defined thing, predicting it from just portraits is wrong". Participants did not consider any other inference as equally subjective as *likable*.

4.6.2 Some subjects state that inferences on 'character and personality traits' cannot be inferred. However, approximately half of subjects highlight that the data is simply insufficient or inadequate. A considerable amount of subjects from all samples stated that a profile picture is "insufficient" data (26%-79% depending on inference, context, and sample) to infer *character and personality traits*. For instance, subjects commented that "[n]o facial features indicate trust", or that *intelligence* "is not quantifiable through visual data". At the same time, a minority (~15%) of the AI-competent, a small percentage of laypeople, and many participants from the validation dataset argued that AI cannot infer specific *character or personality traits*. An AI-competent participant explained that the "problem here is ill-posed", there "is no general understanding", and "no clear" or "objective definition of intelligence that everyone agrees with!" Given the lack of shared definitions, some asked "how is this measured? How is it implemented during training?", and "What are the parameters for identifying someone as intelligent?" These findings suggest that some participants evaluated inferences such as *intelligent* and *trustworthy* as social conceptualizations that require a common understanding before being used as inference in facial analysis AI.

4.6.3 Participants with AI competence believe that stereotypical judgments enable AI to draw 'character and personality traits'. Other people with AI competence worried about "stereotypes" embedded in the training data. They elaborated that, e.g., "a categorization of intelligence based on looks seems to correlate features that are not correlated" or that "the training data for trustworthiness depends on societal stereotypes and not actual trustworthyness" [sic]. Conversely, the existence of "stereotypes" was also used to argue in favor of AI being able to make an inference. For instance, a participant explained that the inference *likable* "makes sense because some people's appearance is appealing to more people. But, this inference can only be made on a statistical basis: Person is or is not likable on average." AI-competent participants stated justifications in relation to "bias, stereotypes and discrimination" most frequently when referring to the inferences *trustworthy*, *assertive*, and *likable*, e.g., one participant commented that "it's an unethical idea to give ai systems the ability to inference something so loosely defined and this will lead to biased choices made in the name of "science"." Laypeople did not show these levels of concern for any of these inferences.

4.6.4 A minority of participants raises concerns regarding the inference skin color. In the HR context, 23% of subjects from all samples raised "ethical concerns" regarding the inference *skin color*. One subject commented that *skin color* "should not be a criterion for job applications. Furthermore, being of a certain skin color should be a matter of self-description and not be determined by a computer program". Some participants also perceived the inference *skin color* to be based on biased data or to lead to discrimination: "Users will get predictions based on race and race-based stereotypes" or "if the model is biased towards skin color, it may not encourage a fair AI agent." Some subjects highlighted that *skin color* can be inferred but should not be done or used: "Color can be detected easily by computer vision frameworks (though this inference imposes certain ethical questions)" or "While it is possible to determine the skin color of a person from a portrait [...], it is ethically incorrect to base any decisions on skin color" or "Detecting skin colour should be trivial for the software, so it is reasonable to expect that inference. It is NOT reasonable that this information should be used to indicate whether someone is suitable for the job." These comments exemplify the diversity of normative evaluation of the inference *skin color*. Although suggesting that AI can infer *skin color*, this inference – which some specifically relate to "race" or "ethnicity" – was perceived as an impermissible inference by a considerable number of subjects.

4.6.5 A minority of participants highlights that binary gender norms are not appropriate and ethically questionable. Referring to the inference *gender*, some participants raised "ethical concerns" in the HR context (AI-competent: 16.1%; laypeople: 11.5%). In both contexts, 9% of participants with AI competence believed that inferences on gender are based on biased data: "The AI might learn to assign gender identity based on a heavily biased training data which are influenced by conventional gender identity norms hence making fateful inferences in the real world. Such inferences are unreasonable". Some subjects across all samples specifically highlighted that "gender norms are not appropriate" anymore: "This used to be a more 'objective' decision, however society has changed and persons can decide by themselves their gender, without being guided by their appearance. The most important part is, again, the inability of an AI system to understand the consequences of deciding something like this". Others commented that gender can be inferred but is not appropriate: "this is very apparent and thus somewhat alright, but then again, gender is a fluid concept". Some participants believed gender to be a social construct that is not binary as is often presupposed by facial analysis AI.

5 KEY OBSERVATIONS AND DISCUSSION

Overall, our study on the ethical perceptions of facial analysis AI suggests that there are no "common sense" facial analysis inferences. In all samples, there are participants who raise concerns, in particular, *ethical concerns* that inferences lack epistemic validity,

should not matter or should not be used for the purpose of an application. In addition, we find that both AI-competent and laypeople express a variety of normative concerns regarding AI facial inferences. At the same time, only a minority of participants concluded that AI cannot, under any circumstance, make an inference from faces.

Regarding the facial inference *emotion expression*, participants note that a profile picture is only a snapshot and thus, "temporary and short-lived". Recently, emotion researchers have argued that emotion expression is more context-dependent and variable than commonly assumed. The *emotional state* of a person cannot be readily inferred from a person's facial expression [6]. Participants in both samples raised similar concerns. For example, one participant stated that there "are numerous people that tend to hide their emotions through pictures [...]".

Our analysis of justifications clearly shows that participants voice concerns regarding the classification of latent traits by facial analysis. Participants pointed out that the inference of attributes such as *intelligence* from facial information presupposed a highly simplified definition of a multidimensional concept. Similarly, participants mentioned potential problems related to the subjectivity associated with inferring attributes such as *likability* from faces.

We found that participants criticized the ethically problematic application of a binary conceptualization of *gender*. This finding aligns with recent critical data science research on computer vision. Here, authors, too, point to the fact that sensitive categories, such as gender and race, are often treated as "common sense categories" in computer vision datasets [25, 70, 80, 85].

On the other hand, a justification theme among both laypeople and people with AI competence pertains to the *possibility* of an AI inference provided that the "data is correct". This line of reasoning resembles narratives behind facial analysis AI research and commercial tools that try to solve issues with predictive power at the level of data *rather than question their epistemic foundations*. Some of the AI-competent and laypeople used entrenched stereotypical heuristics to evaluate AI facial inferences. While heuristics and stereotypes may initially help humans navigate through complex social interactions, research on the validity of human inferences from faces demonstrates that faces are no "strong and reliable indicator of people's underlying traits" [95, p.569].

Some specific differences between the two main samples could be observed. Both laypeople samples applied more pragmatic justifications referring to the "(ir)relevance" of the visual data for a decision-making procedure. For inferences on *character and personality traits*, more than half of laypeople (MT) described the data as "insufficient" for the inference task. People with AI competence mentioned themes related to "(ir)relevance" and "insufficiency" less frequently than laypeople, but raised "ethical concerns" more frequently than laypeople.

The complexities behind participants' justifications indicate a "struggle" for the power over the creation and attribution of meaning for visual data. Our study asks who can and should participate in this discourse. AI experts currently have free rein over the meaning that their datasets should be attributed with. However, politicians are aware of the complexities behind the meaning of visual data [e.g., 32] and we highlight again that more and more critics are voicing ethical concerns [e.g., 25, 42, 51, 80, 85, 89]. One of our main concerns is that the inference of perceived traits or features, e.g., "perceived trustworthiness" [e.g., 84] as opposed to "actual trustworthiness" by an AI system ultimately contributes to society remaining trapped in a cycle of stereotypes.

Taken together, we note that participants in all samples showed a tendency *to oppose* facial AI inference-making. Participants' evaluations underline many of the ethical complications of facial analysis AI that have recently been raised by critical data scientists and other scholars. Moreover, we see that people do not apply a consistent and universal justification profile for each of the facial inferences. Facial inferences are not simple constructs but overloaded with epistemic and pragmatic intuitions that are likely influenced by factors including cultural background.

We end by wondering how a justifiable ethical framework for facial AI inference-making could look like. What "standards" would a satisfactory justification fulfill? Given that we deal with *visual* inferences, we believe that they should first achieve reasonable epistemic validity and that this validity should be supported by scientific agreement over the quality of the evidence. The question then is what a reasonable level of scientific agreement should look like. We have pointed out that while a large majority of researchers underline the invalidity of first facial impressions, there is an ongoing stream of research publications that claim to present evidence on the validity of first impressions.

Participants in our samples disagreed with inferences common in human first impression-making (e.g., trustworthiness, likability etc.) by algorithmic systems. Indeed, one of the core findings of this work is that neither individuals with AI competence nor laypeople trust many of the inferences of facial analysis technology. With legislative attempts seeking to ban certain facial processing technologies, with a plethora of scholars pointing to the dangers of an automated version of physiognomy, and the different sample populations expressing their lack of trust toward such AI inference-making, we ask in what context and under what circumstances such facial analysis AI can be justified at all. It appears that, more often than not, there are better *reasons not to develop and deploy AI* that analyzes human faces to draw a variety of inferences that are then used for a particular decision-making context. Weaving together the argumentation threads from our previous results [31], critical remarks of data scientists and policy-makers, we take it that there is a strong case to be made that such AI inference-making is epistemically invalid, pragmatically of little use, and, overall, contributes and perpetuates stereotypes that stand in conflict with a society's welfare.

6 LIMITATIONS AND FUTURE DIRECTION

Our samples were composed of comparatively young people with AI competence that are not representative of all AI researchers. This may have introduced a bias in terms of the participants' understanding of and critiques on social constructs such as gender identities. In addition, this study does not include voices from industry. Future research should also survey corporate AI developers.

This research makes a methodological contribution by providing an AI knowledge instrument as an alternative to self-reported AI knowledge measures. We hope that the results from the application of the AI knowledge test will act as a starting point for the utilization

of a more objective and reliable measure of knowledge on AI. It should be noted that given rapid advances in AI, the questions contained in the AI quiz should be regularly updated.

Our sample included participants from the United States (laypeople sample) and Europe (AI-competent and validation laypeople sample). We addressed the limitation of comparability of the two main samples by creating a validation dataset that shows substantial similarity in terms of demographics with the AI-competent sample. Given the international application of AI systems, diverse study participants are vital. Hence, future studies should explore whether cultural differences influence ethical concerns of facial processing technologies such as facial analysis AI. If there are no such cross-cultural differences then this could serve as evidence for the existence of culturally-universal ethical perceptions of facial inferences.

Whereas we evaluated the perception of AI inferences from profile pictures, future research should also evaluate perceptions of AI inferences from videos. Given that videos are used for a variety of inference tasks [11], the perception of somewhat more accurate results can be expected. However, it remains to be seen whether video data will influence whether such traits *should* be inferred.

7 CONCLUSION

As the use of AI grows in popularity and as the impact of AI inference-making on societies increases, so does the responsibility of those who develop such AI systems. A special focus must be placed on exploring the perspectives of a diverse group of people both who are potentially driving the implementation of computer vision and AI and those that are subjected to its inference-making.

This work provides insights into perceptions of AI inference-making by the general public compared to perceptions of individuals with high knowledge of AI. It suggests that, by and large, people with AI competence and the general public share many perceptions about AI inference-making and have distinct context- and task-dependent perceptual differences. Being aware of the perceptions and judgments of people with AI competence, on the one side, and users, on the other side, is essential to develop AI systems that are based on democratic discourse, accepted by society, and sustainable.

Concluding this research, we summarize that the application context does have an effect on how people perceive AI inference-making from faces. While differences in AI competence did not have an effect on the inference ratings, specific differences were observable for the ethical justifications. We found that both laypeople and people with AI knowledge showed more agreement with AI inference-making in the low-stake AD context than in the high-stake HR context. In both contexts, people with AI competence – although only a small minority – raised ethical and discriminatory concerns more frequently than laypeople. Laypeople made more references to themes related to the (ir)relevance of the inference for the context of application.

Having explored the question whether differences in AI knowledge account for changes in the perceptions of AI inference-making across two contexts, this work extends research in the field of perceptions of algorithmic systems and contributes to the nascent literature on AI experts' perceptions on AI inference-making. The results invite a deeper reflection on the similarities and differences in the perceptions of AI among different people within the general population. With this work, we aim to ultimately contribute to the development of sustainable AI systems that are supported, not only by their developers, but also by the general public.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments that improved the paper. We thank the study participants for taking part in this study. For their valuable feedback, we thank the participants of the 2021 CEPE/International Association of Computing and Philosophy Conference, the participants of the 2021 Ethics and Technology Lecture Series of the Munich Center for Technology in Society, and the participants of the Venice 2019 Metaethics of AI & Self-learning Robots Workshop.

FUNDING & SUPPORT

This research was conducted with the help of a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

REFERENCES

[1] Noura Al Moubayed, Yolanda Vazquez-Alvarez, Alex McKay, and Alessandro Vinciarelli. 2014. Face-based automatic personality perception. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. Association for Computing Machinery, New York, NY, USA, 1153–1156. https://doi.org/10.1145/2647868.2655014

[2] Kwame Anthony Appiah. 2008. *Experiments in Ethics*. Harvard University Press, Cambridge, Massachusetts, USA.

[3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[4] Christiane Atzmüller and Peter M Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138.

[5] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences* 124 (2018), 150–159. https://doi.org/10.1016/j.paid.2017.12.018

[6] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[7] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. 2016. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers, New Jersey, USA, 5562–5570. https://doi.org/10.1109/CVPR.2016.600

[8] Betaface. 2021. Betaface API. https://www.betafaceapi.com/wpa/ Accessed: 2022-02-27.

[9] Sky Biometry. 2021. Face Recognition Demo. https://skybiometry.com/demo/face-recognition-demo/ Accessed: 2022-02-27.

[10] Jean-François Bonnefon, Astrid Hopfensitz, and Wim De Neys. 2015. Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences* 19, 8 (2015), 421–422. https://doi.org/10.1016/j.tics.2015.05.002

[11] Grzegorz Brodny, Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. 2016. Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. In *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, Portsmouth, UK, 397–404. https://doi.org/10.1109/HSI.2016.7529664

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91.

[13] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2019. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2019), 220–239.

[14] Sarah Castell, Daniel Cameron, Steven Ginnis, Glenn Gottfried, and Kelly Maguire. 2017. Public views of machine learning – Findings from public research and engagement. 92 pages. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

[15] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. "Scary robots": Examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 331–337. https://doi.org/10.1145/3306618.3314232

[16] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and perceptions of AI and why they matter. 28 pages. https://royalsociety.org/~/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf

[17] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2 (2019), 74–78.

[18] Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. Association for Computing Machinery, New York, NY, USA, 1101–1104. https://doi.org/10.1145/2647868.2654977

[19] François Chollet. 2018. *Deep Learning with Python*. Manning Publications, Shelter Island, NY, USA. https://books.google.de/books?id=mjVKEAAAQBAJ

[20] Ching-Hua Chuan, Wan-Hsiu Sunny Tsai, and Su Yeon Cho. 2019. Framing artificial intelligence in American newspapers. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 339–344. https://doi.org/10.1145/3306618.3314285

[21] Cory J Clark, Jamie B Luguri, Peter H Ditto, Joshua Knobe, Azim F Shariff, and Roy F Baumeister. 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106, 4 (2014), 501.

[22] Clearview.ai. 2022. Overview. https://www.clearview.ai/overview Accessed: 2022-02-24.

[23] Kate Crawford. 2021. Time to regulate AI that interprets human emotions. *Nature* 592, 7853 (2021), 167–167.

[24] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, A Sánchez, et al. 2019. AI Now 2019 Report. AI Now Institute.

[25] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY* 36 (2021), 1105–1116.

[26] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 423–426. https://doi.org/10.1145/2818346.2829994

[27] Charles Efferson and Sonja Vogt. 2013. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports* 3, 1 (2013), 1–7. https://doi.org/10.1038/srep01047

[28] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129. https://doi.org/10.1037/h0030377

[29] EmoVu. n.d.. EmoVu Mobile. https://www.programmableweb.com/sdk/emovu-mobile Accessed: 2022-02-25.

[30] Severin Engelmann and Jens Grossklags. 2019. Setting the stage: Towards principles for reasonable image inferences. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 301–307. https://doi.org/10.1145/3314183.3323846

[31] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What people think AI should infer from faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 128–141. https://doi.org/10.1145/3531146.3533080

[32] European Parliament. 2021. Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1678184&t=e&l=en

[33] Face++. 2022. Face Attributes. https://www.faceplusplus.com/attributes/ Accessed: 2022-02-25.

[34] Faception. 2021. Our technology. https://www.faception.com/our-technology Accessed: 2022-02-24.

[35] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, Palo Alto, CA, USA, 963–969.

[36] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with Instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015 (EMPIRE '15)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/2809643.2809644

[37] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2016. Using Instagram picture features to predict users' personality. In *MultiMedia Modeling*, Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu (Eds.). Springer International Publishing, Cham, 850–861. https://doi.org/10.1007/978-3-319-27671-7_71

[38] Bruce Ferwerda and Marko Tkalcic. 2018. Predicting users' personality from Instagram pictures: Using visual and/or content features? In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 157–161. https://doi.org/10.1145/3209219.3209248

[39] Jake Goldenfein. 2019. The profiling potential of computer vision and the challenge of computational empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/3287560.3287568

[40] Armin Granulo, Christoph Fuchs, and Stefano Puntoni. 2019. Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour* 3, 10 (2019), 1062–1069.

[41] Edward Guadagnoli and Wayne F Velicer. 1988. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103, 2 (1988), 265—-275. https://doi.org/10.1037/0033-2909.103.2.265

[42] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated Alt Text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 543–554. https://doi.org/10.1145/3461702.3462620

[43] Isabella Hermann. 2020. Beware of fictional AI narratives. *Nature Machine Intelligence* 2, 11 (2020), 654–654.

[44] Michel Hohendanner, Chiara Ullstein, and Mizuno Daijiro. 2021. Designing the exploration of common good within digital environments: A deliberative speculative design framework and the analysis of resulting narratives. In *Proceedings of the Swiss Design Network Symposium 2021 on Design as Common Good – Framing Design through Pluralism and Social Values*. SUPSI, HSLU, swissdesignnetwork, Lucerne, Switzerland, 566–580.

[45] Graeme D Hutcheson and Nick Sofroniou. 1999. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. Sage Publications, New York City, NY, USA.

[46] Michael R Hyman and Susan D Steiner. 1996. The vignette method in business ethics research: Current uses, limitations, and recommendations. *Studies* 20, 100.0 (1996), 74–100.

[47] Alexander Kachur, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. 2020. Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports* 10, 1 (2020), 1–11.

[48] Henry F Kaiser. 1974. An index of factorial simplicity. *Psychometrika* 39 (1974), 31–36. https://doi.org/10.1007/BF02291575

[49] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 690–700. https://doi.org/10.1145/3442188.3445931

[50] Esther Kaufmann. 2021. Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence* 2 (2021), 100028. https://doi.org/10.1016/j.caeai.2021.100028

[51] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 587–597. https://doi.org/10.1145/3442188.3445920

[52] Owen C King. 2019. Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*. Springer, Cham, 265–282.

[53] Owen C King. 2020. Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. *Ethics and Information Technology* 22, 1 (2020), 25–37.

[54] Karel Kleisner, Veronika Chvátalová, and Jaroslav Flegr. 2014. Perceived intelligence is associated with measured intelligence in men but not women. *PLOS ONE* 9, 3 (2014), 1–7. https://doi.org/10.1371/journal.pone.0081237

[55] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194.

[56] Joshua Knobe and Shaun Nichols. 2017. Experimental Philosophy. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, Stanford, CA, USA.

[57] Michal Kosinski. 2021. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11, 1 (2021), 1–7.

[58] Steven R Kraaijeveld. 2021. Experimental philosophy of technology. *Philosophy & Technology* 34 (2021), 993–1012.

[59] Klaus Krippendorff. 2004. Reliability in content analysis. *Human Communication Research* 30, 3 (2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[60] Udo Kuckartz. 2009. *Evaluation online: internetgestützte Befragung in der Praxis.* Springer VS Wiesbaden, Wiesbaden, Germany.

[61] Udo Kuckartz. 2014. *Mixed methods: Methodologie, Forschungsdesigns und Analyseverfahren.* Springer VS Wiesbaden, Wiesbaden, Germany.

[62] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* Association for Computing Machinery, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

[63] John Leuner. 2019. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. arXiv:cs.CV/1902.10739

[64] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, Vol. 10. AAAI, Cologne, Germany, 211–220. https://ojs.aaai.org/index.php/ICWSM/article/view/14738

[65] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[66] [House Of Lords]. 2018. AI in the UK: Ready, willing and able? https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[67] Philipp Mayring. 2015. *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12 ed.). Weinheim/Basel: Beltz Verlag.

[68] David E Melnikoff and Nina Strohminger. 2020. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour* 4 (2020), 1–7.

[69] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[70] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[71] Laura Naumann, Simine Vazire, Peter Rentfrow, and Samuel Gosling. 2009. Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin* 35, 12 (2009), 1661–1671. https://doi.org/10.1177%2F0146167209346309

[72] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41, 4 (2007), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

[73] Noldus. [n.d.]. Facial Expression Analysis. https://www.noldus.com/facereader/facial-expression-analysis Accessed: 2022-02-27.

[74] Harriet Over and Richard Cook. 2018. Where do spontaneous first impressions of faces come from? *Cognition* 170 (2018), 190–200. https://doi.org/10.1016/j.cognition.2017.10.002

[75] Harriet Over, Adam Eggleston, and Richard Cook. 2020. Ritual and the origins of first impressions. *Philosophical Transactions of the Royal Society B* 375, 1805 (2020), 20190435. https://doi.org/10.1098/rstb.2019.0435

[76] Ian S Penton-Voak, Nicholas Pound, Anthony C Little, and David I Perrett. 2006. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition* 24, 5 (2006), 607–640. https://doi.org/10.1521/soco.2006.24.5.607

[77] Emma Pierson. 2018. Demographics and Discussion Influence Views on Algorithmic Fairness. arXiv:cs.CY/1712.09124

[78] Lin Qiu, Jiahui Lu, Shanshan Yang, Weina Qu, and Tingshao Zhu. 2015. What does your selfie say about you? *Computers in Human Behavior* (2015), 443–449. https://doi.org/10.1016/j.chb.2015.06.032

[79] Stefan Rädiker and Udo Kuckartz. 2019. *Analyse qualitativer Daten mit MAXQDA.* Springer VS Wiesbaden, Wiesbaden, Germany.

[80] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[81] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[82] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 145–151. https://doi.org/10.1145/3375627.3375820

[83] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an online intervention to teach artificial intelligence with LearningML to 10-16-year-old students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education.* Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3408877.3432393

[84] Lou Safra, Coralie Chevallier, Julie Grèzes, and Nicolas Baumard. 2020. Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature Communications* 11, 1 (2020), 1–7. https://doi.org/10.1038/s41467-020-18566-7

[85] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–35. https://doi.org/10.1145/3392866

[86] Cristina Segalin, Dong Seon Cheng, and Marco Cristani. 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding* 156 (2017), 34–50. https://doi.org/10.1016/j.cviu.2016.10.013

[87] Crisitina Segalin, Alessandro Perina, Marco Cristani, and Alessandro Vinciarelli. 2016. The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing* 8, 2 (2016), 268–285. https://doi.org/10.1109/TAFFC.2016.2516994

[88] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129

[89] Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 782–793. https://doi.org/10.1145/3442188.3445939

[90] Luke Stark and Jevan Hutson. 2022. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* 32, 4 (2022). https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2

[91] Louis Leon Thurstone. 1947. *Multiple-factor Analysis; A Development and Expansion of The Vectors of Mind.* University of Chicago Press.

[92] Alexander Todorov. 2017. *Face Value: The Irresistible Influence of First Impressions.* Princeton University Press.

[93] Alexander Todorov, Sean G Baron, and Nikolaas N Oosterhof. 2008. Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience* 3, 2 (2008), 119–127. https://doi.org/10.1093/scan/nsn009

[94] Alexander Todorov, Friederike Funk, and Christopher Y Olivola. 2015. Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences* 19, 8 (2015), 422–423. https://doi.org/10.1016/j.tics.2015.05.013

[95] Alexander Todorov, Christopher Y Olivola, Ron Dotsch, and Peter Mende-Siedlecki. 2015. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology* 66 (2015), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

[96] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114, 2 (2018), 246–257. https://psycnet.apa.org/doi/10.1037/pspa0000098

[97] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 726–737. https://ojs.aaai.org/index.php/ICWSM/article/view/7338

[98] Yan Yan, Jie Nie, Lei Huang, Zhen Li, Qinglei Cao, and Zhiqiang Wei. 2015. Is your first impression reliable? Trustworthy analysis Using facial traits in portraits. In *MultiMedia Modeling*, Xiangjian He, Suhuai Luo, Dacheng Tao, Changsheng Xu, Jie Yang, and Muhammad Abul Hasan (Eds.). Springer International Publishing, Cham, 148–158. https://doi.org/10.1007/978-3-319-14442-9_13

[99] Ansgar Zerfass, Jens Hagelstein, and Ralph Tench. 2020. Artificial intelligence in communication management: A cross-national study on adoption and knowledge, impact, challenges and risks. *Journal of Communication Management* 24, 4 (2020), 377–389. https://doi.org/10.1108/JCOM-10-2019-0137

[100] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C Horowitz, and Allan Dafoe. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research* 71 (2021), 591–666. https://doi.org/10.1613/jair.1.12895

[101] Cornelia Züll and Natalja Menold. 2019. Offene Fragen. In *Handbuch Methoden der empirischen Sozialforschung*, Nina Baur and Jörg Blasius (Eds.). Springer, 855–862.

Appendix: AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI

CHIARA ULLSTEIN*, Technical University of Munich, Chair of Cyber Trust, Germany
SEVERIN ENGELMANN*, Technical University of Munich, Chair of Cyber Trust, Germany
ORESTIS PAPAKYRIAKOPOULOS†, Princeton University, Center for Information Technology Policy, USA
MICHEL HOHENDANNER‡, University of Applied Sciences Munich, Center for Digital Sciences and AI & Faculty of Design, Germany
JENS GROSSKLAGS, Technical University of Munich, Chair of Cyber Trust, Germany

ACM Reference Format:
Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Michel Hohendanner, and Jens Grossklags. 2022. Appendix: AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3551624.3555294

A RESEARCH DESIGN AND METHODS

A.1 Overview of Methods Applied in Studies to Measure AI Knowledge

Zerfass et al. [17] measured Artificial Intelligence (AI) expertise of practitioners in the communications professions using an 8-item quiz, and AI adoption by asking whether participants were using specific AI applications (e.g., Siri) on their phones or AI devices (e.g., Alexa) in their homes or offices. Knowledge on expert models was measured based on the quality of a definition participants were asked to provide in response to an open-end question [10]. Technical knowledge was measured by means of three questions about self-reported knowledge on computer programming, algorithms and AI [1].

Instead of measuring AI knowledge, [18] surveyed researchers who published in leading AI/ML conferences and assumed them to have high AI knowledge. Others surveyed students studying AI [9, 12]. Kasinidou et al. [9] additionally measured their level of knowledge on fairness in algorithmic decision-making or prior training on topics such as algorithm accountability, transparency and fairness through a self-reported 5-point Likert scale.

Again other studies used knowledge in computer programming and numeracy, as measured by Logg et al. [11] using a 11-item numeracy scale by Schwartz et al. [14] as a proxy. Logg et al. [11] measured familiarity with algorithms by asking participants how certain they were to know what an algorithm is.

*Denotes equal contribution.
†Currently affiliated with Sony AI, Switzerland
‡Also affiliated with University of Wuppertal, Industrial Design, Germany.

A.2 Survey Vignette

Figure 1 shows the vignette presented to the participants of the AI-competent and validation laypeople samples, which was based on a the vignette presented in [3]. The same wording and order of text passages were used.

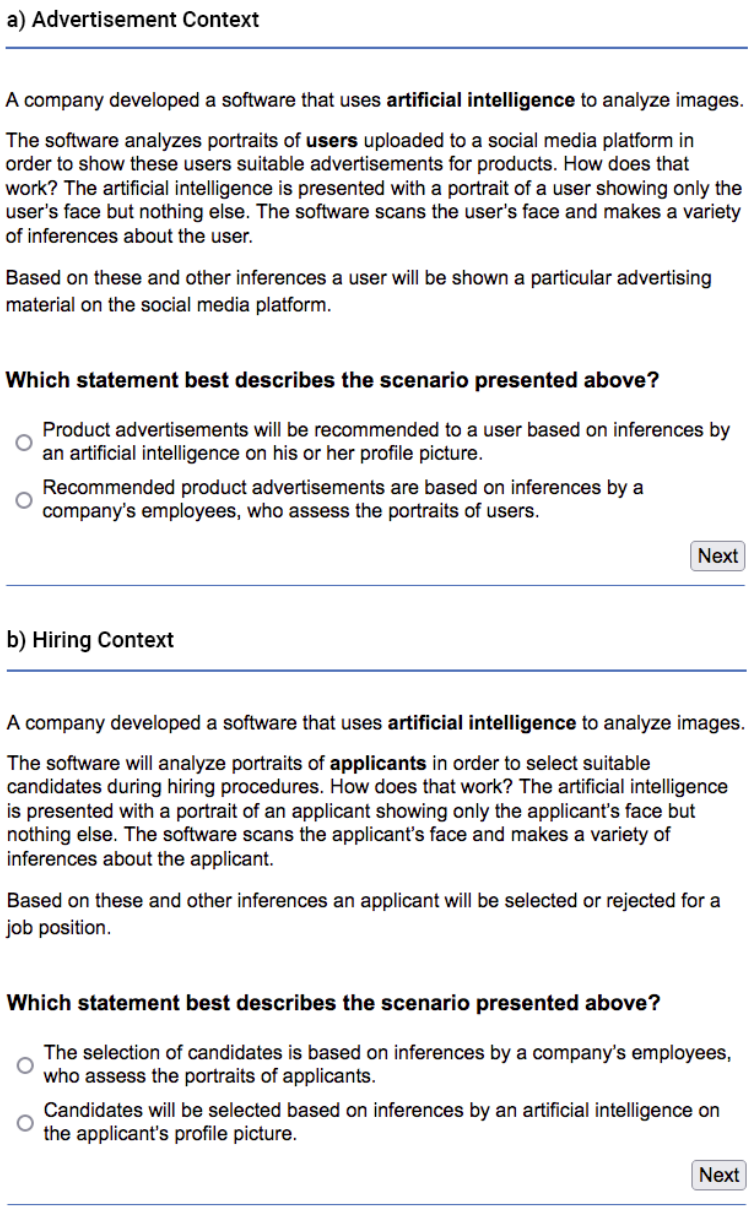


Fig. 1. Scenario presented to study participants in a) the advertisement context or b) the hiring context.

A.3 AI Knowledge Measure

A.3.1 Construction. In order to better assess respondent's AI knowledge, we complemented a self-rated AI knowledge level instrument (one item 5-point Likert scale) by an AI knowledge measure (see Table 1). This measure is based on an instrument used to asses students' AI and Machine Learning (ML) knowledge by Rodríguez-García et al. [13]. The developed AI knowledge measure contains single-choice questions on ML of varying degrees of difficulty. The measure by Rodríguez-García et al. [13] was adapted to the purposes of this study as follows:

Four questions that originated from Estevez et al. [4] were excluded. Those four questions were originally intended to measure the change in knowledge of AI after a workshop-based intervention. Additionally, four questions were removed that did not seem to be fitting for the purpose of this research study. One item in P4 was replaced by an item that is less philosophically disputable. The wording of P5 was changed slightly to make the items shorter. Furthermore, three questions that were not perceived to be fitting for the purposes of this study were removed. Finally, for all four-item questions, one wrong item was exchanged with the answer "I don't know". For the question with two items, an "I don't know" option was added. Without the option "I don't know", respondents would have had either to guess or to choose one answer at random, which would have introduced a bias. Given that this research focuses on computer vision, three self-constructed computer vision specific questions (Q7 - Q9) were added, based on Chollet [2]. In summary, Q1 through Q5 reflect general questions on ML whereas Q6 through Q9 focus specifically on computer vision and are expected to be answerable by less respondents. In Table 1, correct items are marked with an "(X)".

A.3.2 Additional survey questions related to AI knowledge. Besides the questions related to the AI knowledge test, we included a number of additional questions to the survey that allowed us to verify the results from the AI knowledge test. We added two questions on the number of AI courses that the participant took part in (with a technical and with a socio-political or ethical focus). Furthermore, we included three questions to control for the knowledge on the presented AI scenarios, the science of first impression-making, and potential external assistance. To control whether specific AI knowledge might have come from their corporate experiences, we asked participants whether they have an (AI-related) job. We also asked participants how they learned about the survey and what research field best described their research (see Table 2).

A.3.3 Validation. Before running our main study, we tested the AI knowledge measure by running a pre-study with three participants. Participants received a survey with the AI knowledge test questions and an additional question designed to indicate the perceived difficulty of each question in the test. Furthermore, the survey asked for an indication of the number of courses with a focus on technical AI, as well as the number of courses with a social-political and/or ethical AI focus. Additionally, participants were asked to indicate their level of AI knowledge on a 5-point Likert-style scale.

Participants were briefed that they were part of a pre-study that helped evaluate the AI knowledge test. Each participant provided feedback on how long it took to complete the survey and whether any questions were misleading. This feedback was gathered and first discussed with the research team. Then, any remaining issues were discussed with an AI expert not part of the research team.

Based on the feedback from the pre-study, the number of mixed examples in P11 for the correct item was increased (from 10 to 1000) to ensure that the strategy described in this item would more clearly result in a the better system. Furthermore, one item was removed from the AI knowledge quiz, because – based on assumptions made by the participant – all of the items might arguably have been correct.

Table 1. AI knowledge test: Questions. Changes to original items are indicated.

Name	Orig.	Item and Anchors
Q1	P4	**When an artificial intelligence (AI) system offers results that discriminate in terms, for example, of sex, this is usually due to:** • (X) That the data that was used to train the system was not balanced, that is, that much more data was used for men than women, or vice versa. • That the system is designed to be used by men to a greater extent than by women, or vice versa. • That the system itself tried to be sexist. (new item) • I don't know. (*originally: That the developers of the system had sexist biases.*) • (*deleted item: That the system reflects the sexist reality of human nature.*)
Q2	P9	**In which of the following tasks, to be performed by a computer, would it be appropriate to apply machine learning (ML) techniques?** • (X) Recognize if an email is spam (junk mail). • Count the number of times a key is pressed. • Inform about the hours of a certain business based on the day of the week. • I don't know. (*originally: Add large numbers.*)
Q3	P11	**Both Alicia and Robert want to train a machine learning (ML) system that serves to recognize whether a certain text is "happy / positive" or "sad / negative". Alicia and Robert follow two different training strategies. Who of the two will get the better system?** • (X) Alicia. She has compiled 1000 mixed examples of happy / positive texts and another 1000 mixed examples of sad / negative texts. • Robert. He has collected 1000 examples of happy / positive texts and another 10 examples of sad / negative texts. • I don't know.
Q4	P5	**Imagine we implement machine learning (ML) techniques in a text recognition system. We present the computer with a set of sample texts and the computer, after processing, is able to recognize ...** • only the texts that exactly match those examples. • (X) texts similar to those examples (that is, to recognize new texts that it has not seen before). • any text, image or sound that we present to it. • I don't know. (*originally: any text we present to it.*)
Q5	P6	**Which of the following statements is true about machine learning (ML)?** • (X) Training data is essential for machine learning, without data it is not possible to do machine learning. • The more data we use to train a system that incorporates machine learning, the worse (more inaccurate) are the results offered by that system. • Machine learning does not need data to function, precisely because it is automatic and does not depend on being fed data of any kind. • I don't know. (originally: *With automatic learning, computers learn to think and can recognize any type of data (text, image, sound ...), in the same way that a human being does.*)

Name	Orig.	Item and Anchors
Q6	P7	**Which of the following strategies would be more appropriate to teach a computer to recognize the photo of any apple?** • (X) Train the computer with several photos of different apples, taken in different places and contexts. • Train the computer with several similar photos of the same apple, taken in the same place. • Train the computer with several identical copies of the same photo of an apple. • I don't know. (*originally: Train the computer with photos of dogs.*)
Q7	–	**Which of the following datasets is a classic in the machine-learning community and classifying its content correctly can be considered the "Hello World" of deep learning:** • ImageNet • (X) MNIST • Open Images Dataset • I don't know.
Q8	–	**The best tool for attacking visual-classification problems are ...** • (X) convnets, because they work by learning a hierarchy of modular patterns and concepts to represent the visual world, and the representations they learn are easy to inspect. • densely connected layers, because they learn global patterns in their input feature space, which makes them data efficient when processing images. • basic neural networks, because they learn to associate images and labels, and are energy efficient due to their simplistic computational structure. • I don't know.
Q9	–	**For a multilabel classification, the typical choice for a loss function is ...** • MSE • categorical cross entropy. • (X) binary cross entropy. • I don't know.

Table 2. Additional validation questions

Question	Scale
How many courses with a focus on technical AI did you take?	6-point (0 to 5+)
How many courses with a focus on socio-political and/or ethical AI did you take?	6-point (0 to 5+)
In your opinion, how realistic was the scenario?	5-point
How much do you know about the scientific validity of first impressions (based on faces)?	4-point
Did you receive any support for the previous AI quiz? For example, did you consult a search engine (e.g. Google, Bing) or were you helped by nearby friends, colleagues or relatives?	yes/no
Are you currently employed? *exact wording:* yes (IT-related job/company)/ yes (non IT-related job/company)	yes (IT)/ yes (not IT)/ no
How did you learn about this survey? (e.g. which course/ social media/ messaging system)	*open*
Please indicate research field/ study program?	*open*

One participant in the pre-study had taken no AI courses and described him-/herself as a novice with regards to AI knowledge. Another person had taken three technical courses on AI and two socio-political and/or ethical AI courses and rated his/her AI knowledge as intermediate. Another person had attended five technical courses on AI and three socio-political and/or ethical AI courses and rated his/her AI knowledge as advanced. All respondents had a Master's degree. The reported time needed to complete the quiz was 5, 8 and 10 minutes (order unrelated to presented subjects).

Based on respondents' answers on the perceived difficulty of a question (easy, medium, difficult), a difficulty score was calculated. A question received zero difficulty points when being rated as easy, one difficulty point when being rated as medium and two difficulty points when being rated as difficult. The sum total of the scores collected was then divided by the number of participants. Thus, the difficulty score ranges from 0 to 2. Figure 2 displays the questions ordered by their difficulty score.

$$DifficultyScore = \frac{\sum(difficultyPoints)}{N_{\text{respondents}}}$$

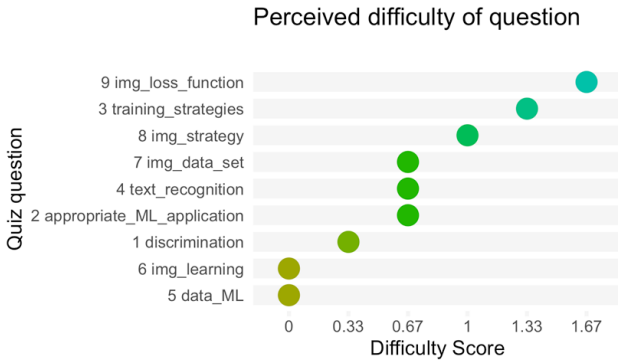


Fig. 2. Perceived difficulty of AI knowledge test question by participants of the pre-study.

People with less knowledge on AI perceived more questions as difficult than people with more knowledge on AI. More specifically, a question that has been perceived as difficult by a respondent with little AI knowledge, was considered as medium by the other two respondents with more AI knowledge. Two questions were rated as easy by all participants: statement about machine learning (training data is essential) and strategy to train an image recognition system (several photos of different apples taken in different places and contexts).

Furthermore, the results from the pre-study hint at a difference in answering behavior, i.e., respondents with a higher self-identified AI knowledge tended to avoid choosing the option "I don't know", and rather risked to select a wrong answer. Instead, the respondent with little AI knowledge tended to select the option "I don't know" more frequently, and in contrast to the other two participants, did not select any incorrect answer.

We observed a positive association between the self-rated AI knowledge and the AI knowledge test. This association is also in line with the number of courses taken, i.e., respondents who took fewer AI courses had fewer correct answers than respondents who took more AI courses.

Overall, the test seems to reflect knowledge on AI. Compared to the self-rated AI knowledge, the AI knowledge test seems to be more objective and less influenced by personal reflections on knowledge or personal characteristics such as diffidence (e.g., one subject had 90% correct answers but indicated to only have intermediate AI knowledge).

A.3.4 AI-competent Dataset. The AI knowledge test was included in the questionnaire when surveying the AI-competent sample. Figure 3 presents the relationships between self-rated AI knowledge, the number of questions in the AI knowledge test answered correctly, and the number of technical courses on AI taken. Figure 3 illustrates that the number of courses taken also influenced self-perception. Participants who attended many courses rated their level of knowledge on average higher than participants who attended fewer courses focusing on technical AI.

Correlations found supported these observations: In order to assess the relationship between the above described AI Knowledge variables, we computed Spearman's rank correlation[1] (not all of the variables were normally distributed). There was a weak positive correlation between the number of correct answers in the AI knowledge test and the self-rated AI knowledge level, r_S=.37, p<.001. There was a moderate positive correlation between the number of correct answers and the number of courses taken on technical AI, r_S=.57, p<.001. There was a strong positive correlation between the self-rated AI knowledge level and the number of courses taken, r_S=.72, p<.001. For this subject pool, we defined participants to be AI-competent when they had correctly answered at least six out of nine questions.

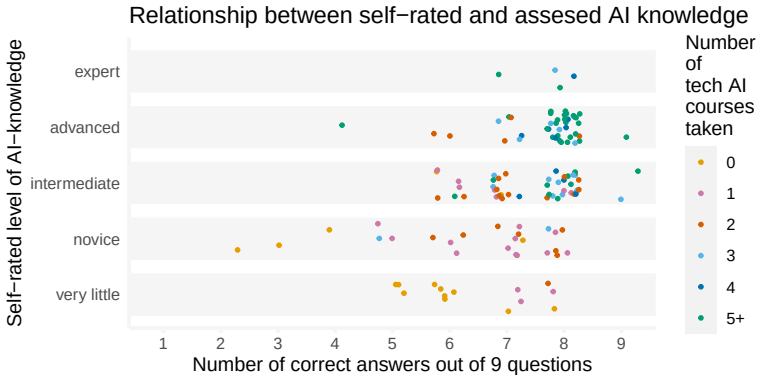


Fig. 3. Knowledge representation based on different measures. The 'number of correct answers' is based on the AI knowledge quiz included in the survey. Participants who did not answer the manipulation check correctly and who consulted external help are not included in the plot. N = 122.

[1]Spearman's rank correlation rho (absolute correlation values): 0-.19: very weak, 20-.39: weak, .40-.59: moderate, .60-.79: strong, .80-1.0: very strong

A.4 Data Cleaning

The AI-competent data sample was cleaned based on the criteria listed in Table 3. Participants who had indicated to have consulted external help for the AI knowledge test were removed from the dataset.

Table 3. Data Cleaning Criteria

	removed cases	N
Original N		160
< 18 years	0	160
Attention check AD	7	153
Attention check HR	14	139
Duration < 120 seconds	0	139
External help	7	132
Low knowledge quiz score	10	122
Final N		**122**

B ANALYSIS OF INFERENCE RATINGS

B.1 Welch Two Sample t-test

The Welch two-sample t-tests produced the following results for the AD context. **Gender** ($mean_{\text{AI-competent}}$ =2.92, $mean_{\text{laypeople}}$ =2.43): t_{Welch} (122.85) =1.72, p>.05, $p_{\text{Bonf.}}$ =0.70, $\hat{g}_{\text{Hedges}}$ =0.30, $CI_{95\%}$ [-0.05,0.65]; **Emotion expression** ($mean_{\text{AI-competent}}$ =2.75, $mean_{\text{laypeople}}$ =3.13): t_{Welch} (120.99) =-1.28, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.23, $CI_{95\%}$ [-0.58,0.12]; **Wearing glasses** ($mean_{\text{AI-competent}}$ =1.67, $mean_{\text{laypeople}}$ =1.57): t_{Welch} (119.85) =0.50, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.09, $CI_{95\%}$ [-0.26,0.44]; **Skin color** ($mean_{\text{AI-competent}}$ =3.22, $mean_{\text{laypeople}}$ =2.65): t_{Welch} (122.67) =1.64, p>.05, $p_{\text{Bonf.}}$ =0.83, $\hat{g}_{\text{Hedges}}$ =0.29, $CI_{95\%}$ [-0.06,0.64]; **Intelligent** ($mean_{\text{AI-competent}}$ =5.60, $mean_{\text{laypeople}}$ =5.58): t_{Welch} (122.38) =0.06, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.01, $CI_{95\%}$ [-0.34,0.36]; **Trustworthy** ($mean_{\text{AI-competent}}$ =5.88, $mean_{\text{laypeople}}$ =5.68): t_{Welch} (121.95) =0.82, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.15, $CI_{95\%}$ [-0.21,0.50]; **Assertive** ($mean_{\text{AI-competent}}$ =4.53, $mean_{\text{laypeople}}$ =5.08): t_{Welch} (120.23) =-1.79, p>.05, $p_{\text{Bonf.}}$ =0.61, $\hat{g}_{\text{Hedges}}$ =-0.32, $CI_{95\%}$ [-0.68,0.04]; **Likable** ($mean_{\text{AI-competent}}$ =4.73, $mean_{\text{laypeople}}$ =5.20): t_{Welch} (120.99) =-1.45, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.26, $CI_{95\%}$ [-0.62,0.10].

The Welch two-sample t-tests produced the following results for the HR context. **Gender** ($mean_{\text{AI-competent}}$ =3.93, $mean_{\text{laypeople}}$ =3.93): t_{Welch} (107.18) =-0.02, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.00, $CI_{95\%}$ [-0.37,0.37]; **Emotion expression** ($mean_{\text{AI-competent}}$ =3.35, $mean_{\text{laypeople}}$ =3.85): t_{Welch} (113.74) =-1.47, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.27, $CI_{95\%}$ [-0.64,0.01]; **Wearing glasses** ($mean_{\text{AI-competent}}$ =2.53, $mean_{\text{laypeople}}$ =3.31): t_{Welch} (112.77) =-1.84, p>.05, $p_{\text{Bonf.}}$ =0.55, $\hat{g}_{\text{Hedges}}$ =-0.34, $CI_{95\%}$ [-0.70,0.03]; **Skin color** ($mean_{\text{AI-competent}}$ =3.98, $mean_{\text{laypeople}}$ =4.10): t_{Welch} (108.79) =-0.27, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =-0.05, $CI_{95\%}$ [-0.42,0.32]; **Intelligent** ($mean_{\text{AI-competent}}$ =6.09, $mean_{\text{laypeople}}$ =5.61): t_{Welch} (111.91) =1.63, p>.05, $p_{\text{Bonf.}}$ =0.85, $\hat{g}_{\text{Hedges}}$ =0.30, $CI_{95\%}$ [-0.07,0.66]; **Trustworthy** ($mean_{\text{AI-competent}}$ =5.93, $mean_{\text{laypeople}}$ =5.20): t_{Welch} (103.70) =2.30, p=.02, $p_{\text{Bonf.}}$ =0.18, $\hat{g}_{\text{Hedges}}$ =0.42, $CI_{95\%}$ [0.05,0.79]; **Assertive** ($mean_{\text{AI-competent}}$ =5.20, $mean_{\text{laypeople}}$ =5.00): t_{Welch} (112.37) =0.63, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.12, $CI_{95\%}$ [-0.25,0.48]; **Likable** ($mean_{\text{AI-competent}}$ =5.40, $mean_{\text{laypeople}}$ =5.26): t_{Welch} (112.72) =0.44, p>.05, $p_{\text{Bonf.}}$ =1, $\hat{g}_{\text{Hedges}}$ =0.08, $CI_{95\%}$ [-0.29,0.45].

B.2 Exploratory Factor Analysis

We conducted an exploratory factor analysis on the eight inferences (items) with orthogonal rotation (varimax) for each of the three samples. For the analysis, cases with missing values, i.e., "Can't Answer" responses, were removed from all three samples, which reduced the sample size for the laypeople sample N=118, for the AI-competent sample to N=112, and for laypeople validation sample to N=91. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis for the laypeople (MT) sample KMO = 0.76, for the AI-competent sample KMO = 0.75, and for the laypeople validation sample KMO = 0.68. All KMO values for individual inferences were ≥ 0.70 for laypeople (MT) sample, ≥ 0.68 for the AI-competent sample, and ≥ 0.64 for the laypeople validation sample. Hence, all values were above the acceptable limit of 0.5 [5, 7, 8]. Bartlett's test of sphericity indicated that correlations between inferences were sufficiently large for the laypeople (MT) sample $\chi^2(28) = 283.9352$, $p < .001$, the AI-competent sample $\chi^2(28) = 227.8268$, $p < .001$, and the laypeople validation sample $\chi^2(28) = 192.1025$, $p < .001$ [5].

Multiple criteria for the identification of the number of factors to extract suggested two factors. For examples, for all three samples two factors had eigenvalues over Kaiser's criterion of 1. The scree plot, very simple structure of complexity 1, as well as the Velicer MAP all suggested two factors for all of the three samples. Given these analysis, we extracted two factors in the final analysis. For all three samples, oblique rotation resulted in factors with correlations <.32 [15], yet the same pattern structure. Hence, orthogonal rotation was chosen. Table 20 shows the factor loadings after rotation for all of the three samples separately. It should be noted that some factor loadings do not exceed .6 and our sample size is rather small [6].

We performed robustness checks with sub-samples of 85% of the data. The results from the robustness checks validate the findings from the main analysis. However, the solutions were not always stable. Some items loaded on two factors, and hence, did not achieve simple structure. This is because there are variables with loadings >.3 on more than one factor [16], e.g., *gender* or *wearing glasses*.

While small factor loadings and unstable factor solutions during the robustness check suggest that the interpretation of the factor analyses should be considered with caution, the structure of the factor loadings replicates findings from [3]. We assume that both, small factor loadings and the lack of simple structure, emerge from the small sample size.

Table 4. Exploratory factor analysis for all three samples: Varimax rotated factor loadings

	Laypeople (MT)		**AI-competent**		**Laypeople (Validation)**	
	Character and personality	Social constructs and features	Character and personality	Social constructs and features	Character and personality	Social constructs and features
gender	-0.01	0.53	0.36	0.67	0.07	0.53
emotion expression	0.15	0.53	0.18	0.49	0.24	0.42
wearing glasses	-0.29	0.75	-0.09	0.64	0	0.76
skin color	-0.13	0.8	0.02	0.68	-0.05	0.83
intelligent	0.69	-0.07	0.6	0.05	0.7	-0.1
trustworthy	0.74	-0.15	0.8	-0.14	0.8	-0.02
assertive	0.72	0.08	0.7	0.2	0.64	0.16
likable	0.69	-0.05	0.56	0.22	0.53	0.22
Eigenvalues	2.17	1.81	1.98	1.67	1.88	1.82
% of variance	0.27	0.23	0.25	0.21	0.23	0.23
α	0.81	0.75	0.75	0.71	0.75	0.73

B.3 ANOVAs for each of the inferences

B.3.1 AI-competent vs. MTurk Laypeople Sample. Table 5 to Table 12 present the results from the Bonferroni corrected ANOVAs for each of the eight inferences.[2]

Using Pillai's trace, there were significant main effects at an α-level of 0.05 for *context* on the inference ratings for gender, emotion expression, wearing glasses and skin color. There were no other significant effects.

B.3.2 AI-competent vs. Validation Laypeople Sample. Table 13 to Table 20 present the results from the Bonferroni corrected validation ANOVAs for each of the eight inferences. For this comparison, we were able to include participant's information on whether they have a job (no; yes, IT-related; yes, not IT related).

Using Pillai's trace, there were significant main effects at an α-level of 0.05 for *context* on the inference ratings for gender, wearing glasses and skin color, but not for emotion expression. There were no other significant effects.

[2]Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5. ANOVA for inference: gender

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	82.63	1	23.60	0.000	0.000	***
gender	6.57	2	0.94	0.393	1.000	
age	16.28	5	0.93	0.463	1.000	
education	15.27	7	0.62	0.736	1.000	
country	105.05	27	1.11	0.330	1.000	
sample	0.43	1	0.12	0.725	1.000	
Residuals	689.67	197				

Table 6. ANOVA for inference: emotion expression

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	28.22	1	9.34	0.003	0.015	*
gender	2.25	2	0.37	0.689	1.000	
age	15.40	5	1.02	0.407	1.000	
education	47.21	7	2.23	0.033	0.199	
country	70.14	27	0.86	0.669	1.000	
sample	1.24	1	0.41	0.523	1.000	
Residuals	598.40	198				

Table 7. ANOVA for inference: wearing glasses

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	85.37	1	26.31	0.000	0.000	***
gender	1.43	2	0.22	0.803	1.000	
age	5.65	5	0.35	0.883	1.000	
education	5.64	7	0.25	0.972	1.000	
country	99.93	27	1.14	0.297	1.000	
sample	0.74	1	0.23	0.633	1.000	
Residuals	645.73	199				

Table 8. ANOVA for inference: skin color

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	68.18	1	14.17	0.000	0.001	**
gender	4.36	2	0.45	0.637	1.000	
age	16.34	5	0.68	0.640	1.000	
education	17.33	7	0.51	0.823	1.000	
country	112.15	27	0.86	0.664	1.000	
sample	0.00	1	0.00	0.984	1.000	
Residuals	933.58	194				

Table 9. ANOVA for inference: intelligent

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	5.41	1	2.11	0.148	0.889
gender	0.01	2	0.00	0.999	1.000
age	28.72	5	2.24	0.052	0.311
education	9.67	7	0.54	0.805	1.000
country	94.04	26	1.41	0.099	0.593
sample	4.11	1	1.60	0.207	1.000
Residuals	507.89	198			

Table 10. ANOVA for inference: trustworthy

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	2.07	1	0.86	0.356	1.000
gender	4.23	2	0.87	0.419	1.000
age	15.44	5	1.28	0.275	1.000
education	5.93	7	0.35	0.929	1.000
country	56.01	25	0.93	0.568	1.000
sample	0.61	1	0.25	0.616	1.000
Residuals	478.66	198			

Table 11. ANOVA for inference: assertive

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	3.76	1	1.24	0.266	1.000
gender	6.14	2	1.01	0.364	1.000
age	9.43	5	0.62	0.682	1.000
education	14.97	7	0.71	0.666	1.000
country	77.99	25	1.03	0.429	1.000
sample	14.54	1	4.80	0.030	0.177
Residuals	593.24	196			

Table 12. ANOVA for inference: likable

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	7.75	1	2.42	0.121	0.728
gender	3.75	2	0.59	0.558	1.000
age	17.35	5	1.08	0.371	1.000
education	20.61	7	0.92	0.492	1.000
country	48.06	26	0.58	0.951	1.000
sample	4.33	1	1.35	0.246	1.000
Residuals	627.60	196			

Table 13. Validation ANOVA for inference: gender

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	70.24	1	19.43	0.000	0.000	***
gender	1.61	2	0.22	0.800	1.000	
age	11.78	4	0.81	0.518	1.000	
education	11.27	6	0.52	0.793	1.000	
country	158.26	35	1.25	0.177	1.000	
student job	7.36	2	1.02	0.364	1.000	
sample	0.25	1	0.07	0.794	1.000	
Residuals	611.07	169				

Table 14. Validation ANOVA for inference: emotion expression

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	3.70	1	1.37	0.244	1.000
gender	0.87	2	0.16	0.851	1.000
age	20.24	4	1.87	0.118	0.824
education	34.05	6	2.10	0.056	0.390
country	127.58	35	1.35	0.110	0.767
student job	1.63	2	0.30	0.740	1.000
sample	0.01	1	0.00	0.959	1.000
Residuals	454.11	168			

Table 15. Validation ANOVA for inference: wearing glasses

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	30.60	1	8.59	0.004	0.027	*
gender	2.16	2	0.30	0.738	1.000	
age	14.38	4	1.01	0.404	1.000	
education	3.84	6	0.18	0.982	1.000	
country	93.19	35	0.75	0.844	1.000	
student job	0.89	2	0.13	0.882	1.000	
sample	4.58	1	1.29	0.258	1.000	
Residuals	601.86	169				

Table 16. Validation ANOVA for inference: skin color

	Sum Sq	Df	F	Pr(>F)	Bonf.	
context	38.58	1	7.80	0.006	0.041	*
gender	13.63	2	1.38	0.255	1.000	
age	11.97	4	0.61	0.659	1.000	
education	13.69	6	0.46	0.836	1.000	
country	178.70	34	1.06	0.386	1.000	
student job	1.61	2	0.16	0.850	1.000	
sample	0.01	1	0.00	0.971	1.000	
Residuals	825.81	167				

Table 17. Validation ANOVA for inference: intelligent

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	14.11	1	5.86	0.017	0.116
gender	0.36	2	0.07	0.928	1.000
age	9.09	4	0.94	0.441	1.000
education	3.78	6	0.26	0.954	1.000
country	90.77	34	1.11	0.327	1.000
student job	6.16	2	1.28	0.281	1.000
sample	5.18	1	2.15	0.144	1.000
Residuals	409.63	170			

Table 18. Validation ANOVA for inference: trustworthy

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	0.00	1	0.00	0.962	1.000
gender	4.00	2	1.02	0.362	1.000
age	2.36	4	0.30	0.877	1.000
education	18.30	6	1.56	0.162	1.000
country	60.96	33	0.94	0.560	1.000
student job	9.12	2	2.33	0.100	0.703
sample	0.85	1	0.43	0.511	1.000
Residuals	330.78	169			

Table 19. Validation ANOVA for inference: assertive

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	9.05	1	3.31	0.071	0.496
gender	5.93	2	1.08	0.341	1.000
age	2.87	4	0.26	0.902	1.000
education	17.37	6	1.06	0.390	1.000
country	104.36	33	1.16	0.273	1.000
student job	17.97	2	3.28	0.040	0.280
sample	8.13	1	2.97	0.087	0.607
Residuals	459.79	168			

Table 20. Validation ANOVA for inference: likable

	Sum Sq	Df	F	Pr(>F)	Bonf.
context	11.75	1	3.50	0.063	0.443
gender	2.54	2	0.38	0.686	1.000
age	5.22	4	0.39	0.817	1.000
education	43.68	6	2.17	0.049	0.341
country	63.58	34	0.56	0.977	1.000
student job	5.70	2	0.85	0.430	1.000
sample	4.93	1	1.46	0.228	1.000
Residuals	571.63	170			

C CODE BOOK

This code book was provided to all researchers who were involved in the process of labeling the datasets. It provides information on the context of the data and guidelines on how to label the data. Figure 1 was added for better understanding of the survey scenarios.

C.1 General Notes and Background Information on the Study

C.1.1 Study Description. The comments to be categorized originate from surveys on the perception of AI inference-making in the context of advertisement and in the context of hiring. After rating how much a participant agrees or disagrees with a certain inference made by a software application using AI, the participant was asked to justify his/her rating in one to two sentences. In total, the participant was asked to repeat this process for eight different inferences: *gender*, *skin color*, *wearing glasses*, *emotion expression*, *intelligent*, *trustworthy*, *assertive*, *likable*.

C.1.2 Experimental Set-up. One participant was either drawn into the context of advertisement (AD) or into the context of hiring (HR). Figure 1 contains examples of two scenarios shown to two different participants (one drawn into the AD context and the other drawn into the HR context).

C.2 Coding Instructions

C.2.1 Case-wise analysis. The answers to the open questions are analyzed case-wise, i.e., one respondent at a time. Given partially very little text per answer and occasional references to previous answers, a case-wise coding of all answers per participant ensures the preservation of participant-based contextual information.

C.2.2 Scope of material. The **unit of evaluation** corresponds to all justification texts by respondents in the samples. The justification texts are answers to eight brief open questions in the survey.

The **unit of context** determines the material that can be consulted for coding. In this study a participant may reference a previous answer; hence, the context unit equals all responses from one participant.

The **unit of coding** resembles the minimal textual element that can be assigned to one category; here, parts of one sentence of a response from one study participant.

C.2.3 Repeated information and multiple codes per justification. Multiple codes can be assigned to one justification of a participant. This approach allows accounting for complex justification patterns, where participants discuss different topics within one comment. The rating is to be considered when assigning a code, because it usually helps understanding the justification better.

C.2.4 Missing responses. Some respondents did not justify their rating ("NA") or wrote, e.g., "None". In these cases, the theme "no justification" is assigned.

C.3 Categories

The following Table 21 summarizes all categories, gives definitions as well as application descriptions and differentiates the categories from related ones. Examples are provided.

Table 21. Code Book: Definition of categories and examples.

Category	Description	Example
AI (in)ability		
technical ability of AI	**Definition:** The AI has the technical ability to draw an inference. **Application:** This code is applied when a participant gives a specific explanation to why s/he believes the AI to be able to draw the inference.	*Obviously an AI can identify the shade of skin | Machine learning can be used to determine whether or not the person is expressing an emotion.*
works well/ accurately	**Definition:** The AI accurately draws a certain inference. This task is known to work well. **Application:** This code is used when a participant highlights high accuracy scores for a specific task or mentions that a specific task is regularly and successfully solved by the AI system.	*Yes, would work, as it's already being done | emotion regonition based on facial expression is a very popular AI Task [sic] | solid results can be achieved by an AI*
easy to infer	**Definition:** The inference task is easy to solve for an AI system. **Application:** This code is used when a participant highlights that a certain inference can easily be drawn by an AI system.	*This is something that should be easy for an AI to determine. | the gender of a person based on a picture is in itself a relative easy classification task*
can infer most/sometimes/ difficult in some situations	**Definition:** The AI systems can most times or sometimes draw the inference. However, for specific cases, such as the gender "other", a correct inference is difficult or accompanied by (many) mistakes. **Application:** This code is assigned when a participants highlights that the system will make mistakes for some cases, e.g, for "other".	*except for very small amount of situations like trans, gender can be analysed easily by ai | The vast majority of men and women have features that make their gender clearly identifiable. However, gender-neutral persons or people who simply don't look like or conform to a gender would be difficult | it won't be perfect as people express emotions differently*
difficult/ not possible to infer	**Definition:** The inference task is difficult or impossible to be solved for an AI system. **Application:** This code is used when a participant highlights that a certain inference is impossible or difficult for an AI system.	*Impossible to infer | Too complex a notion to quantify, even for humans | Hard for AI to decide.*
Inference task		
inference is objective	**Definition:** The inference task is objective. **Application:** This code is used when a participant highlights that the evaluation of the inference is not dependent on the observer or specifically uses the word "objective".	*This can also be easily and objectively answered by an ML algorithm. | can be determined objectively*
inference is subjective	**Definition:** The inference task is subjective. **Application:** This code is used when a participant highlights that the evaluation of the inference is dependent on the observer or specifically uses the word "subjective".	*trustworthy is a subjective trait | An image can not show whether a person is likeable or not. Likeability is largely subjective and can not be judged in objective terms.*

Category	Description	Example
Reference to data		
indicative distinct facial/visual features	**Definition:** Distinct facial tendencies or features indicate a certain inference. **Application:** The code is used when a participant highlights certain facial properties as key for drawing a correct inference.	*Learning to recognize the traits that show specific emotions is possible | I supposed a computer can be programmed to detect certain facial tendencies that express emotions*
profile picture good evidence (data)	**Definition:** The profile picture provides good evidence for an AI system to draw a certain inference. **Application:** The code is used when a participant highlights that the inference can be drawn based on the provided data, here, the profile picture, or comments that a certain inference can be "seen".	*for the most part it's fairly obvious to see if someone is expressing anger, hate, love, etc. | visually can be determined by the AI | It seems reasonable that an AI could figure out whether someone is wearing glasses by their picture.*
data quality/ variety of high relevance	**Definition:** Data quality, including a varied dataset, is of high importance to train an AI system to draw specific inferences. **Application:** The code is used when a participant highlights that the success of the AI system is based on the quality of the data. This code is also used when a participant mentions that the data can be manipulated in such a way that it is difficult to draw correct inferences.	*diversity on the dataset would be required to make sure no skin tone under different lighting is left out | If there's valid, reliable data to support it, it's reasonable | If the data is valid and reliable, it's reasonable.*
not sufficient/ good evidence (data) for task	**Definition:** The profile picture does not provide sufficient or good evidence for an AI system to draw a certain inference. **Application:** The code is used when a participant highlights that further data or different data would be required to properly draw the inference, e.g., because the image only captures a single moment. This code is also used when it is mentioned that facial expressions do not resemble how a person actually feels or what they identify with.	*A personality cannot be inferred from facial traits. It is inferred by actions, which cannot be shown in a profile pic | There are numerous of people that tend to hid their emotions through pictures and everyday lifestyle but end up taking their own lives. Nothing looks like it seems.* [sic] *| [it's no] real indicator whether they are nice.*
Reference to (ir)relevance of inference for purpose of AI system		
inference relevant	**Definition:** The inference is relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Application:** The code is used when a participant highlights that drawing an inference is useful/helpful/relevant to the purpose of the AI system.	*I agree that different genders need different products and services, so this would be reasonable | There are products that target just men or just women so i can see this being helpful*
inference sometimes relevant	**Definition:** The inference is (only) sometimes relevant to the decision of the AI, e.g. advertisement choice or applicant selection. **Application:** The code is used when a participant highlights that drawing an inference is not always, but only sometimes, useful/helpful/relevant to the purpose of the AI system.	*There are cases where a certain gender could be preferable to another (babysitters, private tutors), but there are also cases where this distinction does not matter (corporate jobs, waiters, sellers).*
inference not relevant	**Definition:** The inference is not relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Application:** The code is used when a subject highlights that drawing the inference is not useful/helpful/relevant or does not have anything to do with the purpose of the AI system.	*does not seem like a valueable information. | skin color doesn't influence a persons consumer behavior* [sic]

Category	Description	Example
Ethics and Norms		
ethically questionable/ should not matter/ should not be used	**Definition:** Drawing the inference is ethically questionable or should not matter. An inference should not be used to make subsequent decisions. **Application:** The code is used when a participant highlights or critiques drawing a specific inference. Some participants stress that an inference should not matter for a decision made by an AI or that an inference, even when drawn, should not further be used in subsequent AI decision-making.	*It should not matter for the job | I don't think that the AI should consider race or skin color when deciding what advertisements to show people | I think this is a bit too touchy of a subject due to being politically correct is very important at this time on history*
bias/ stereotypes/ discrimination	**Definition:** Drawing the specific inference leads to bias or discrimination. Making decisions based on the inference is based on stereotypes. **Application:** The code is used when a participant highlights or critiques drawing specific inferences because the resulting AI decision-making would be biased, be based on stereotypes or discriminate.	*I think this can be racist. | This will end badly, if white people are more likable than black people. Won't to that.* [sic]
binary gender system not appropriate	**Definition:** A binary concept of the inference gender is not appropriate and does not reflect today's society. **Application:** The code is used when a participant highlights or critiques drawing the inference gender based only on two categories, females and males.	*Gender norms are a thing of past! | gender is a fluid concept | many people do not identify with their sex and birth or with a binary gender system which may lead to incorrect classification*
Comparison to human		
easy for human to identify	**Definition:** The inference task is easy to solve for a human. **Application:** This code is used when a participant makes a comparison to a human being and highlights that a certain inference can easily be drawn by a human.	*This is an easy task even for a human*
difficult for human to identify	**Definition:** The inference task is difficult to solve for a human. **Application:** This code is used when a participant makes a comparison to a human being and highlights that a certain inference is difficult to be drawn by a human.	*Disagree, because its also hard for humans to guess that from experience | there are some difficulties (androgynous), which even humans have problems with*
Miscellaneous		
person not sure/ indecisive	**Definition:** A respondent is indecisive whether to agree or disagree and/or does not have any opinion. **Application:** This category is assigned if a respondent is unsure. In such cases the respondent gave a rating "4", i.e. neither agree nor disagree. Occasionally, the text-field is left empty.	*What is skin color? Do you mean race? What are the categories? | Not sure if a single picture can determine the assertiveness of a person. People generally put a brave face on social media.*
no justification	**Definition:** A respondent did not provide an open-text response.	*NA | none*

D ANALYSIS OF JUSTIFICATION THEMES

D.1 Comparison of Usage of Justification Themes

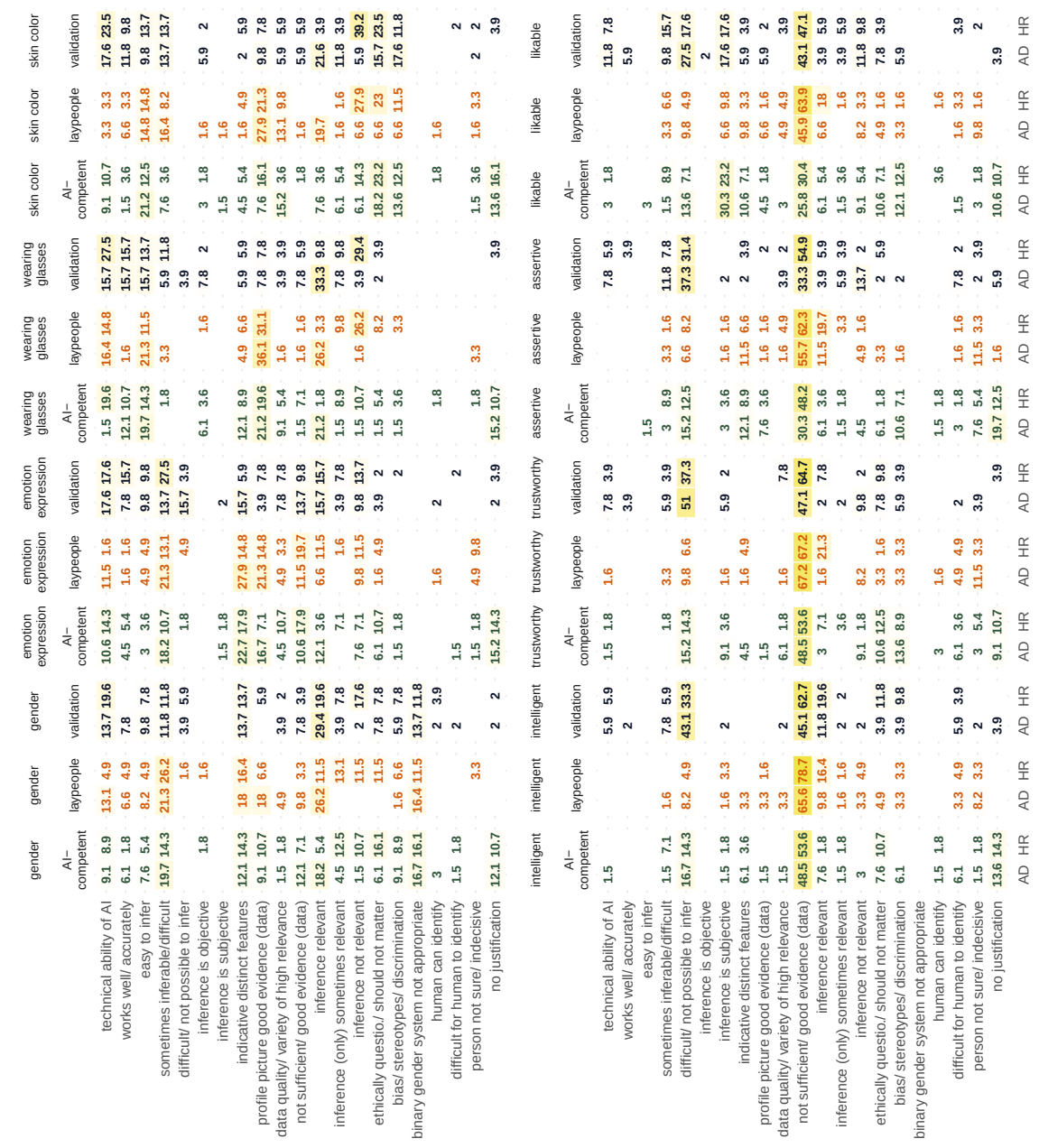


Fig. 4. Percentages of justification themes used by participants for each inference (number of participants per context and inference as baseline). Each column adds up to more than 100%, because participants used up to four themes in one justification.

D.2 Co-occurrence Analysis: AI (in)ability and data-related themes are the themes most frequently used in combination with other themes.

We analyzed the co-occurrence of themes with each other to identify patterns in the use of multiple justification themes. Figure 5a) depicts the frequencies of two themes used in combination. Please note that this analysis includes only the AI-competent and laypeople (MT) sample, and refers to all justifications containing two or more themes, i.e. 20% to 35% of justifications (see Table 1 in main article). Figure 5b) illustrates networks of the co-occurrences of themes by sample and by inference type.

D.2.1 For inferences on 'constructs and features', the AI-competent raise concerns but recognize AI to be able to make certain inferences. Referring to inferences on *constructs and features*, people with AI-competence raise "ethical and discriminatory concerns" in combination with almost all other justification themes, however, most frequently in combinations with themes on "AI ability" (11.5%) or the "sufficiency" of the profile picture as evidence (10%; Figure 5a) and 5b-1)). The "sufficiency" of the profile picture as evidence is also often mentioned in combinations with themes on AI ability (10%).

This relationship changes for justifications on the inferences on *character and personality traits*. "Ethical and discriminatory concerns" are most frequently brought forward in combination with themes on the "insufficiency" of a profile picture as evidence (11.4%; Figure 5a) and 5b-3)).

D.2.2 For inferences on 'character and personality traits', laypeople often pair comments on the (in)sufficiency or (in)adequacy of the data with another theme. Referring to inferences on *constructs and features*, laypeople highlight the "sufficiency" of the data in combination with comments on the "ability" (17.2%) and "inability" of AI (15.2%) (Figure 5a) and 5b-2)). With reference to the inferences on *character and personality traits*, laypeople frequently mention themes related to "insufficiency" of the data in combination with "inability" of AI to make such an inference (12.5%), "ethical and discriminatory concerns" (15.5%), and "comparison(s) to human" abilities (12.5%). For instance, a comment on the inference *assertive* states: "I can't see how even a person could determine this from a picture." However, the "insufficiency" of the data is also frequently mentioned in combination with the "relevance" of the inference (10.6%), e.g., a comment on the inference *intelligent* states: "You want to hire smart people but i dont think that can be analyzed from a photo" [sic] (Figure 5a) and b-4)).

D.2.3 More theme combinations are used to explain ratings on inferences referring to 'constructs and features' and by individuals with AI-competence. Generally, a greater variety of combinations are used to justify ratings on inference ratings referring to *constructs and features*. This applies to both main samples (see Figure 5a), 5b-1) and b-2)). Figure 5a), 5b-3) and b-4) show a smaller variety of theme combinations for justifications on inference ratings referring to *character and personality traits*. This implies that opinions on inferences referring to *character and personality traits* are clearer. In contrast, opinions on inferences referring to *constructs and features* are more varied. People with AI competence use a greater variety of theme combinations than laypeople (MT) for both types of combinations (Figure 5a), 5b-1) and 5b-3)).

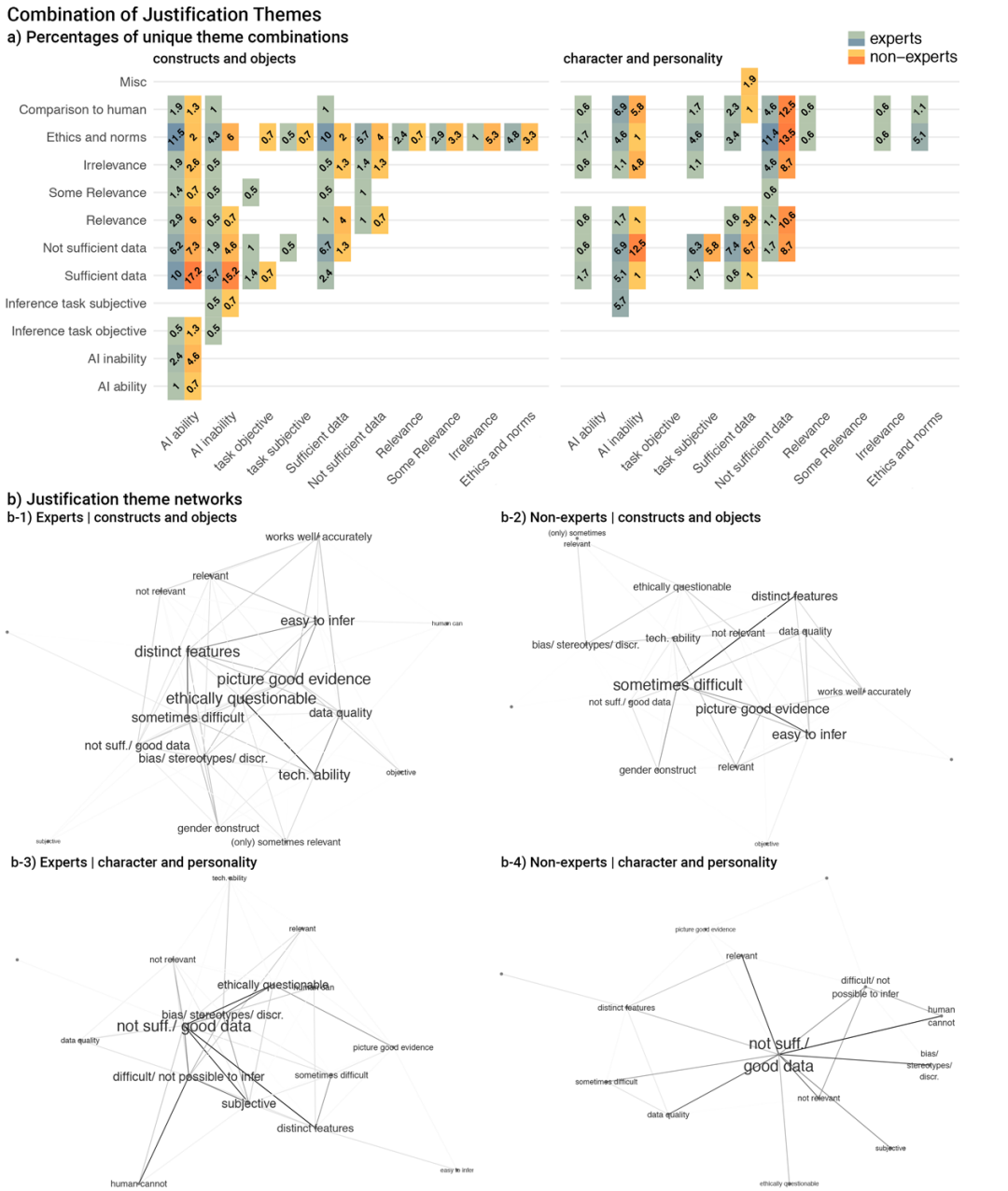


Fig. 5. Combination of justification themes by inference type and sample. a) Unique combinations of two themes (i.e. super-ordinate theme topic) by inference type and sample. Analysis refers to justifications containing more than one theme. b) Network analysis of co-occurrences of themes. We calculated undirected weighted one-mode networks.

REFERENCES

[1] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[2] François Chollet. 2018. *Deep Learning with Python.* Manning Publications, Shelter Island, NY, USA. https://books.google.de/books?id=mjVKEAAAQBAJ

[3] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What people think AI should infer from faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 128–141. https://doi.org/10.1145/3531146.3533080

[4] Julian Estevez, Gorka Garate, and Manuel Graña. 2019. Gentle introduction to artificial intelligence for high-school students using scratch. *IEEE Access* 7 (2019), 179027–179036. https://doi.org/10.1109/ACCESS.2019.2956136

[5] Andy Fields, Jeremy Miles, and Zoë Fields. 2012. *Discovering Statistics Using R.* Sage Publishing, London.

[6] Edward Guadagnoli and Wayne F Velicer. 1988. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103, 2 (1988), 265—-275. https://doi.org/10.1037/0033-2909.103.2.265

[7] Graeme D Hutcheson and Nick Sofroniou. 1999. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models.* Sage Publications, New York City, NY, USA.

[8] Henry F Kaiser. 1974. An index of factorial simplicity. *Psychometrika* 39 (1974), 31–36. https://doi.org/10.1007/BF02291575

[9] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 690–700. https://doi.org/10.1145/3442188.3445931

[10] Esther Kaufmann. 2021. Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence* 2 (2021), 100028. https://doi.org/10.1016/j.caeai.2021.100028

[11] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[12] Emma Pierson. 2018. Demographics and discussion influence views on algorithmic fairness. arXiv:1712.09124 [cs.CY]

[13] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3408877.3432393

[14] Lisa M Schwartz, Steven Woloshin, William C Black, and H Gilbert Welch. 1997. The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine* 127, 11 (1997), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

[15] Barbara G Tabachnick and Linda S Fidell. 2013. *Using Multivariate Statistics.* Pearson Education.

[16] Louis Leon Thurstone. 1947. Multiple-factor analysis; a development and expansion of The Vectors of Mind. (1947).

[17] Ansgar Zerfass, Jens Hagelstein, and Ralph Tench. 2020. Artificial intelligence in communication management: A cross-national study on adoption and knowledge, impact, challenges and risks. *Journal of Communication Management* (2020). https://doi.org/10.1108/JCOM-10-2019-0137

[18] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C Horowitz, and Allan Dafoe. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research* 71 (2021), 591–666. https://doi.org/10.1613/jair.1.12895

List of Figures

List of Tables

Bibliography

[1] G. C. Bowker and S. L. Star. *Sorting things out: Classification and its consequences*. MIT Press, 2000.

[2] J. C. Scott. "Seeing like a state". In: *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 2008.

[3] J.-E. Mai. "The modernity of classification". In: *Journal of Documentation* 67.4 (2011), pp. 710–730.

[4] J. R. Beniger and D. L. Robyn. "Quantitative graphics in statistics: A brief history". In: *The American Statistician* 32.1 (1978), pp. 1–11.

[5] W. N. Espeland and M. L. Stevens. "A sociology of quantification". In: *European Journal of Sociology/Archives Européennes de Sociologie* 49.3 (2008), pp. 401–436.

[6] T. J. Misa, P. Brey, and A. Feenberg. *Modernity and technology*. MIT Press, 2003.

[7] S. H. Kim. "Max Weber". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University, 2021.

[8] S. Kalberg. "Max Weber's types of rationality: Cornerstones for the analysis of rationalization processes in history". In: *American Journal of Sociology* 85.5 (1980), pp. 1145–1179.

[9] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. "Algorithmic fairness: Choices, assumptions, and definitions". In: *Annual Review of Statistics and Its Application* 8 (2021), pp. 141–163.

[10] D. Majstorovic, M. A. Wimmer, R. Lay-Yee, P. Davis, and P. Ahrweiler. "Features and added value of simulation models using different modelling approaches supporting policy-making: A comparative analysis". In: *Policy Practice and Digital Science*. Springer, 2015, pp. 95–123.

[11] G. F. Khan, B. Swar, and S. K. Lee. "Social media risks and benefits: A public sector perspective". In: *Social Science Computer Review* 32.5 (2014), pp. 606–627.

[12] Y. T. Uhls, N. B. Ellison, and K. Subrahmanyam. "Benefits and costs of social media in adolescence". In: *Pediatrics* 140.Supplement_2 (2017), S67–S70.

[13] N. Ellison and d. boyd. "Sociality through social network sites". In: *The Oxford Handbook of Internet Studies*. Ed. by W. H. Dutton. Oxford University Press, 2013, pp. 151–172.

[14] C. H. Smith. "Corporatised Identities≠ Digital Identities: Algorithmic Filtering on Social Media and the Commercialisation of Presentations of Self". In: *Ethics of Digital Well-Being*. Springer, 2020, pp. 55–80.

[15] S. Engelmann, V. Scheibe, F. Battaglia, and J. Grossklags. "Social Media Profiling Continues to Partake in the Development of Formalistic Self-Concepts. Social Media Users Think So, Too." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 238–252.

[16] D. Evans, S. Bratton, and J. McKee. *Social media marketing*. AG Printing & Publishing, 2021.

[17] Y. K. Dwivedi, K. K. Kapoor, and H. Chen. "Social media marketing and advertising". In: *The Marketing Review* 15.3 (2015), pp. 289–309.

[18] S. Engelmann, J. Grossklags, and L. Herzog. "Should users participate in governing social media? Philosophical and technical considerations of democratic social media". In: *First Monday* (2020).

[19] S. Engelmann, J. Grossklags, and O. Papakyriakopoulos. "A democracy called Facebook? Participation as a privacy strategy on social media". In: *Privacy Technologies and Policy: 6th Annual Privacy Forum, APF 2018, Barcelona, Spain, June 13-14, 2018, Revised Selected Papers 6*. Springer. 2018, pp. 91–108.

[20] A. Mager. "Algorithmic ideology: How capitalist society shapes search engines". In: *Information, Communication & Society* 15.5 (2012), pp. 769–787.

[21] D. H. Park, H. K. Kim, I. Y. Choi, and J. K. Kim. "A literature review and classification of recommender systems research". In: *Expert Systems with Applications* 39.11 (2012), pp. 10059–10072.

[22] Q. Zhang, J. Lu, and Y. Jin. "Artificial intelligence in recommender systems". In: *Complex & Intelligent Systems* 7.1 (2021), pp. 439–457.

[23] C. Onay and E. Öztürk. "A review of credit scoring research in the age of Big Data". In: *Journal of Financial Regulation and Compliance* 26.3 (2018), pp. 382–405.

[24] T. Berg, V. Burg, A. Gombović, and M. Puri. "On the rise of fintechs: Credit scoring using digital footprints". In: *The Review of Financial Studies* 33.7 (2020), pp. 2845–2897.

[25] L. Gambacorta, Y. Huang, H. Qiu, and J. Wang. *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm*. BIS Working Paper. 2019.

[26] N. Aggarwal. "The norms of algorithmic credit scoring". In: *The Cambridge Law Journal* 80.1 (2021), pp. 42–73.

[27] J. Isaak and M. J. Hanna. "User data privacy: Facebook, Cambridge Analytica, and privacy protection". In: *Computer* 51.8 (2018), pp. 56–59.

[28] A. N. Novak and M. O. Vilceanu. ""The internet is not pleased": twitter and the 2017 Equifax data breach". In: *The Communication Review* 22.3 (2019), pp. 196–221.

[29] J. Verble. "The NSA and Edward Snowden: surveillance in the 21st century". In: *ACM Sigcas Computers and Society* 44.3 (2014), pp. 14–20.

[30] J. X. Chen. "The evolution of computing: AlphaGo". In: *Computing in Science & Engineering* 18.4 (2016), pp. 4–7.

[31] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements". In: *Psychological Science in the Public Interest* 20.1 (2019), pp. 1–68.

[32] A. Strasser, M. Crosby, and E. Schwitzgebel. "Limits and risks of GPT-3 applications". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 44. 44. 2022.

[33] A. Granulo, C. Fuchs, and S. Puntoni. "Psychological reactions to human versus robotic job replacement". In: *Nature Human Behaviour* 3.10 (2019), pp. 1062–1069. DOI: https://doi.org/10.1038/s41562-019-0670-y.

[34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.

[35] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. self-published; fairmlbook.org, 2019.

[36] S. Harding. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press, 1991.

[37] M. Fourcade and K. Healy. "Seeing like a market". In: *Socio-economic Review* 15.1 (2017), pp. 9–29.

[38] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove. "Potential for discrimination in online targeted advertising". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 5–19.

[39] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. "Social media, political polarization, and political disinformation: A review of the scientific literature". In: *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[40] A. Guiora and E. A. Park. "Hate speech on social media". In: *Philosophia* 45.3 (2017), pp. 957–971.

[41] M. K. Scheuerman, A. Hanna, and E. Denton. "Do datasets have politics? disciplinary values in computer vision dataset development". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–37.

[42] C. Y. Olivola, F. Funk, and A. Todorov. "Social attributions from faces bias human choices". In: *Trends in Cognitive Sciences* 18.11 (2014), pp. 566–570. DOI: 10.1016/j.tics.2014.09.007.

[43] G. S. Lenz and C. Lawson. "Looking the part: Television leads less informed citizens to vote based on candidates' appearance". In: *American Journal of Political Science* 55.3 (2011), pp. 574–589. DOI: 10.1111/j.1540-5907.2011.00511.x.

[44] A. Todorov. *Face Value: The Irresistible Influence of First Impressions*. Princeton University Press, 2017.

[45] I. S. Penton-Voak, N. Pound, A. C. Little, and D. I. Perrett. "Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception". In: *Social Cognition* 24.5 (2006), pp. 607–640. DOI: 10.1521/soco.2006.24.5.607.

[46] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao. "A facial expression emotion recognition based human-robot interaction system". In: *IEEE/CAA Journal of Automatica Sinica* 4.4 (2017), pp. 668–676.

[47] D. K. Mulligan, C. Koopman, and N. Doty. "Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083 (2016), Article No. 20160118. DOI: https://doi.org/10.1098/rsta.2016.0118.

[48] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.

[49] S. Barocas and A. D. Selbst. "Big data's disparate impact". In: *Calif. L. Rev.* 104 (2016), p. 671.

[50] B. Friedman and H. Nissenbaum. "Bias in computer systems". In: *ACM Transactions on Information Systems (TOIS)* 14.3 (1996), pp. 330–347.

[51] D. Pedreshi, S. Ruggieri, and F. Turini. "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008, pp. 560–568.

[52] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments". In: *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.

[53] J. Buolamwini and T. Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 77–91.

[54] J. Kleinberg, S. Mullainathan, and M. Raghavan. *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807. 2016.

[55] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. "Machine bias". In: *Ethics of Data and Analytics: Concepts and Cases*. Ed. by K. Martin. Auerbach Publications, 2022, pp. 254–264.

[56] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan. "Robust optimization for fairness with noisy protected groups". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5190–5203.

[57] W. Sinnott-Armstrong. "Consequentialism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University, 2021.

[58] C. Dwork. "Differential privacy: A survey of results". In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19.

[59] B. Mittelstadt. "Principles alone cannot guarantee ethical AI". In: *Nature Machine Intelligence* 1.11 (2019), pp. 501–507.

[60] T. Hagendorff. "The ethics of AI ethics: An evaluation of guidelines". In: *Minds and Machines* 30.1 (2020), pp. 99–120.

[61] A. Jobin, M. Ienca, and E. Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9 (2019), pp. 389–399.

[62] E. Bietti. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 210–219.

[63] A. Rességuier and R. Rodrigues. "AI ethics should not remain toothless! A call to bring back the teeth of ethics". In: *Big Data & Society* 7.2 (2020), p. 2053951720942541.

[64] L. Munn. "The uselessness of AI ethics". In: *AI and Ethics* (2022), pp. 1–9.

[65] N. A. Ahad and S. S. S. Yahaya. "Sensitivity analysis of Welch'st-test". In: *AIP Conference proceedings*. Vol. 1605. 1. American Institute of Physics. 2014, pp. 888–893.

[66] R. S. Nickerson. "Null hypothesis significance testing: a review of an old and continuing controversy." In: *Psychological methods* 5.2 (2000), p. 241.

[67] Y.-G. Lee and S.-Y. Kim. "Introduction to statistics". In: *Yulgokbooks, Korea* (2008), pp. 342–351.

[68] P. Vik. *Regression, ANOVA, and the general linear model: A statistics primer*. Sage Publications, 2013.

[69] A. Fields, J. Miles, and Z. Fields. *Discovering statistics using R*. London: Sage, 2012.

[70] R. D. Ledesma and P. Valero-Mora. "Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis". In: *Practical assessment, research, and evaluation* 12.1 (2007), p. 2.

[71] B. Thompson. "Exploratory and confirmatory factor analysis: Understanding concepts and applications". In: *Washington, DC* 10694.000 (2004).

[72] R. T. Warne. "A primer on multivariate analysis of variance (MANOVA) for behavioral scientists." In: *Practical Assessment, Research & Evaluation* 19 (2014).

[73] M. W. Watkins. "Exploratory factor analysis: A guide to best practice". In: *Journal of Black Psychology* 44.3 (2018), pp. 219–246.

[74] P. Mayring. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 12th ed. Weinheim/Basel: Beltz Verlag, 2015.

[75] P. Mayring and T. Fenzl. "Qualitative Inhaltsanalyse". In: *Handbuch Methoden der empirischen Sozialforschung*. Ed. by N. Baur and J. Blasius. Springer, 2019, pp. 633–648.

[76] H.-F. Hsieh and S. E. Shannon. "Three approaches to qualitative content analysis". In: *Qualitative health research* 15.9 (2005), pp. 1277–1288.

[77] K. Krippendorff. "Reliability in content analysis: Some common misconceptions and recommendations". In: *Human communication research* 30.3 (2004), pp. 411–433.

[78] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A robustly optimized BERT pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[79] A. Assarroudi, F. Heshmati Nabavi, M. R. Armat, A. Ebadi, and M. Vaismoradi. "Directed qualitative content analysis: the description and elaboration of its underpinning methods and data analysis process". In: *Journal of Research in Nursing* 23.1 (2018), pp. 42–55.

[80] C. Olston, M. Najork, et al. "Web crawling". In: *Foundations and Trends® in Information Retrieval* 4.3 (2010), pp. 175–246.

[81] M. A. Khder. "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application." In: *International Journal of Advances in Soft Computing & Its Applications* 13.3 (2021).

[82] R. Lawson. *Web scraping with Python*. Packt Publishing Ltd, 2015.

[83] R. S. Devi, D. Manjula, and R. Siddharth. "An efficient approach for web indexing of big data through hyperlinks in web crawling". In: *The Scientific World Journal* 2015 (2015).

[84] S. M. Mirtaheri, M. E. Dinçktürk, S. Hooshmand, G. V. Bochmann, G.-V. Jourdan, and I. V. Onut. "A brief history of web crawlers". In: *arXiv preprint arXiv:1405.0749* (2014).

[85] M. A. Kausar, V. Dhaka, and S. K. Singh. "Web crawler: a review". In: *International Journal of Computer Applications* 63.2 (2013).

[86] D. Myers and J. W. McGuffee. "Choosing scrapy". In: *Journal of Computing Sciences in Colleges* 31.1 (2015), pp. 83–89.

[87] D. Kouzis-Loukas. *Learning scrapy*. Packt Publishing Ltd, 2016.

[88] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso, and S. N. Mbaye. "Web scraping: state-of-the-art and areas of application". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 6040–6042.

[89] E. Cambria and B. White. "Jumping NLP curves: A review of natural language processing research". In: *IEEE Computational intelligence magazine* 9.2 (2014), pp. 48–57.

[90] A. Aizawa. "An information-theoretic perspective of tf–idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45–65.

[91] J. Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[92] S. Qaiser and R. Ali. "Text mining: use of TF-IDF to examine the relevance of words to documents". In: *International Journal of Computer Applications* 181.1 (2018), pp. 25–29.

[93] J. W. Mohr and P. Bogdanov. *Introduction—Topic models: What they are and why they matter*. 2013.

[94] J. Mcauliffe and D. Blei. "Supervised topic models". In: *Advances in neural information processing systems* 20 (2007).

[95] Z. Tong and H. Zhang. "A text mining research based on LDA topic modelling". In: *International conference on computer science, engineering and information technology*. 2016, pp. 201–210.

[96] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia Tools and Applications* 78.11 (2019), pp. 15169–15211.

[97] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[98] R. Arun, V. Suresh, C. Veni Madhavan, and N. Murthy. "On finding the natural number of topics with latent dirichlet allocation: Some observations". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2010, pp. 391–402.

[99] Y. Wang, E. Agichtein, and M. Benzi. "TM-LDA: efficient online modeling of latent topic transitions in social media". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, pp. 123–131.

[100] D. M. Blei and J. D. Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

[101] H. Li. "Language models: past, present, and future". In: *Communications of the ACM* 65.7 (2022), pp. 56–63.

[102] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 2020, pp. 38–45.

[103] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[105] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (2019).

[106] C. Atzmüller and P. M. Steiner. "Experimental vignette studies in survey research". In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6.3 (2010), pp. 128–138. DOI: https://doi.org/10.1027/1614-2241/a000014.

[107] M. R. Hyman and S. D. Steiner. "The vignette method in business ethics research: Current uses, limitations, and recommendations". In: *Proceedings of the Annual Meeting of the Southern Marketing Association*. 1996, pp. 261–265.

[108] J. Knobe and S. Nichols. "Experimental philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University, 2017.

[109] P. M. Steiner, C. Atzmüller, and D. Su. "Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap". In: *Journal of Methods and Measurement in the Social Sciences* 7.2 (2016), pp. 52–94.

[110] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. "Likert scale: Explored and explained". In: *British journal of applied science & technology* 7.4 (2015), p. 396.

[111] C. J. Clark, J. B. Luguri, P. H. Ditto, J. Knobe, A. F. Shariff, and R. F. Baumeister. "Free to punish: a motivated account of free will belief." In: *Journal of personality and social psychology* 106.4 (2014), p. 501.

[112] A. Feltz. "The Knobe effect: A brief overview". In: *The Journal of Mind and Behavior* (2007), pp. 265–277.

[113] B. Gert and J. Gert. "The Definition of Morality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Metaphysics Research Lab, Stanford University, 2017.

[114] J. Zigon. *Morality: An Anthropological Perspective*. Berg, 2008.

[115] L. Floridi. *The Ethics of Information*. Oxford University Press, 2013.

[116] State Council. *Notice of the State Council on issuing the outline of the plan for building a Social Credit System (2014-2020); (in Chinese)*. Apr. 2014.

[117] S. Mistreanu. *Life inside China's Social Credit laboratory: The party's massive experiment in ranking and monitoring Chinese citizens has already started*. last accessed on May 26, 2022. 2018. URL: https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/.

[118] M. Turilli and L. Floridi. "The ethics of information transparency". In: *Ethics of Information Technology* (2009), pp. 105–112.

[119] Nature Editorial. "China sets a strong example on how to address scientific fraud". In: *Nature* 558.162 (2018).

[120] Y. Huang. *Why is not there a bottom line for food security issue in China (Zhongguo Shipin Anquan Weihe Meiyou Dixian?) (in Chinese)*. 2012. URL: https://cn.nytimes.com/opinion/20120821/c21huang/.

[121] Ipsos Public Affairs. *What Worries the World?* https://www.ipsos.com/sites/default/files/2017-08/What_worries_the_world-July-2017.pdf. 2017.

[122] A. Demirguc-Kunt, L. Klapper, D. Singer, S. Ansar, and J. Hess. *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. 2018.

[123] P.-H. Wong. "Dao, harmony and personhood: Towards a Confucian ethics of technology". In: *Philosophy & Technology* 25.1 (2012), pp. 67–86.

[124] C. Ess. "Ethical pluralism and global information ethics". In: *Ethics and Information Technology* 8.4 (2006), pp. 215–226.

[125] J. Xu. "Evolving legal frameworks for protecting the right to Internet privacy in China". In: *China and Cybersecurity: Espionage, Strategy, and Politics in the Digital Domain* (2015), pp. 242–259.

[126] Y. Wu, T. Lau, D. Atkin, and C. Lin. "A comparative study of online privacy regulations in the US and China". In: *Telecommunications Policy* 35.7 (2011), pp. 603–616.

[127] Y. Chen and A. Cheung. "The transparent self under big data profiling: Privacy and Chinese legislation on the Social Credit System". In: *The Journal of Comparative Law* 12.2 (2017), pp. 356–378.

[128] Y.-H. Lü. "Privacy and data privacy issues in contemporary China". In: *Ethics and Information Technology* 7.1 (2005), pp. 7–15.

[129] G. Kostka. "China's social credit systems and public opinion: Explaining high levels of approval". In: *New Media & Society* 21.7 (2019), pp. 1565–1593.

[130] Z. Zhao and F. Ding. "Shanghai, Zhejiang, Shenzhen Social Credit System Models, Problems and Revelation (Shanghai, Zhejiang, Shenzhen Shehui Xinyong Tixi Jianshe Moshi, Wenti yu Qishi) (in Chinese)". In: *Wei Shi* 10 (2007). in Chinese, pp. 70–73.

[131] J. Liu. "Building Social Credit System: the Content, the Model, and the Trajectory (Shehui Xinyong Tixi Jianshe: Neihan, Moshi yu Lujing Xuanze; in Chinese)". In: *Journal of the Party School of the Central Committee of the C.P.C.* 15.3 (2011), pp. 50–53.

[132] M. Ohlberg, S. Ahmed, and B. Lang. "Central planning, local experiments: The complex implementation of China's Social Credit System". In: *Mercator Inst. China Studies* (Apr. 2018).

[133] A. Marczewski. "The Ethics of Gamification". In: *XRDS* 24.1 (Sept. 2017), pp. 56–59.

[134] Z. Ramadan. "The gamification of trust: The case of China's "social credit"". In: *Marketing Intelligence & Planning* 36.1 (2018), pp. 93–107.

[135] A. Acquisti and J. Grossklags. "Privacy and rationality in individual decision making". In: *IEEE Security & Privacy* 3.1 (2005), pp. 26–33.

[136] M. Meissner and J. Wübbeke. "IT-backed authoritarianism: Information technology enhances central authority and control capacity under Xi Jinping". In: *China's Core Executive: Leadership Styles, Structures and Processes under Xi Jinping* (2016), pp. 52–57.

[137] N. Jentzsch. *The Economics and Regulation of Financial Privacy: An International Comparison of Credit Reporting Systems*. Springer Science & Business Media, 2006.

[138] C. Hoofnagle. "Big Brother's little helpers: How ChoicePoint and other commercial data brokers collect and package your data for law enforcement". In: *North Carolina Journal of International Law and Commercial Regulation* 29 (2003), pp. 595–637.

[139] M. Röder, A. Both, and A. Hinneburg. "Exploring the space of topic coherence measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM. 2015, pp. 399–408.

[140] J. Chuang, C. Manning, and J. Heer. "Termite: Visualization techniques for assessing textual topic models". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. 2012, pp. 74–77.

[141] C. Sievert and K. Shirley. "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 2014, pp. 63–70.

[142] S. Wachter, B. Mittelstadt, and L. Floridi. "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation". In: *International Data Privacy Law* 7.2 (2017), pp. 76–99.

[143] A. Selbst and J. Powles. "Meaningful information and the right to explanation". In: *International Data Privacy Law* 7.4 (2017), pp. 233–242.

[144] U. Gneezy and A. Rustichini. "A fine is a price". In: *The Journal of Legal Studies* 29.1 (2000), pp. 1–17.

[145] U. Gneezy and A. Rustichini. "Pay enough or don't pay at all". In: *The Quarterly Journal of Economics* 115.3 (2000), pp. 791–810.

[146] R. Bénabou and J. Tirole. "Incentives and prosocial behavior". In: *American Economic Review* 96.5 (2006), pp. 1652–1678.

[147] R. Ryan and E. Deci. "Intrinsic and extrinsic motivations: Classic definitions and new directions". In: *Contemporary Educational Psychology* 25.1 (2000), pp. 54–67.

[148] U. Gneezy, S. Meier, and P. Rey-Biel. "When and why incentives (don't) work to modify behavior". In: *Journal of Economic Perspectives* 25.4 (2011), pp. 191–210.

[149] C. Campbell. "How China is using "Social Credit Scores" to reward and punish its citizens". In: *Time* (2019). Available at: `https://time.com/collection/davos-2019/5502592/china-social-credit-score/`.

[150] A. Ma. "China has started ranking citizens with a creepy 'Social Credit' system – Here's what you can do wrong, and the embarrassing, demeaning ways they can punish you". In: *Business Insider* 29 (2018).

[151] L. Matsakis. "How the West got China's Social Credit System wrong". In: *Wired (29 July 2019)* (2019).

[152] J. Horsley. "China's Orwellian social credit score isn't real". In: *Foreign Policy* 16 (2018).

[153] S. Ahmed. "The messy truth about social credit". In: *Logic* 7 (2019).

[154] M. Chorzempa, P. Triolo, and S. Sacks. *China's Social Credit System: A mark of progress or a threat to privacy?* Policy Briefs, Peterson Institute for International Economics No. PB18-14. 2018.

[155] F. Liang, V. Das, N. Kostyuk, and M. M. Hussain. "Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure". In: *Policy & Internet* 10.4 (2018), pp. 415–453.

[156] X. Dai. *Toward a reputation state: The Social Credit System project of China*. SSRN Working Paper No. 3193577. 2018.

[157] D. Mac Sıthigh and M. Siems. "The Chinese Social Credit System: A model for other countries?" In: *The Modern Law Review* 82.6 (2019), pp. 1034–1071.

[158] M. von Blomberg. "The Social Credit System and China's rule of law". In: *Social Credit Rating*. Ed. by O. Everling. 2018, pp. 111–137.

[159] A. Trauth-Goik. ""Constructing a culture of honesty and integrity": The evolution of China's Han-centric surveillance system". In: *IEEE Technology and Society Magazine* 38.4 (2019), pp. 75–81.

[160] M. Chen and J. Grossklags. "An empirical analysis of the commercial arm of the Chinese Social Credit System". In: *Proceedings on Privacy Enhancing Technologies* 4 (2020), pp. 89–110.

[161] S. Heilmann. *Red Swan: How Unorthodox Policy-Making Facilitated China's Rise*. Chinese University Press, 2018.

[162] S. Engelmann, M. Chen, F. Fischer, C.-Y. Kao, and J. Grossklags. "Clear sanctions, vague rewards: How China's Social Credit System currently defines "good" and "bad" behavior". In: *Proceedings of the Second ACM Conference on Fairness, Accountability, and Transparency*. 2019, pp. 69–78.

[163] S. R. Hoffman. "Programming China: The Communist Party's autonomic approach to managing state security". PhD thesis. University of Nottingham, 2017.

[164] T. Krause and D. Fischer. "An economic approach to China's Social Credit System". In: *Social Credit Rating*. Springer, 2020, pp. 437–453.

[165] R. Creemers. *China's Social Credit System: An evolving practice of control*. SSRN Working Paper Nr. 3175792. 2018.

[166] K. Hao. "The Biden administration's AI plans: what we might expect". In: *MIT Technology Review (22 January 2021)* (2021).

[167] S. Heilmann. "From local experiments to national policy: The origins of China's distinctive policy process". In: *The China Journal* 59 (2008), pp. 1–30.

[168] S. Heilmann. "Maximum tinkering under uncertainty: Unorthodox lessons from China". In: *Modern China* 35.4 (2009), pp. 450–462.

[169] G. Schubert and B. Alpermann. "Studying the Chinese policy process in the era of 'top-level design': The contribution of 'political steering' theory". In: *Journal of Chinese Political Science* 24.2 (2019), pp. 199–224.

[170] A. Mertha. ""Fragmented authoritarianism 2.0": Political pluralization in the Chinese policy process". In: *The China Quarterly* 200 (2009), pp. 995–1012.

[171] J. Jager, D. Putnick, and M. Bornstein. "II. More than just convenient: The scientific merits of homogeneous convenience samples". In: *Monographs of the Society for Research in Child Development* 82.2 (2017), pp. 13–30.

[172] S. Marti and H. Garcia-Molina. "Taxonomy of trust: Categorizing P2P reputation systems". In: *Computer Networks* 50.4 (2006), pp. 472–484.

[173] G. Brennan and P. Pettit. "Hands invisible and intangible". In: *Synthese* 94.2 (1993), pp. 191–225.

[174] Y. Sun, I. Councill, and L. Giles. "The ethicality of web crawlers". In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2010, pp. 668–675.

[175] V. Krotov and L. Silva. "Legality and ethics of web scraping". In: *Proceedings of the Twenty-fourth Americas Conference on Information Systems (AMCIS)*. 2018.

[176] M. Thelwall and D. Stuart. "Web crawling ethics revisited: Cost, privacy, and denial of service". In: *Journal of the American Society for Information Science and Technology* 57.13 (2006), pp. 1771–1779.

[177] Z. Mneimneh, J. Pasek, L. Singh, R. Best, L. Bode, E. Bruch, C. Budak, P. Davis-Kean, K. Donato, N. Ellison, et al. *Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data*. Working Paper. 2021.

[178] Y. T. Yan. *Exporting China's Social Credit System to Central Asia*. Jan. 2020. URL: https://thediplomat.com/2020/01/exporting-chinas-social-credit-system-to-central-asia/.

[179] M. Chen, K. Bogner, J. Becheva, and J. Grossklags. "The transparency of the Chinese Social Credit System from the perspective of German organizations". In: *Proceedings of the 29th European Conference on Information Systems (ECIS)*. Completed research paper. 2021.

[180] D. Lazer, A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, et al. "Computational social science: Obstacles and opportunities". In: *Science* 369.6507 (2020), pp. 1060–1062.

[181] P. Spence. "How to cheat at Xi Jinping Thought". In: *Foreign Policy* (2019). Last accessed on February 4, 2022. URL: https://foreignpolicy.com/2019/03/06/how-to-cheat-at-xi-jinping-thought/.

[182] N. R. Lardy. "China: Toward a consumption-driven growth path". In: *Seeking Changes: The Economic Development in Contemporary China*. Ed. by Y. Zhou. World Scientific, 2016, pp. 85–111.

[183] R. Tyers. "China and global macroeconomic interdependence". In: *The World Economy* 39.11 (2016), pp. 1674–1702.

[184] M. Chen and J. Grossklags. "An analysis of the current state of the Consumer Credit Reporting System in China". In: *Proceedings on Privacy Enhancing Technologies* 2020.4 (2020), pp. 89–110.

[185] M. Chen, K. Bogner, J. Becheva, and J. Grossklags. "The transparency of the Chinese Social Credit System from the perspective of German organizations". In: *Proceedings of the 29th European Conference on Information Systems (ECIS)*. 2021.

[186] S. Engelmann, M. Chen, L. Dang, and J. Grossklags. "Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 78–88.

[187] R. Creemers. *Is big data increasing Beijing's capacity for control?* Last accessed on May 26, 2022. 2016. URL: http://www.chinafile.com/conversation/Is-Big-Data-Increasing-Beijing-Capacity-Control/.

[188] G. Fellner, R. Sausgruber, and C. Traxler. "Testing enforcement strategies in the field: Threat, moral appeal and social information". In: *Journal of the European Economic Association* 11.3 (2013), pp. 634–660.

[189] M. Chudek and J. Henrich. "Culture-gene coevolution, norm-psychology and the emergence of human prosociality". In: *Trends in Cognitive Sciences* 15.5 (2011), pp. 218–226.

[190] B. Bakken. *The exemplary society: Human improvement, social control, and the dangers of modernity in China*. Oxford University Press, 2000.

[191] S. Lindtner, K. Anderson, and P. Dourish. "Cultural appropriation: Information technologies as sites of transnational imagination". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*. 2012, pp. 77–86.

[192] R. Botsman. "Big data meets Big Brother as China moves to rate its citizens". In: *Wired UK* (Oct. 2017).

[193] P. Farrell and P. Tyrrell. *China's 'Social Credit' monitoring: Big Brother's frightening new tool for repression*. last accessed on May 26, 2022. Nov. 2018. URL: https://www.dailysignal.com/2018/11/02/chinas-social-credit-monitoring-big-brothers-frightening-new-tool-for-repression/.

[194] S. Hoffman. "Programming China". In: *Merics China Monitor* 44 (2017), pp. 1–12.

[195] F. Liang, V. Das, N. Kostyuk, and M. M. Hussain. "Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure". In: *Policy & Internet* 10.4 (2018), pp. 415–453.

[196] C. Campbell. "How China is using "Social Credit Scores" to reward and punish its citizens". In: *Time* (Jan. 2019).

[197] S. Hoffman. "Managing the state: Social credit, surveillance, and the Chinese Communist Party's plan for China". In: *Artificial Intelligence, China, Russia, and the Global Order*. Ed. by N. D. Wright. Air University Press, 2018, pp. 48–54.

[198] M. Persson, M. Vlaskamp, and F. Obbema. *China rates its own citizens – Including online behavior*. last accessed on May 26, 2022. Apr. 2015. URL: https://www.volkskrant.nl/buitenland/china-rates-its-own-citizens-including-online-behaviour~a3979668/..

[199] M. Chorzempa, P. Triolo, S. Sacks, et al. *China's Social Credit System: A mark of progress or a threat to privacy?* Peterson Institute for International Economics, Policy Brief 18-14. 2018.

[200] S. Mosher. *China's new 'Social Credit System' is a dystopian nightmare*. last accessed on May 26, 2022. May 2019. URL: https://nypost.com/2019/05/18/chinas-new-social-credit-system-turns-orwells-1984-into-reality/.

[201] W. W. Moss. "Dang'an: Contemporary Chinese archives". In: *The China Quarterly* 145 (1996), pp. 112–129.

[202] M. Chen and J. Grossklags. "Social control in the digital transformation of society: A case study of the Chinese Social Credit System". In: *Social Sciences* 11.6 (2022), Article No. 229.

[203] Z. T. Chen and M. Cheung. "Privacy perception and protection on Chinese social media: A case study of WeChat". In: *Ethics and Information Technology* 20.4 (2018), pp. 279–289.

[204] M. W. Berkowitz and F. Oser. *Moral education: Theory and application*. Lawrence Erlbaum Associates, 1985.

[205] P. Singer and R. Singer. *The moral of the story: An anthology of ethics through literature*. Wiley, 2005.

[206] R. H. Hersh, J. P. Miller, and G. D. Fielding. *Models of Moral Education: An Appraisal*. Longman, 1980.

[207] D. Resnick. "The role of heroes in Jewish education". In: *Religious Education* 97.2 (2002), pp. 108–123.

[208] D. Wong. "Reasons and analogical reasoning in Mengzi". In: *Essays on the Moral Philosophy of Mengzi*. Ed. by X. Liu and P. J. Ivanhoe. Hackett Publishing Indianapolis, 2002, pp. 187–220.

[209] D. Wong. "Chinese ethics". In: *Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. last accessed on May 26, 2022. Stanford University, 1995. URL: https://plato.stanford.edu/archives/sum2021/entries/ethics-chinese/.

[210] G. G. Reed. "Moral/political education in the People's Republic of China: Learning through role models". In: *Journal of Moral Education* 24.2 (1995), pp. 99–111.

[211] B. McLaughlin and J. A. Velez. "Imagined politics: How different media platforms transport citizens into political narratives". In: *Social Science Computer Review* 37.1 (2019), pp. 22–37.

[212] J. Daniel. "Choosing the type of probability sampling". In: *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Sage Publications, Inc, 2012. Chap. 5, pp. 125–175.

[213] V. Krotov and L. Silva. "Legality and ethics of web scraping". In: *Proceedings of the Twenty-fourth Americas Conference on Information Systems*. 2018.

[214] H.-F. Hsieh and S. Shannon. "Three approaches to qualitative content analysis". In: *Qualitative Health Research* 15.9 (2005), pp. 1277–1288.

[215] M. B. Tappan and L. M. Brown. "Stories told and lessons learned: Toward a narrative approach to moral development and moral education". In: *Harvard Educational Review* 59.2 (May 1989), pp. 182–205.

[216] D. B. Bernheim. "A theory of conformity". In: *Journal of Political Economy* 102.5 (1994), pp. 841–877.

[217] R. B. Cialdini and R. R. Reno. "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places". In: *Journal of Personality and Social Psychology* 58.6 (1990), pp. 1015–1026.

[218] R. R. Reno, R. B. Cialdini, and C. A. Kallgren. "The transsituational influence of social norms". In: *Journal of Personality and Social Psychology* 64.1 (1993), pp. 104–112.

[219] R. F. Baumeister. "The self". In: *Handbook of Social Psychology*. Ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey. 4th. Vol. 1. McGraw Hill, 1998, pp. 680–740.

[220] N. Mazar and D. Ariely. "The dishonesty of honest people: A theory of self-concept maintenance". In: *Journal of Marketing Research* 45.6 (2008), pp. 633–644.

[221] A. Polk. "Chinese need to learn to save again". In: *Bloomberg.com* (Feb. 2018).

[222] Sohu Business. *The Young Chinese with Zero Saving and Unaffordable Credit Cards Alert China's Economy*. `http://www.sohu.com/a/249634274_100063243`, last accessed on May 26, 2022 (in Chinese). Aug. 2018.

[223] W. Garcia-Fontes. *Small and medium enterprises financing in China*. Central Bank of Malaysia Working Paper. 2005.

[224] C. Huang and Z. Liu. "Analysis on financing difficulties for SMEs due to asymmetric information". In: *Global Disclosure of Economics and Business* 3.1 (2014), pp. 77–80.

[225] G. Turner. *Ordinary people and the media: The demotic turn*. Sage Publications, 2010.

[226] M. Murray. "Narrative psychology and narrative analysis". In: *Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design*. Ed. by P. M. Camic, J. E. Rhodes, and L. Yardley. American Psychological Association, 2003, pp. 95–112.

[227] E. Jeffreys. "Modern China's idols: Heroes, role models, stars and celebrities". In: *Portal: Journal of Multidisciplinary International Studies* 9.1 (2012), pp. 1–32.

[228] W.-t. Chan. "Exploring the Confucian tradition". In: *Philosophy East and West* 38.3 (1988), pp. 234–250.

[229] A. Olberding. "Dreaming of the Duke of Zhou: Exemplarism and the Analects". In: *Journal of Chinese Philosophy* 35.4 (2008), pp. 625–639.

[230] E. Slingerland. "The situationist critique and early Confucian virtue ethics". In: *Ethics* 121.2 (2011), pp. 390–419.

[231] M. Rokeach. *The Open and Closed Mind: Investigations into the Nature of Belief Systems and Personality Systems*. Basic Books, 1960.

[232] D. Bar-Tal. *Group beliefs: A conception for analyzing group structure, processes, and behavior*. Springer Science & Business Media, 2012.

[233] E. Goffman. *The presentation of self in everyday life*. Doubleday, 1959.

[234] S. Michalopoulos and M. M. Xue. "Folklore". In: *The Quarterly Journal of Economics* 136.4 (2021), pp. 1993–2046.

[235] G. Venkatadri, P. Sapiezynski, E. M. Redmiles, A. Mislove, O. Goga, M. Mazurek, and K. P. Gummadi. "Auditing Offline Data Brokers via Facebook's Advertising Platform". In: *The World Wide Web Conference*. 2019, pp. 1920–1930. DOI: https://doi.org/10.1145/3308558.3313666.

[236] A. Andreou, G. Venkatadri, O. Goga, K. Gummadi, P. Loiseau, and A. Mislove. "Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations". In: *Network and Distributed System Security Symposium (NDSS)*. 2018, pp. 1–15. URL: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_10-1_Andreou_paper.pdf.

[237] S. Engelmann and J. Grossklags. "Setting the Stage: Towards Principles for Reasonable Image Inferences". In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP)*. 2019, pp. 301–307. DOI: https://doi.org/10.1145/3314183.3323846.

[238] P. Sapiezynski, A. Ghosh, L. Kaplan, A. Mislove, and A. Rieke. "Algorithms that" Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences". In: *arXiv preprint arXiv:1912.07579* (2019).

[239] Q. Zhao, M. Willemsen, G. Adomavicius, M. Harper, and J. Konstan. "Interpreting user inaction in recommender systems". In: *12th ACM Conference on Recommender Systems (RecSys)*. 2018, pp. 40–48. DOI: https://doi.org/10.1145/3240323.3240366.

[240] D. Sovilj, S. Sanner, H. Soh, and H. Li. "Collaborative filtering with behavioral models". In: *Conference on User Modeling, Adaptation and Personalization (UMAP)*. 2018, pp. 91–99. DOI: https://doi.org/10.1145/3209219.3209235.

[241] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. "Our Twitter profiles, our selves: Predicting personality with Twitter". In: *IEEE Third International Conference on Social Computing*. 2011, pp. 180–185. DOI: 10.1109/PASSAT/SocialCom.2011.26.

[242] A. Pak and P. Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In: *Seventh International Conference on Language Resources and Evaluation (LREC)*. 2010, pp. 1320–1326. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf.

[243] B. Ferwerda, M. Schedl, and M. Tkalcic. "Predicting personality traits with Instagram pictures". In: *Workshop on Emotions and Personality in Personalized Systems*. 2015, pp. 7–10. DOI: https://doi.org/10.1145/2809643.2809644.

[244] W. Ma, M. Zhang, C. Wang, C. Luo, Y. Liu, and S. Ma. "Your Tweets reveal what you like: Introducing cross-media content information into multi-domain recommendation". In: *International Joint Conferences on Artificial Intelligence (IJCAI)*. 2018, pp. 3484–3490. URL: https://www.ijcai.org/proceedings/2018/0484.pdf.

[245] B. Chang, Y. Park, D. Park, S. Kim, and J. Kang. "Content-aware hierarchical point-of-interest embedding model for successive POI recommendation". In: *International Joint Conferences on Artificial Intelligence (IJCAI)*. 2018, pp. 3301–3307. URL: https://www.ijcai.org/proceedings/2018/0458.pdf.

[246] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer. "Personalized emotion recognition by personality-aware high-order learning of physiological signals". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 15.1s (2019), Article No. 14. DOI: https://doi.org/10.1145/3233184.

[247] G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens. "User profiling through deep multimodal fusion". In: *Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. 2018, pp. 171–179. DOI: https://doi.org/10.1145/3159652.3159691.

[248] M. Hildebrandt. "Privacy as protection of the incomputable self: From agnostic to agonistic machine learning". In: *Theoretical Inquiries in Law* 20.1 (2019), pp. 83–121. DOI: https://doi.org/10.1515/til-2019-0004.

[249] J. Van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. "Visual word ambiguity". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.7 (2010), pp. 1271–1283. DOI: https://doi.org/10.1109/TPAMI.2009.132.

[250] A. Rouvroy and Y. Poullet. "The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy". In: *Reinventing Data Protection?* Ed. by S. Gutwirth, Y. Poullet, P. De Hert, C. de Terwangne, and S. Nouwt. Springer, 2009, pp. 45–76. DOI: https://doi.org/10.1007/978-1-4020-9498-9_2.

[251] J. Kupfer. "Privacy, autonomy, and self-concept". In: *American Philosophical Quarterly* 24.1 (1987), pp. 81–89. URL: https://www.jstor.org/stable/20014176.

[252] D. Susser, B. Roessler, and H. Nissenbaum. "Technology, autonomy, and manipulation". In: *Internet Policy Review* 8.2 (2019), pp. 1–22. DOI: 10.14763/2019.2.1410.

[253] E. J. Bloustein. "Privacy as an aspect of human dignity: An answer to Dean Prosser". In: *New York University Law Review* 39.6 (1964), pp. 962–1007. URL: https://heinonline.org/HOL/LandingPage?handle=hein.journals/nylr39&div=71.

[254] C. Prunkl. "Human autonomy in the age of artificial intelligence". In: *Nature Machine Intelligence* 4.2 (2022), pp. 99–101. DOI: https://doi.org/10.1038/s42256-022-00449-9.

[255] L. Stark. "Facial recognition is the plutonium of AI". In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (2019), pp. 50–55. DOI: https://doi.org/10.1145/3313129.

[256] L. Stark and J. Hutson. "Physiognomic artificial intelligence". In: *Fordham Intellectual Property, Media & Entertainment Law Journal* 32.4 (2022), pp. 922–978. URL: https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2/.

[257] S. Engelmann, C. Ullstein, O. Papakyriakopoulos, and J. Grossklags. "What People Think AI Should Infer From Faces". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022. DOI: https://doi.org/10.1145/3531146.3533080.

[258] E. Aımeur. "Personalisation and privacy issues in the age of exposure". In: *Conference on User Modeling, Adaptation and Personalization (UMAP)*. 2018, pp. 375–376. DOI: https://doi.org/10.1145/3209219.3209271.

[259] A. D. Selbst, d. boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. "Fairness and abstraction in sociotechnical systems". In: *Second Annual Conference on Fairness, Accountability, and Transparency*. 2019, pp. 59–68. DOI: https://doi.org/10.1145/3287560.3287598.

[260] J. Gama, R. Sebastião, and P. P. Rodrigues. "On evaluating stream learning algorithms". In: *Machine Learning* 90.3 (2013), pp. 317–346. DOI: https://doi.org/10.1007/s10994-012-5320-9.

[261] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. "Learning under Concept Drift: A Review". In: *IEEE Transactions on Knowledge and Data Engineering* 31.12 (2019), pp. 2346–2363. DOI: 10.1109/TKDE.2018.2876857.

[262] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. "YALE: Rapid prototyping for complex data mining tasks". In: *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, pp. 935–940. DOI: https://doi.org/10.1145/1150402.1150531.

[263] I. Žliobaitė. "Learning under concept drift: An overview". In: *arXiv preprint arXiv:1010.4784* (2010). DOI: https://doi.org/10.48550/arXiv.1010.4784.

[264] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. "An algorithmic framework for performing collaborative filtering". In: *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1999, pp. 230–237. DOI: https://doi.org/10.1145/312624.312682.

[265] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. "User profiles for personalized information access". In: *The Adaptive Web*. Ed. by P. Brusilovsky, A. Kobsa, and W. Nejdl. Springer, 2007, pp. 54–89. DOI: 10.1007/978-3-540-72079-9_2.

[266] F. Zarrinkalam, H. Fani, and E. Bagheri. "Extracting, mining and predicting users' interests from social networks". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 1407–1408. DOI: https://doi.org/10.1145/3331184.3331383.

[267] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke. "A survey of user profiling: State-of-the-art, challenges, and solutions". In: *IEEE Access* 7 (2019), pp. 144907–144924. DOI: 10.1109/ACCESS.2019.2944243.

[268] J. Chen, Y. Liu, and M. Zou. "Home location profiling for users in social media". In: *Information & Management* 53.1 (2016), pp. 135–143. DOI: https://doi.org/10.1016/j.im.2015.09.008.

[269] A. Abdel-Hafez and Y. Xu. "A survey of user modelling in social media websites". In: *Computer and Information Science* 6.4 (2013), pp. 59–71. DOI: 10.5539/cis.v6n4p59.

[270] J. Van Dijck. "'You have one identity': Performing the self on Facebook and LinkedIn". In: *Media, Culture & Society* 35.2 (2013), pp. 199–215. DOI: https://doi.org/10.1177/0163443712468605.

[271] L. Stark. "Algorithmic psychometrics and the scalable subject". In: *Social Studies of Science* 48.2 (2018), pp. 204–231. DOI: https://doi.org/10.1177/0306312718772094.

[272] J. Blass. "Algorithmic advertising discrimination". In: *Northwestern University Law Review* 114.2 (2019), pp. 415–467. URL: https://scholarlycommons.law.northwestern.edu/nulr/vol114/iss2/3/.

[273] K. A. Appiah. *The ethics of identity*. Princeton University Press, 2010.

[274] S. Gallagher. "Philosophical conceptions of the self: Implications for cognitive science". In: *Trends in Cognitive Sciences* 4.1 (2000), pp. 14–21. DOI: https://doi.org/10.1016/S1364-6613(99)01417-5.

[275] S. Buss and L. Overton, eds. *Contours of agency: Essays on themes from Harry Frankfurt*. MIT Press, 2002.

[276] D. Zahavi. "Self and other: The limits of narrative understanding". In: *Royal Institute of Philosophy Supplements* 60 (2007), pp. 179–202. DOI: https://doi.org/10.1017/S1358246107000094.

[277] O. John, L. Naumann, and C. Soto. "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues". In: *Handbook of Personality: Theory and Research*. Ed. by O. P. John, R. W. Robins, and L. A. Pervin. 3rd. The Guilford Press, 2008, pp. 114–158.

[278] M. Schechtman. "The narrative self". In: *The Oxford Handbook of the Self*. Ed. by S. Gallagher. Oxford University Press, 2011.

[279] M. Schechtman. *Staying alive: Personal identity, practical concerns, and the unity of a life*. Oxford University Press, 2014.

[280] M. Schechtman. *The constitution of selves*. Cornell University Press, 1996.

[281] H. G. Frankfurt. "Freedom of the will and the concept of a person". In: *The Journal of Philosophy* 68.1 (1971), pp. 5–20. DOI: https://doi.org/10.2307/2024717.

[282] C. Taylor. "Responsibility for self". In: *The Identities of Persons*. Ed. by A. Rorty. University of California Press, 1976.

[283] C. Taylor. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, 1989.

[284] J. D. Velleman. *Self to self: Selected essays*. Cambridge University Press, 2006.

[285] S. Gallagher, ed. *The Oxford handbook of the self*. Oxford University Press, 2011.

[286] J. Locke. *An essay concerning human understanding*. Printed for Thomas Basset, 1690.

[287] S. Gallagher and D. Zahavi. *The phenomenological mind*. Routledge, 2020. DOI: https://doi.org/10.4324/9780429319792.

[288] H. Yin, B. Cui, L. Chen, Z. Hu, and X. Zhou. "Dynamic user modeling in social media systems". In: *ACM Transactions on Information Systems* 33.3 (2015), pp. 1–44. DOI: https://doi.org/10.1145/2699670.

[289] D. Rafailidis, P. Kefalas, and Y. Manolopoulos. "Preference dynamics with multimodal user-item interactions in social media recommendation". In: *Expert Systems with Applications* 74 (2017), pp. 11–18. DOI: https://doi.org/10.1016/j.eswa.2017.01.005.

[290] M. Chessa, J. Grossklags, and P. Loiseau. "A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications". In: *IEEE 28th Computer Security Foundations Symposium (CSF)*. 2015, pp. 90–104. DOI: 10.1109/CSF.2015.14.

[291] J. Brunk, J. Mattern, and D. M. Riehle. "Effect of transparency and trust on acceptance of automatic online comment moderation systems". In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. 2019, pp. 429–435. DOI: 10.1109/CBI.2019.00056.

[292] J. K. Choi and Y. G. Ji. "Investigating the importance of trust on adopting an autonomous vehicle". In: *International Journal of Human-Computer Interaction* 31.10 (2015), pp. 692–702. DOI: https://doi.org/10.1080/10447318.2015.1070549.

[293] N. El Bekri, J. Kling, and M. F. Huber. "A study on trust in black box models and post-hoc explanations". In: *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO)*. Ed. by F. Martínez Alvarez. Vol. 950. Advances in Intelligent Systems and Computing. Springer, 2020, pp. 35–46. DOI: 10.1007/978-3-030-20055-8_4.

[294] K. F. Oduor and E. N. Wiebe. "The effects of automated decision algorithm modality and transparency on reported trust and task performance". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 52. 4. SAGE Publications Sage CA: Los Angeles, CA. 2008, pp. 302–306.

[295] K. A. Appiah. *Experiments in ethics*. Harvard University Press, 2008.

[296] S. Nichols and J. Knobe. "Moral responsibility and determinism: The cognitive science of folk intuitions". In: *Noûs* 41.4 (2007), pp. 663–685. URL: https://www.jstor.org/stable/4494554.

[297] J. Knobe. "Intentional action and side effects in ordinary language". In: *Analysis* 63.3 (2003), pp. 190–194.

[298] D. E. Melnikoff and N. Strohminger. "The automatic influence of advocacy on lawyers and novices". In: *Nature Human Behaviour* 4.12 (2020), pp. 1258–1264. DOI: https://doi.org/10.1038/s41562-020-00943-3.

[299] S. R. Kraaijeveld. "Experimental philosophy of technology". In: *Philosophy & Technology* 34.4 (2021), pp. 993–1012. DOI: https://doi.org/10.1007/s13347-021-00447-6.

[300] L. Floridi. “Web 2.0 vs. the semantic web: A philosophical assessment”. In: *Episteme* 6.1 (2009), pp. 25–37.

[301] S. Wachter and B. Mittelstadt. “A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI”. In: *Columbia Business Law Review* (2019).

[302] S. Zhao, S. Grasmuck, and J. Martin. “Identity construction on Facebook: Digital empowerment in anchored relationships”. In: *Computers in Human Behavior* 24.5 (2008), pp. 1816–1836.

[303] N. Ellison, C. Steinfield, and C. Lampe. “Connection strategies: Social capital implications of Facebook-enabled communication practices”. In: *New Media & Society* 13.6 (2011), pp. 873–892.

[304] J. Kang and L. Wei. “Let me be at my funniest: Instagram users’ motivations for using Finsta (aka, fake Instagram)”. In: *The Social Science Journal* (2019).

[305] A. MacIntyre. *After Virtue: A Study in Moral Theology*. University of Notre Dame Press, 1981.

[306] M. Leary and J. P. Tangney. *Handbook of Self and Identity*. Guilford Press, 2011.

[307] W. Swann, A. Stein-Seroussi, and B. Giesler. “Why people self-verify”. In: *Journal of Personality and Social Psychology* 62.3 (1992), pp. 392–401.

[308] W. Swann, P. Rentfrow, and J. Guinn. “Self-verification: The search for coherence”. In: *Handbook of Self and Identity*. Ed. by M. Leary and J. P. Tangney. 2003, pp. 367–383.

[309] W. Swann. “Identity negotiation: Where two roads meet”. In: *Journal of Personality and Social Psychology* 53.6 (1987), p. 1038.

[310] V. Lavrenko, R. Manmatha, and J. Jeon. “A model for learning the semantics of pictures”. In: *Advances in Neural Information Processing Systems*. 2004, pp. 553–560.

[311] S. Göring, K. Brand, and A. Raake. “Extended Features using Machine Learning Techniques for Photo Liking Prediction”. In: *International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2018, pp. 1–6.

[312] E. Massip, S. C. Hidayati, W.-H. Cheng, and K.-L. Hua. “Exploiting Category-Specific Information for Image Popularity Prediction in Social Media”. In: *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 2018, pp. 45–46.

[313] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei. “Sequential prediction of social media popularity with deep temporal context networks”. In: *International Joint Conferences on Artificial Intelligence (IJCAI)*. 2017, pp. 3062–3068.

[314] Z. Zhang, T. Chen, Z. Zhou, J. Li, and J. Luo. “How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention”. In: *IEEE International Conference on Big Data (Big Data)*. 2018, pp. 2383–2392.

[315] G. Seidman. “Expressing the “true self” on Facebook”. In: *Computers in Human Behavior* 31 (2014), pp. 367–372.

[316] L. P. Tosun. “Motives for Facebook use and expressing “true self” on the Internet”. In: *Computers in Human Behavior* 28.4 (2012), pp. 1510–1517.

[317] C. Jernigan and B. Mistree. “Gaydar: Facebook friendships expose sexual orientation”. In: *First Monday* 14.10 (2009).

[318] B. Ferwerda, M. Schedl, and M. Tkalcic. “Predicting personality traits with Instagram pictures”. In: *Workshop on Emotions and Personality in Personalized Systems*. ACM. 2015, pp. 7–10.

[319] B. Ferwerda, M. Schedl, and M. Tkalcic. “Using Instagram picture features to predict users’ personality”. In: *International Conference on Multimedia Modeling*. 2016.

[320] B. Ferwerda and M. Tkalcic. “Predicting Users’ Personality from Instagram Pictures: Using Visual and/or Content Features?” In: *Conference on User Modeling, Adaptation and Personalization*. ACM. 2018, pp. 157–161.

[321] K. Han, Y. Jo, Y. Jeon, B. Kim, J. Song, and S.-W. Kim. “Photos Don’t Have Me, But How Do You Know Me? Analyzing and Predicting Users on Instagram”. In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. UMAP ’18. ACM, 2018, pp. 251–256.

[322] K. Han, S. Lee, J. Y. Jang, Y. Jung, and D. Lee. “Teens are from mars, adults are from venus: Analyzing and predicting age groups with behavioral characteristics in Instagram”. In: *Conference on Web Science*. ACM. 2016, pp. 35–44.

[323] J. Song, K. Han, D. Lee, and S.-W. Kim. ““Is a picture really worth a thousand words?”: A case study on classifying user attributes on Instagram”. In: *PloS One* 13.10 (2018), e0204938.

[324] S. V. Jin and A. Muqaddam. ““Narcissism 2.0! Would narcissists follow fellow narcissists on Instagram?” The mediating effects of narcissists personality similarity and envy, and the moderating effects of popularity”. In: *Computers in Human Behavior* 81 (2018), pp. 31–41.

[325] R. Frost and D. Rickwood. “A systematic review of the mental health outcomes associated with Facebook use”. In: *Computers in Human Behavior* 76 (2017), pp. 576–600.

[326] Z. Brown and M. Tiggemann. “Attractive celebrity and peer images on Instagram: Effect on women’s mood and body image”. In: *Body Image* 19 (2016), pp. 37–43.

[327] J. Hendrickse, L. Arpan, R. Clayton, and J. Ridgway. “Instagram and college women’s body image: Investigating the roles of appearance-related comparisons and intrasexual competition”. In: *Computers in Human Behavior* 74 (2017), pp. 92–100.

[328] A. Reece and C. Danforth. “Instagram photos reveal predictive markers of depression”. In: *EPJ Data Science* 6.1 (2017), p. 15.

[329] K. Stanovich. *Decision Making and Rationality in the Modern World (Fundamentals in Cognition)*. Oxford University Press, 2009.

[330] D. Ariely, G. Loewenstein, and D. Prelec. ""Coherent arbitrariness": Stable demand curves without stable preferences". In: *The Quarterly Journal of Economics* 118.1 (2003), pp. 73–106.

[331] A. Tversky and D. Kahneman. "The framing of decisions and the psychology of choice". In: *Science* 211.4481 (1981), pp. 453–458.

[332] J. Pollock and J. Cruz. *Contemporary Theories of Knowledge*. Rowman & Littlefield, 1999.

[333] J. Cohen. "Can human irrationality be experimentally demonstrated?" In: *Behavioral and Brain Sciences* 4.3 (1981), pp. 317–331.

[334] M. Argyle and R. McHenry. "Do spectacles really affect judgements of intelligence?" In: *British Journal of Social and Clinical Psychology* 10.1 (1971), pp. 27–29.

[335] F. Moore, D. Filippou, and D. I. Perrett. "Intelligence and attractiveness in the face: Beyond the attractiveness halo effect". In: *Journal of Evolutionary Psychology* 9.3 (2011), pp. 205–217.

[336] O. King. "Machine Learning and Irresponsible Inference: Morally Assessing the Training Data for Image Recognition Systems". In: *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*. Ed. by D. Berkich and M. V. d'Alfonso. Springer, 2019, pp. 265–282.

[337] H. Over and R. Cook. "Where do spontaneous first impressions of faces come from?" In: *Cognition* 170 (2018), pp. 190–200.

[338] C. Y. Olivola, D. L. Eubanks, and J. B. Lovelace. "The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance". In: *The Leadership Quarterly* 25.5 (2014), pp. 817–834.

[339] N. O. Rule and N. Ambady. "The face of success: Inferences from chief executive officers' appearance predict company profits". In: *Psychological Science* 19.2 (2008), pp. 109–111.

[340] C. C. Ballew and A. Todorov. "Predicting political elections from rapid and unreflective face judgments". In: *Proceedings of the National Academy of Sciences* 104.46 (2007), pp. 17948–17953.

[341] C. Y. Olivola, A. B. Sussman, K. Tsetsos, O. E. Kang, and A. Todorov. "Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters". In: *Social Psychological and Personality Science* 3.5 (2012), pp. 605–613.

[342] J. P. Wilson and N. O. Rule. "Facial trustworthiness predicts extreme criminal-sentencing outcomes". In: *Psychological Science* 26.8 (2015), pp. 1325–1331.

[343] R. Dumas and B. Testé. "The influence of criminal facial stereotypes on juridic judgments." In: *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie* 65.4 (2006), p. 237.

[344] L. A. Zebrowitz and S. M. McDonald. "The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts". In: *Law and Human Behavior* 15.6 (1991), pp. 603–623.

[345] K. Crawford and T. Paglen. "Excavating AI: The politics of images in machine learning training sets". In: *AI & SOCIETY* (2021), pp. 1–12.

[346] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. "Saving face: Investigating the ethical concerns of facial recognition auditing". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 145–151.

[347] L. Stark and J. Hoey. "The ethics of emotion in artificial intelligence systems". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 782–793.

[348] J. Goldenfein. "The profiling potential of computer vision and the challenge of computational empiricism". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 110–119.

[349] J.-F. Bonnefon, A. Hopfensitz, W. De Neys, et al. "Face-ism and kernels of truth in facial inferences". In: *Trends in Cognitive Sciences* 19.8 (2015), pp. 421–422.

[350] A. Todorov, S. G. Baron, and N. N. Oosterhof. "Evaluating face trustworthiness: A model based approach". In: *Social Cognitive and Affective Neuroscience* 3.2 (2008), pp. 119–127.

[351] C. Efferson and S. Vogt. "Viewing men's faces does not lead to accurate predictions of trustworthiness". In: *Scientific Reports* 3.1 (2013), pp. 1–7.

[352] R. S. Kramer and R. Ward. "Internal facial features are signals of personality and health". In: *The Quarterly Journal of Experimental Psychology* 63.11 (2010), pp. 2273–2287.

[353] K. Kleisner, V. Chvátalová, and J. Flegr. "Perceived intelligence is associated with measured intelligence in men but not women". In: *PloS ONE* 9.3 (2014), e81237.

[354] A. Kachur, E. Osin, D. Davydov, K. Shutilov, and A. Novokshonov. "Assessing the Big Five personality traits using real-life static facial images". In: *Scientific Reports* 10.1 (2020), pp. 1–11.

[355] L. Naumann, S. Vazire, P. Rentfrow, and S. Gosling. "Personality judgments based on physical appearance". In: *Personality and Social Psychology Bulletin* 35.12 (2009), pp. 1661–1671.

[356] A. C. Little and D. I. Perrett. "Using composite images to assess accuracy in personality attribution to faces". In: *British Journal of Psychology* 98.1 (2007), pp. 111–126.

[357] J. F. Cohn and F. De la Torre. "Automated face analysis for affective computing." In: (2015).

[358] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita. "Perceiving the person and their interactions with the others for social robotics–a review". In: *Pattern Recognition Letters* 118 (2019), pp. 3–13.

[359] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario. "Emotion modelling for social robotics applications: a review". In: *Journal of Bionic Engineering* 15.2 (2018), pp. 185–203.

[360] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5562–5570.

[361] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. "Video and image based emotion recognition challenges in the wild: Emotiw 2015". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015, pp. 423–426.

[362] Y. Wang and M. Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images". In: *Journal of Personality and Social Psychology* 114.2 (2018), pp. 246–257.

[363] J. Leuner. "A Replication Study: Machine Learning Models Are Capable of Predicting Sexual Orientation From Facial Images". In: *arXiv preprint arXiv:1902.10739* (2019).

[364] M. Kosinski. "Facial recognition technology can expose political orientation from naturalistic facial images". In: *Scientific Reports* 11.1 (2021), pp. 1–7.

[365] N. Xi, D. Ma, M. Liou, Z. C. Steinert-Threlkeld, J. Anastasopoulos, and J. Joo. "Understanding the political ideology of legislators from social media images". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 726–737.

[366] C. Segalin, A. Perina, M. Cristani, and A. Vinciarelli. "The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits". In: *IEEE Transactions on Affective Computing* (2016).

[367] C. Segalin, D. S. Cheng, and M. Cristani. "Social profiling through image understanding: Personality inference using convolutional neural networks". In: *Computer Vision and Image Understanding* 156 (2017), pp. 34–50.

[368] Y. Yan, J. Nie, L. Huang, Z. Li, Q. Cao, and Z. Wei. "Is your first impression reliable? Trustworthy analysis using facial traits in portraits". In: *International Conference on Multimedia Modeling*. Springer. 2015, pp. 148–158.

[369] L. Qiu, J. Lu, S. Yang, W. Qu, and T. Zhu. "What does your selfie say about you?" In: *Computers in Human Behavior* (2015).

[370] F. Celli, E. Bruni, and B. Lepri. "Automatic personality and interaction style recognition from Facebook profile pictures". In: *Conf. ACM Multimedia*. 2014.

[371] D. Azucar, D. Marengo, and M. Settanni. "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis". In: *Personality and Individual Differences* (2018).

[372] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli. "Face-based automatic personality perception". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014, pp. 1153–1156.

[373] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. "Facetube: Predicting personality from facial expressions of emotion in online conversational video". In: *Conf. Multimodal Interaction*. 2012, pp. 53–56.

[374] K. Crawford. "Time to regulate AI that interprets human emotions". In: *Nature* 592.7853 (2021), pp. 167–167.

[375] S. Nowak and S. Rüger. "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation". In: *Proceedings of the International Conference on Multimedia Information Retrieval*. 2010, pp. 557–566.

[376] H. Su, J. Deng, and L. Fei-Fei. "Crowdsourcing annotations for visual object detection". In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[377] M. Miceli and J. Posada. "Wisdom for the Crowd: Discoursive Power in Annotation Instructions for Computer Vision". In: *arXiv preprint arXiv:2105.10990* (2021).

[378] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. Sánchez, et al. *AI Now 2019 Report. AI Now Institute*. 2019.

[379] M. Miceli, M. Schuessler, and T. Yang. "Between subjectivity and imposition: Power dynamics in data annotation for computer vision". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–25.

[380] M. Miceli, T. Yang, L. Naudts, M. Schuessler, D. Serbanescu, and A. Hanna. "Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 161–172.

[381] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 325–336.

[382] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker. "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–35.

[383] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 469–481.

[384] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[385] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 220–229.

[386] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. "The dataset nutrition label: A framework to drive higher data quality standards". In: *arXiv preprint arXiv:1805.03677* (2018).

[387] Z. Khan and Y. Fu. "One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 587–597.

[388] M. Hanley, S. Barocas, K. Levy, S. Azenkot, and H. Nissenbaum. "Computer Vision and Conflicting Values: Describing People with Automated Alt Text". In: *arXiv preprint arXiv:2105.12754* (2021).

[389] A. Todorov, C. Y. Olivola, R. Dotsch, and P. Mende-Siedlecki. "Social attributions from faces: Determinants, consequences, accuracy, and functional significance". In: *Annual Review of Psychology* 66 (2015), pp. 519–545.

[390] H. Over, A. Eggleston, and R. Cook. "Ritual and the origins of first impressions". In: *Philosophical Transactions of the Royal Society B* 375.1805 (2020), p. 20190435.

[391] L. A. Zebrowitz and J. M. Montepare. "Social psychological face perception: Why appearance matters". In: *Social and Personality Psychology Compass* 2.3 (2008), pp. 1497–1517.

[392] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. "Facebook profiles reflect actual personality, not self-idealization". In: *Psych. Science* (2010).

[393] T. Teixeira, M. Wedel, and R. Pieters. "Emotion-induced engagement in internet video advertisements". In: *Journal of Marketing Research* 49.2 (2012), pp. 144–159.

[394] S. C. Guntuku, W. Lin, J. Carpenter, W. K. Ng, L. H. Ungar, and D. Preoţiuc-Pietro. "Studying personality through the content of posted and liked images on Twitter". In: *Web Science Conf.* 2017.

[395] M. K. Scheuerman, M. Pape, and A. Hanna. "Auto-essentialization: Gender in automated facial analysis as extended colonial project". In: *Big Data & Society* 8.2 (2021), p. 20539517211053712.

[396] O. Keyes. "The misgendering machines: Trans/HCI implications of automatic gender recognition". In: *Proceedings of the ACM on human-computer interaction* 2.CSCW (2018), pp. 1–22.

[397] R. J. Vernon, C. A. Sutherland, A. W. Young, and T. Hartley. "Modeling first impressions from highly variable facial images". In: *Proceedings of the National Academy of Sciences* 111.32 (2014), E3353–E3361.

[398] O. C. King. "Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems". In: *On the cognitive, ethical, and scientific dimensions of artificial intelligence*. Springer, 2019, pp. 265–282.

[399] O. C. King. "Presumptuous aim attribution, conformity, and the ethics of artificial social cognition". In: *Ethics and Information Technology* 22.1 (2020), pp. 25–37. DOI: https://doi.org/10.1007/s10676-019-09512-3.

[400] A. Sorokin and D. Forsyth. "Utility data annotation with Amazon Mechanical Turk". In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2008, pp. 1–8.

[401] G. Wohlgenannt. "A comparison of domain experts and crowdsourcing regarding concept relevance evaluation in ontology learning". In: *International Workshop on Multidisciplinary Trends in Artificial Intelligence*. Springer. 2016, pp. 243–254. DOI: 10.1007/978-3-319-49397-821.

[402] V. Williamson. "On the ethics of crowdsourced research". In: *PS: Political Science & Politics* 49.1 (2016), pp. 77–81.

[403] B. F. Malle and J. Knobe. "The folk concept of intentionality". In: *Journal of Experimental Social Psychology* 33.2 (1997), pp. 101–121.

[404] K. Kieslich, M. Lünich, and F. Marcinkowski. "The Threats of Artificial Intelligence Scale (TAI)". In: *International Journal of Social Robotics* (2021), pp. 1–15.

[405] A. Smith, L. Rainie, K. Olmstead, J. Jiang, A. Perrin, P. Hitlin, and M. Hefferon. "Public attitudes toward computer algorithms". In: *Pew Research Center* 16 (2018). URL: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/11/PI_2018.11.19_algorithms_FINAL.pdf.

[406] D. Chong and J. N. Druckman. "Framing theory". In: *Annual Review of Political Science* 10 (2007), pp. 103–126.

[407] I. Grossmann, R. P. Eibach, J. Koyama, and Q. B. Sahi. "Folk standards of sound judgment: Rationality Versus Reasonableness". In: *Science Advances* 6.2 (2020), eaaz0289.

[408] A. Bear and J. Knobe. "Normality: Part descriptive, part prescriptive". In: *Cognition* 167 (2017), pp. 25–37.

[409] J. Demaree-Cotton. "Do framing effects make moral intuitions unreliable?" In: *Philosophical Psychology* 29.1 (2016), pp. 1–22.

[410] M. Kutlu, T. McDonnell, Y. Barkallah, T. Elsayed, and M. Lease. "Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement?" In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018, pp. 805–814.

[411] M. Kutlu, T. McDonnell, T. Elsayed, and M. Lease. "Annotator rationales for labeling tasks in crowdsourcing". In: *Journal of Artificial Intelligence Research* 69 (2020), pp. 143–189.

[412] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Vol. 10. Ishk, 2003.

[413] P. Tambe, P. Cappelli, and V. Yakubovich. "Artificial intelligence in human resources management: Challenges and a path forward". In: *California Management Review* 61.4 (2019), pp. 15–42.

[414] J. M. McCarthy, T. N. Bauer, D. M. Truxillo, N. R. Anderson, A. C. Costa, and S. M. Ahmed. "Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?"" In: *Journal of Management* 43.6 (2017), pp. 1693–1725.

[415] L. Li, T. Lassiter, J. Oh, and M. K. Lee. "Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021.

[416] S. K. Jha, S. Jha, and M. K. Gupta. "Leveraging Artificial Intelligence for Effective Recruitment and Selection Processes". In: *International Conference on Communication, Computing and Electronics Systems*. Springer, Singapore. 2020, pp. 287–293.

[417] J. M. Basch and K. G. Melchers. "Fair and flexible?! Explanations can improve applicant reactions toward asynchronous video interviews". In: *Personnel Assessment and Decisions* 5.3 (2019), p. 2.

[418] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier. "Multimodal first impression analysis with deep residual networks". In: *IEEE Transactions on Affective Computing* 9.3 (2017), pp. 316–329.

[419] H. F. Kaiser. "A second generation little jiffy". In: *Psychometrika* 35.4 (1970), pp. 401–415.

[420] G. D. Hutcheson and N. Sofroniou. *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage, 1999.

[421] M. C. Howard. "A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?" In: *International Journal of Human-Computer Interaction* 32.1 (2016), pp. 51–62.

[422] A. B. Costello and J. Osborne. "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis". In: *Practical Assessment, Research, and Evaluation* 10.1 (2005), p. 7. DOI: 10.7275/jyj1-4868.

[423] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. 2013.

[424] C. Zygmont and M. R. Smith. "Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions". In: *The Quantitative Methods for Psychology* 10.1 (2014), pp. 40–55. DOI: 10.20982/tqmp.10.1.p040.

[425] R. K. Henson and J. K. Roberts. "Use of exploratory factor analysis in published research: Common errors and some comment on improved practice". In: *Educational and Psychological measurement* 66.3 (2006), pp. 393–416.

[426] Faception. *Our technology*. Accessed: 2022-02-24. 2021. URL: https://www.faception.com/our-technology (visited on 02/24/2022).

[427] Clearview.ai. *Overview*. Accessed: 2022-02-24. 2022. URL: https://www.clearview.ai/overview.

[428] Face++. *Face Attributes*. Accessed: 2022-02-25. n.d. URL: https://www.faceplusplus.com/attributes/.

[429] Y. Yan, J. Nie, L. Huang, Z. Li, Q. Cao, and Z. Wei. "Is Your First Impression Reliable? Trustworthy Analysis Using Facial Traits in Portraits". In: *MultiMedia Modeling*. Ed. by X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. A. Hasan. Cham: Springer International Publishing, 2015, pp. 148–158. ISBN: 978-3-319-14442-9. DOI: 10.1007/978-3-319-14442-9_13.

[430] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli. "Face-Based Automatic Personality Perception". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 1153–1156. ISBN: 9781450330633. DOI: 10.1145/2647868.2655014. URL: https://doi.org/10.1145/2647868.2655014.

[431] I. D. Raji and J. Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 429–435.

[432] A. Todorov, F. Funk, and C. Y. Olivola. "Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences". In: *Trends in Cognitive Sciences* 19.8 (2015), pp. 422–423. DOI: 10.1016/j.tics.2015.05.013.

[433] European Parliament. *Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters*. 2021. URL: https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1678184&t=e&l=en.

[434] S. Castell, D. Cameron, S. Ginnis, G. Gottfried, and K. Maguire. "Public views of Machine Learning Findings from public research and engagement conducted on behalf of the Royal Society". In: (2017), p. 92. URL: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.

[435] J. W. Burton, M.-K. Stein, and T. B. Jensen. "A systematic review of algorithm aversion in augmented decision making". In: *Journal of Behavioral Decision Making* 33.2 (2019), pp. 220–239.

[436] S. Cave and K. Dihal. "Hopes and fears for intelligent machines in fiction and reality". In: *Nature Machine Intelligence* 1.2 (2019), pp. 74–78.

[437] S. Cave, K. Coughlan, and K. Dihal. ""Scary Robots": Examining Public Responses to AI". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 331–337. ISBN: 9781450363242. DOI: 10.1145/3306618.3314232. URL: https://doi.org/10.1145/3306618.3314232.

[438] C.-H. Chuan, W.-H. S. Tsai, and S. Y. Cho. "Framing Artificial Intelligence in American Newspapers". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 339–344. ISBN: 9781450363242. DOI: 10.1145/3306618.3314285. URL: https://doi.org/10.1145/3306618.3314285.

[439] E. Fast and E. Horvitz. "Long-Term Trends in the Public Perception of Artificial Intelligence". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 963–969.

[440] I. Hermann. "Beware of fictional AI narratives". In: *Nature Machine Intelligence* 2.11 (2020), pp. 654–654.

[441] [. O. Lords]. *AI in the UK: ready, willing and able?* 2018. URL: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

[442] S. Cave, C. Craig, K. Dihal, S. Dillon, J. Montgomery, B. Singler, and L. Taylor. "Portrayals and perceptions of AI and why they matter". In: (2018).

[443] M. Hohendanner, C. Ullstein, and M. Daijiro. "Designing the Exploration of Common Good within Digital Environments: A Deliberative Speculative Design Framework and the Analysis of Resulting Narratives". In: *Proceedings of the Swiss Design Network Symposium 2021 on Design as Common Good - Framing Design through Pluralism and Social Values*. Lucerne, Switzerland: SUPSI, HSLU, swissdesignnetwork, 2021, pp. 566–580. ISBN: 978-88-7595-108-5.

[444] E. Pierson. *Demographics and discussion influence views on algorithmic fairness*. 2018. arXiv: 1712.09124 [cs.CY].

[445] S. Engelmann, C. Ullstein, O. Papakyriakopoulos, and J. Grossklags. "What People Think AI Should Infer From Faces". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 128–141. ISBN: 9781450393522. DOI: 10.1145/3531146.3533080. URL: https://doi.org/10.1145/3531146.3533080.

[446] EmoVu. *EmoVu Mobile*. Accessed: 2022-02-25. n.d. URL: https://www.programmableweb.com/sdk/emovu-mobile.

[447] P. Ekman and W. V. Friesen. "Constants across cultures in the face and emotion". In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129. DOI: 10.1037/h0030377.

[448] Noldus. *Facial Expression Analysis*. Accessed: 2022-02-27. URL: https://www.noldus.com/facereader/facial-expression-analysis.

[449] Betaface. *Betaface API*. Accessed: 2022-02-27. 2021. URL: https://www.betafaceapi.com/wpa/.

[450] S. Biometry. *Face Recognition Demo*. Accessed: 2022-02-27. 2021. URL: https://skybiometry.com/demo/face-recognition-demo/.

[451] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. Moghaddam, and L. Ungar. "Analyzing personality through social media profile picture choice". In: *Conf. ICWSM*. 2016.

[452] M. Kasinidou, S. Kleanthous, P. Barlas, and J. Otterbacher. "I Agree with the Decision, but They Didn't Deserve This: Future Developers' Perception of Fairness in Algorithmic Decisions". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 690–700. ISBN: 9781450383097. DOI: 10.1145/3442188.3445931. URL: https://doi.org/10.1145/3442188.3445931.

[453] B. Zhang, M. Anderljung, L. Kahn, N. Dreksler, M. C. Horowitz, and A. Dafoe. *Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers*. 2021. arXiv: 2105.02117.

[454] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese. "In AI we trust? Perceptions about automated decision-making by artificial intelligence". In: *AI & SOCIETY* 35.3 (2020), pp. 611–623.

[455] E. Kaufmann. "Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models". In: *Computers and Education: Artificial Intelligence* 2 (2021), p. 100028. ISSN: 2666-920X. DOI: 10.1016/j.caeai.2021.100028. URL: https://www.sciencedirect.com/science/article/pii/S2666920X21000229.

[456] J. M. Logg, J. A. Minson, and D. A. Moore. "Algorithm appreciation: People prefer algorithmic to human judgment". In: *Organizational Behavior and Human Decision Processes* 151 (2019), pp. 90–103. DOI: 10.1016/j.obhdp.2018.12.005.

[457] A. Zerfass, J. Hagelstein, and R. Tench. "Artificial intelligence in communication management: a cross-national study on adoption and knowledge, impact, challenges and risks". In: *Journal of Communication Management* (2020). DOI: 10.1108/JCOM-10-2019-0137.

[458] M. K. Lee and S. Baykal. "Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division". In: *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 2017, pp. 1035–1048.

[459] F. Chollet. *Deep Learning with Python*. MANNING, 2018. ISBN: 9781617294433. URL: https://books.google.de/books?id=mjVKEAAAQBAJ.

[460] J. D. Rodrıguez-Garcıa, J. Moreno-León, M. Román-González, and G. Robles. "Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students". In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 177–183. ISBN: 9781450380621. URL: https://doi.org/10.1145/3408877.3432393.

[461] S. Rädiker and U. Kuckartz. *Analyse qualitativer Daten mit MAXQDA*. Wiesbaden, Germany: Springer VS Wiesbaden, 2019.

[462] U. Kuckartz. *Mixed methods: methodologie, Forschungsdesigns und Analyseverfahren*. Wiesbaden, Germany: Springer VS Wiesbaden, 2014.

[463] C. Züll and N. Menold. "Offene Fragen". In: *Handbuch Methoden der empirischen Sozialforschung*. Ed. by N. Baur and J. Blasius. Springer, 2019, pp. 855–862.

[464] U. Kuckartz. *Evaluation online: internetgestützte Befragung in der Praxis*. Wiesbaden, Germany: Springer VS Wiesbaden, 2009.

[465] K. Krippendorff. "Reliability in Content Analysis". In: *Human Communication Research* 30.3 (2004), pp. 411–433. DOI: 10.1111/j.1468-2958.2004.tb00738.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-2958.2004.tb00738.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2958.2004.tb00738.x.

[466] H. Kaiser. "An index of factorial simplicity". In: *Psychometrika* 39 (1974), pp. 31–36. DOI: 10.1007/BF02291575.

[467] E. Guadagnoli and W. F. Velicer. "Relation of sample size to the stability of component patterns." In: *Psychological Bulletin* 103.2 (1988), pp. 265–275. DOI: 10.1037/0033-2909.103.2.265.

[468] L. L. Thurstone. "Multiple-factor analysis; a development and expansion of The Vectors of Mind." In: (1947).

[469] L. Safra, C. Chevallier, J. Grèzes, and N. Baumard. "Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings". In: *Nature Communications* 11.1 (2020), pp. 1–7. DOI: 10.1038/s41467-020-18566-7.

[470] G. Brodny, A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. "Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions". In: *2016 9th International Conference on Human System Interactions (HSI)*. Portsmouth, UK: IEEE, 2016, pp. 397–404. DOI: 10.1109/HSI.2016.7529664.

[471] L. M. Schwartz, S. Woloshin, W. C. Black, and H. G. Welch. "The role of numeracy in understanding the benefit of screening mammography". In: *Annals of Internal Medicine* 127.11 (1997), pp. 966–972. DOI: 10.7326/0003-4819-127-11-199712010-00003.

[472] J. Estevez, G. Garate, and M. Graña. "Gentle introduction to artificial intelligence for high-school students using scratch". In: *IEEE Access* 7 (2019), pp. 179027–179036. DOI: 10.1109/ACCESS.2019.2956136.

[473] R. Hursthouse and G. Pettigrove. "Virtue Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University, 2018.

[474] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, and A. Rahmim. "A brief history of AI: how to prevent another winter (a critical review)". In: *PET clinics* 16.4 (2021), pp. 449–469.

[475] A. Chignell. "The Ethics of Belief". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2018. Metaphysics Research Lab, Stanford University, 2018.