Technische Universität München
TUM School of Management

Operationalizing AI Governance:

A Conceptualization and Implementation of Responsible AI in Technical and Organizational Processes

Ellen Karenina Victoria-Regina Hohma

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität München zur Erlangung einer

Doktorin der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Philipp Maume

Prüfende der Dissertation:

1. Prof. Dr. Christoph Lütge
2. Prof. Dr. Joachim Henkel

Die Dissertation wurde am 30.04.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Management am 15.10.2024 angenommen.

Acknowledgments

This dissertation not only marks the peak of one of the most intensive periods of my academic career to date, but also depicts the irreplaceable value of a network of support, guidance and encouragement from a group of remarkable people to whom I would like to express my deepest gratitude.

First, my endeavor would not have been possible, without my professor and supervisor, Prof. Dr. Christoph Lütge. His trust gave me the freedom to explore all the angles of academic opportunities while providing me with an invaluable platform and network for professional growth. I would also like to thank my second supervisor, Prof. Dr. Joachim Henkel, who joined my journey in the end and provided valuable feedback and support to give my thesis the finishing touches. Furthermore, my sincere appreciation goes to my coauthors, Dr. Anna Beer, Dr. Christian M. M. Frey, Prof. Dr. Thomas Seidl, Dr. Ryan Burnell, Caitlin Corrigan, Ph. D., and Prof. Dr. Christoph Lütge for supporting me through collaboration, guidance and, above all, their deep expertise. Thanks to you, all the effort never felt like real workload. A special thanks is also dedicated to the TUM Institute for Ethics in AI, particularly its executive director, Caitlin Corrigan, Ph. D. and all of my colleagues. The research, projects, events and, above all, travels we embarked on together have not only enriched my professional experience but have also left me with unforgettable memories.

A very special word of thanks goes to my mentor, Dr. Sebastian Planck, who has been a cornerstone of support and understanding throughout this process. Sebastian has always been there to listen to any self-doubts, providing the reassurance and perspectives needed to realize that these doubts were unfounded. Without your untiring support, this dissertation would not have been possible!

No less important than professional guidance, is the support I received from all my family and friends who supported me on this journey. Above all, my parents, Anni and Günter Hohma, who deeply supported me even when one or the other holiday was spent at the desk – my father with his clear and focused views that have been a major guidance in times of tough pressure, and my mother with her boundless empathy and warm-heartedness that have been my foundation and a source of peace. Furthermore, I would like to thank my sister, Julia Hohma, who always has an open ear, knowing my thoughts even before I know them and my grandma Lotte Hohma – danke, dass du mir immer so fleißig alle Zeitungsartikel zum Thema KI mitbringst!

Finally, words cannot express my gratitude to Lennart Rümmler, my most loyal reader, most reliable source of inspiration and, even more important, my best friend, partner in crime and life companion. Without you neither my dissertation nor myself could grow to the needed level of excellence to realize this as well as the many life goals I have achieved with you by my side. Thank you for making me the best version of myself!

To all who have supported me on this journey, please accept my deepest thanks. This dissertation is not just a reflection of my work, but a testament to the incredible support system that I am fortunate enough to have by my side.

Abstract

As both the capabilities of as well as interest in Artificial Intelligence (AI) have increased and have boosted its prevalence, calls for effective AI governance tend to dominate the conversations about this promising technology. To follow those, steps have been initiated on a regulatory as well as standardization level to govern risks related to AI practices. While many of the developed approaches have been well received, practitioners continue to encounter difficulties in applying them in the field as they face the challenge of operationalizing the often-abstract concepts to concrete and actionable measures. The aim of this dissertation is therefore to support practitioners in this urgent task by contributing through three published research articles, each aimed at supporting the clarification and contextualization of existing fundamentals. The first article synthesizes and analyzes obligations and measures proposed in current regulation, standardization or other research initiatives, with the underlying goal of translating them to a trustworthy process for AI development. Articles 2 and 3 advance and contextualize identified methods under certain thematic and technological contexts. Specifically, Article 2 explores the theoretical definition of one responsible AI principle fairness, and its translation into applicable statistical models in the field of public health surveillance. Article 3 studies a second trustworthy AI principle, robustness, showcasing its technical implementation in the context of enhanced spectral clustering techniques. The value of this thesis hence lies in the demonstration that operationalization of AI Governance is both essential and achievable, with developed principles already nearing practical alignment, and offering roads to foster contextualization proving that they are accessible. In this way, the ultimate aim of this thesis is to support moving responsible AI concepts beyond mere conversations and become what they should be seen as – necessary tools in the development of AI.

Table of content

1 Introduction 1

2 Emerging Foundations of AI Governance 7

2.1 Regulatory Initiatives on AI Governance 7

2.1.1 Fundamental Approaches in AI Regulation 8

2.1.2 The EU Regulatory Landscape Linked to AI 11

2.2 Industrial Standards for AI Governance 21

2.2.1 Foundational Standardization Work 22

2.2.2 Current Efforts in Standardization of AI Governance Operationalization 23

3 Implications, Problems and a Way Forward for the Operationalization of AI Governance Concepts 26

4 Contribution to Advancing AI Governance to Business Practice 29

4.1 Extended Abstract Article 1 – From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems 30

4.2 Extended Abstract Article 2 – Individuality and Fairness in Public Health Surveillance Technology: A Survey of User Perceptions in Contact Tracing Apps 32

4.3 Extended Abstract Article 3 – SCAR: Spectral Clustering Accelerated and Robustified 34

5 Discussion 37

6 Conclusion 41

1 | Introduction

> *“In an era where the mere mention of AI evokes a blend of excitement and trepidation, the relentless ascent of artificial intelligence persists as both a captivating marvel and a source of apprehension. The tech realm has witnessed the unfolding of AI's prowess, with generative AI pushing the boundaries of innovation. Yet, amidst the promises of groundbreaking advancements lie the shadows of risks and uncertainties. AI, like any powerful force, demands our attention and diligence in steering its trajectory. In this age of unprecedented technological acceleration, the call to address and manage the inherent risks echoes louder than ever.”*
>
> *– ChatGPT 3.5*

In times when catchy introductory quotes for a thesis no longer need to be searched for among the statements of clever minds but can be generated precisely and on-demand in just a few seconds, the potential and capabilities of artificial intelligence (AI) technologies are well understood. At the same time, while the potential of human-like but machine-produced outputs reach seemingly limitless spheres, the risks, as ChatGPT itself has acknowledged, that such technologies can entail are almost as undeniable. It is therefore not surprising that while businesses, the media and even general public are rushing to experiment with astonishing capabilities of the latest AI functionalities, a second movement has evolved in parallel, which is nearly as rapidly gaining momentum as the technological innovation itself. The call to steer the developed AI capabilities in a desirable direction is rarely not mentioned within the same breath as its benefits – its resolution, however, is seen among the biggest challenges that AI brings. The core topic of this dissertation, the responsible development of AI, is therefore both a major AI buzzword and the much-needed solution to a central AI conundrum, to which this work aims to contribute.

The idea of managing innovation in order to promote good practices and minimize the resulting risks yet – of course – did not first emerge with AI. Rather, this fundamental principle lies at the core of the field of technology or IT governance. Long before the rise of AI, continuous innovation and in particular rapid progression in new technologies has facilitated the need to develop structured frameworks and methodologies for their effective management. Developing approaches for aligning information technology initiatives with overarching corporate goals and strategies has been seen as a critical point for ensuring that IT investments and activities contribute directly to the achievement of an organization's mission and vision. Their value lies in the optimization of IT practices and resources, seeking efficiency and cost-effectiveness while proactively identifying and mitigating associated risks.

The resulting discipline of technology governance hence describes the allocation of decision-making authority and responsibilities related to information technology among various stakeholders within an enterprise, as well as the establishment of procedures and mechanisms for formulating and overseeing strategic IT decisions (Peterson, 2004). Consequently, it serves as the framework through which an organization can guide and oversee both the present and future use of IT resources. Thereby, technology governance extends beyond the purview of individual organizations and regulatory bodies, encompassing a broad spectrum of institutional and normative mechanisms aimed at steering the development of technology (OECD, 2024).

More practically, concepts for organizational technology governance have been formulated and are in use – currently mostly without an explicit reference to AI technologies. For example, Mohamed and Kaur a/p Gian Singh (2012) identified three fundamental dimensions of IT Governance: structure, process and relational elements. The structure dimension encompasses the organizational units, roles and responsibilities involved in the decision-making processes related to IT, including the overall organizational setup of IT and its associated departments (Mohamed & Kaur a/p Gian Singh, 2012). Processes encompass the various activities that ultimately lead to the formulation and execution of IT strategies. These activities include, for example, risk management, performance management, including the establishment and monitoring of Key Performance Indicators, project management and information security (Alreemy et al., 2016; Mohamed & Kaur a/p Gian Singh, 2012). Finally, relational aspects of IT Governance include all dissemination and communication efforts to share the strategic goals, principles and policies with related actors (Mohamed & Kaur a/p Gian Singh, 2012). Numerous frameworks to guide the implementation of these dimensions have been developed, with ITIL or ISO 20000, ISO 17799, ISO 27000 or other security frameworks, Six Sigma, COBIT, PMI/PMBOK, Risk IT (ISACA) and IT Assurance Framework (ISACA) among the most popular ones (Smits & Hillegersberg, 2013).

While IT governance thus appears like a well-researched – even though certainly not yet exhausted – field, the question arises as to why AI governance in the beginning of this thesis has been introduced as one of the grand challenges of AI development. To clarify this, looking closer into what distinguishes AI technologies from ‘traditional’ may help. Many have tried to find a precise definition for AI, and few have succeeded – a surprising number of scholars even argues for ‘no one’. At the heart of proposed definitions often lies the “capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning” (ISO/IEC, 2015). A core common ground is hence found in the imitation of human cognitive capabilities through machines (McCarthy, 2007) combined with certain degrees of automation and progression. Such capabilities are often linked to computer vision, computer audition, computer linguistics, advanced robotics and control, forecasting, discovery, planning and creation (EIT Community, 2021; Samoili et al., 2020). It is precisely this power to predict and learn, and thus the capability of inferring or generating new knowledge, that makes AI technologies special compared to traditional ones. Autonomous decision-making and unpredictable learning behavior require special provisions for, for example, transparency and human oversight (Bartneck et al. 2021). These characteristics have become increasingly important in the context of AI and have therefore not been adequately considered in traditional IT governance frameworks, yet. In other words, while traditional IT governance concepts are not completely outdated, there are specific characteristics of AI for which the traditional concepts do not provide the necessary specification and guidance. Identifying these and incorporating appropriate additions into existing technology governance concepts is hence a main objective of current AI governance initiatives.

In order to define suitable objectives and the related mechanisms and measures required to mitigate AI-specific risks, it is necessary to define the principles underlying these key fundamentals. Within the EU a high focus is placed on the cultivation of an ecosystem characterized by excellence and trust in order to facilitate the responsible and secure utilization of AI within Europe, as emphasized by the European Commission (AI HLEG, 2019). To fulfill the primary objective of preventing violations of fundamental rights and Union values, particularly AI applications that fulfill the requirements for lawful, safe and trustworthy AI are deemed acceptable (AI HLEG, 2019). Similarly, the US White House Office of Science and Technology Policy (OSTP) in its "Blueprint for an AI Bill of Rights" places civil liberties and democratic principles at its core, with a particular emphasis on safeguarding against threats to civil rights, personal liberties, privacy, equitable opportunities, and access to essential resources and services (OSTP, 2022). Such foundational values are also supported by international organizations, such as in the UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021). This recommendation underscores the imperative that AI systems must operate for the benefit of humanity, individuals, societies, the environment and ecosystems while preventing harm (UNESCO, 2021).

These examples signify a collective consensus – at least within Western-oriented societies – regarding the foundations of anticipated behavior of AI, revolving around advancing AI for the greater good while mitigating the potential for AI-induced harm. Followingly, several efforts have been made to translate this foundational concept into a more practical definition. Within the EU, particularly the ethical guidelines for trustworthy AI set forth by the High-level Expert Group on Artificial Intelligence (AI HLEG, 2019) have gained large traction and are therefore often cited as a primary point of reference. They introduce the concept of trustworthy AI and outline three important components: legal, ethical and robust AI. Trustworthy AI is thereby seen as the primary goal to clarify and unify ethical considerations and to uphold "fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter)" (European Parliament, 2000, p. 6) as well as international human rights.

Such core ideas are further concretized with the large field of research that has revolved around the creation of actionable principles to realize ethical AI. Many frameworks to detail the ethical or responsible design, development and use have been set forth by a variety of actors, such as the AI4People (Floridi et al., 2018), OECD (2019) or UNESCO (2021). A high-level consensus around the principles of transparency, justice and fairness, non-maleficence, responsibility and privacy has been detected (Jobin et al., 2019). In a similar way, these principles are reflected in the considerations of the AI HLEG, which suggest 7 key requirements for trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability (AI HLEG, 2019).

From this introductory digression into the need and evolution of AI governance it can be discerned that the impact of AI-specific risks has been acknowledged, and fundamental frameworks to define principles alongside desired anticipated system behavior have been proposed, discussed and agreed to provide the basis for duly developed AI systems. As we are now in a stage where required foundations exist, the primary objective of current research in AI governance has shifted towards their operationalization in practical and business contexts. Therefore, the objective of this thesis is, based on the current state of AI governance operationalization, to determine the necessary, however, up-to-now unfulfilled paths for the practical application of AI governance particularly from the perspective of AI providers as well as to contribute to their fulfillment. More specifically, through exploring the tangible approaches for addressing responsible AI considerations under certain thematic and technological contexts, that and how this facilitates their practical implementation shall be demonstrated. The findings are presented across three published research articles. In Article 1, the foundation is set out by deriving essential obligations that AI providers hold in the implementation of responsible AI as well as measures to address them. Articles 2 and 3 dive deeper into contextualization, studying identified methods under certain thematic and technological contexts. The significance of this thesis hence lies in demonstrating that operationalizing AI Governance is not only vital but also feasible, with established principles already approaching practical application. Moreover, it offers pathways for their contextualization. Thus, this thesis supports the advancement of responsible AI from abstract discussions to practical tools in AI development.

To present the value and contribution of the work in this thesis, the emerging fundamentals of AI governance are outlined in Section 2. This incorporates the foundations that have been laid by regulation and standard developing organizations. A particular focus in this overview is placed on the implications such fundamentals have for the AI provider to determine the essentials for AI governance operationalization in practice. The issues that remain in this regard and the demanded specifications to address those are summarized in Section 3. They are the point of reference for the contribution which is outlined in Section 4. Finally, in Section 5, the magnitude and significance of the contribution is discussed and concluded in the form of an outlook to fundamental challenges in Section 6.

2 | Emerging Foundations of AI Governance

The particular challenges and risks associated with AI are increasing attention to AI governance approaches. To address the complexities that arise with their resolution, essentially two key types of players provide the groundwork for unifying standardized approaches to guide AI providers in their AI governance efforts. Governmental and regulatory efforts establish the foundational principles on which Standardization Developing Organizations (SDOs) build industry-specific guidance detailing the proposed or mandatory requirements.

To examine the current state-of-the-art in AI governance, in the following, I review recent work from the mentioned two types of key player to discern the specific requirements they impose on organizations' AI governance concepts. The objective here is to present the foundational aspects of AI governance to comprehend the groundwork upon which the work of this thesis is based. Consequently, the analysis in this section places particular focus on outlining the implications that regulation and standardization hold for AI providers, which builds the basis for a more profound understanding of the distinctive contribution that this thesis makes to the transition of AI governance from theory to practice.

2.1 Regulatory Initiatives on AI Governance

The field of AI is evolving rapidly and, consequently, a growing demand for regulatory measures to govern it is being articulated. While an increasing number of propositions for AI regulation and guidance are being put forward, no uniform approach has (yet) prevailed. AI providers hence find themselves in a challenging position with the need to operate within the framework of existing regulations, which are only gradually adapting to accommodating the new AI-based use cases and specialized regulations emerging in the AI domain. These regulatory efforts, while intended to provide clarity and structure, also impose specific responsibilities on AI providers as they work to establish robust AI governance frameworks.

The broader global initiatives aimed at regulating AI and their potential implications for the governance of AI providers are outlined in the following, to provide an overview of fundamental governmental approaches to guiding AI development. Furthermore, a detailed examination of the regulatory net is presented along with its implications for AI providers within the EU – a leading player in the area of responsible AI having proposed advanced and extensive regulatory provisions – to better understand its concrete impacts on the governance practices of AI providers.

2.1.1 Fundamental Approaches in AI Regulation

With the urgent need and frequent calls for more unified guidance on approaching ethical and societal concerns around AI, governments around the globe have initiated policy efforts to recommend or regulate the responsible development and use of AI. A survey from Stanford University's Institute for Human-Centered Artificial Intelligence has found that in 2022, legislative bodies from the examined 127 countries have passed 37 laws that at least contained the expression 'Artificial Intelligence' (Lynch, 2023). Many countries have published national AI strategies to set out a roadmap for innovation and regulation regarding this emerging technology. Many of them emphasize the important and disruptive role of AI (Bareis & Katzenbach, 2022) and lay down strategic goals on how to strengthen scientific innovation, retain and attract AI-skilled for the future and promote industrial uptake, for example, through sectoral programs (Radu, 2021). Naturally, however, they differ in the readiness and progressiveness of approaches as well as in priority-setting in the envisaged pathways to reach the aspired goals (Bareis & Katzenbach, 2022; Wilson, 2022). For example, while some countries, such as Germany, Finland, France, or South Korea, prioritize security and societal values by introducing supervisory or control measures, others such as China, the UK, the US, or the UAE place more emphasis on innovation by pursuing goals such as becoming leaders or first-buyers in AI technology (Radu, 2021).

Particularly, countries that take a more cautious approach regarding AI technologies are likely to act by building up a net of regulatory efforts to adapt to AI. However, also those striving for technological leadership acknowledge the requirement of clearly defined pathways to allow innovators to operate in accordance. Of the planned and communicated regulatory efforts, thus, two fundamental streams can be observed: a *strong regulation approach*, where the countries envision public intervention and thus the establishment of hard rules around the use of AI, and a *soft regulation approach*, where no specific regulatory action is foreseen or provisions only have minor binding character, as visualized in Figure 1.

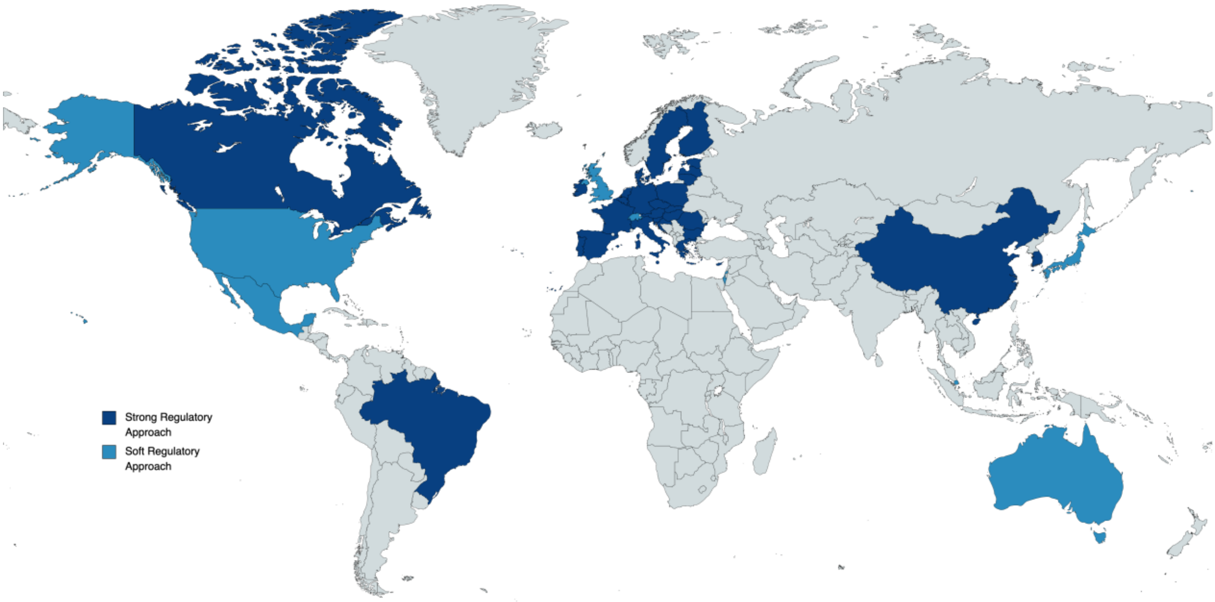


Figure 1: Selection of AI regulation approaches around the globe as of late 2023.

Strong Regulatory Approach

Several countries around the world have opted for strict regulation of AI technologies and developed dedicated laws to control the associated risks. Highly propelling approaches came, for example, from Brazil, Canada, China and the EU. Often a risk-based approach is chosen to account for potential harm to safety or societal values that might arise with the introduction of AI systems. On an EU level, the developed AI Act categorizes AI systems into 4 risk levels, banning AI systems that pose unaccepted risks and demanding further risk prevention measures alongside a conformity assessment for systems that are determined as highly risky. The Brazilian government in its AI Bill suggests a highly similar approach by mandating a preliminary assessment of an AI system from AI providers, evaluating the system's potential to create harm into the categories of 'excessive' or 'high' (Access Partnership, 2023). Similar to the EU approach, AI systems classified as posing excessive risks are not permitted and additional obligations are required for systems of high risk. A less strict, but still restrictive approach is taken by Canada in its proposed Artificial Intelligence and Data

Act (AIDA) to ensure safe, non-discriminatory and accountable AI systems (Government of Canada, 2023). A particular focus is set on holding businesses responsible for their AI activities and obliging them to implement appropriate risk governance mechanisms as well as enabling users to make informed decisions based on suitable information and transparency. Finally, the People's Republic of China has proposed and already put into force several provisions to regulate certain use cases of AI, such as algorithmic recommendations or generative AI, as well as promoting the development of the AI industry (Latham & Watkins, 2023). Their primary objective is to address risks that come along with AI-generated content, such as deep fakes, and to protect national and social security in China.

Governments with a strong regulatory approach to AI governance, therefore, provide substantive guidance, as they clarify fundamental risks or red flags as well as how to approach them, supporting AI providers in the creation of AI governance concepts. At the same time, however, potentially binding rules are imposed on AI providers, creating additional barriers and burdens. Precautionary countermeasures may be obligatory to mitigate potential risks and ensure responsible development, deployment and use of their AI systems. A standardized approach with public oversight could thus come at the price of higher costs for AI providers or impeded innovation. On top, such strong and strictly enforced rules create additional market entry barriers for new players, potentially reinforcing the formation of some form of AI monopolies. These trade-offs have hence led to a second set of less stringent regulatory approaches.

Soft Regulatory Approach

A softer regulatory approach can be seen with governments opting not to produce AI-dedicated regulations that set out legally binding measures but instead focus on updating existing legislation to also account for AI-related risks. In addition, some develop specific guidelines for the development and use of AI that, however, do not foresee any enforceable mandates.

Switzerland, for example, has chosen to selectively update existing regulations. Adjustments are made, for instance, to data protection laws aiming to demand the increase of transparency within AI systems, or further legislations, such as the General Equal Treatment Act, competition law, product liability law and general civil law to establish the necessary guidance for AI technologies (Kohn & Pieper, 2023). Similarly, although the country has taken a proactive stance in the regulation of AI, there is currently no dedicated AI regulation foreseen in Japan, and recently the authorities have unofficially indicated that they are leaning towards a softer approach to regulating AI (Nussey & Kelly, 2023). Such an approach could be modelled on the United States, which has published a draft AI Bill of Rights at the national level (OSTP, 2022). This sets out five principles for mitigating AI-related harms, accompanied by a technical manual that provides non-binding guidance on implementation and relies on voluntary compliance by AI providers. While further regulatory efforts have been made to create an Algorithmic Accountability Act that prescribes concrete and legally binding measures, it has not yet been enacted and its passage in the House and Senate is in doubt (Holistic AI, 2023).

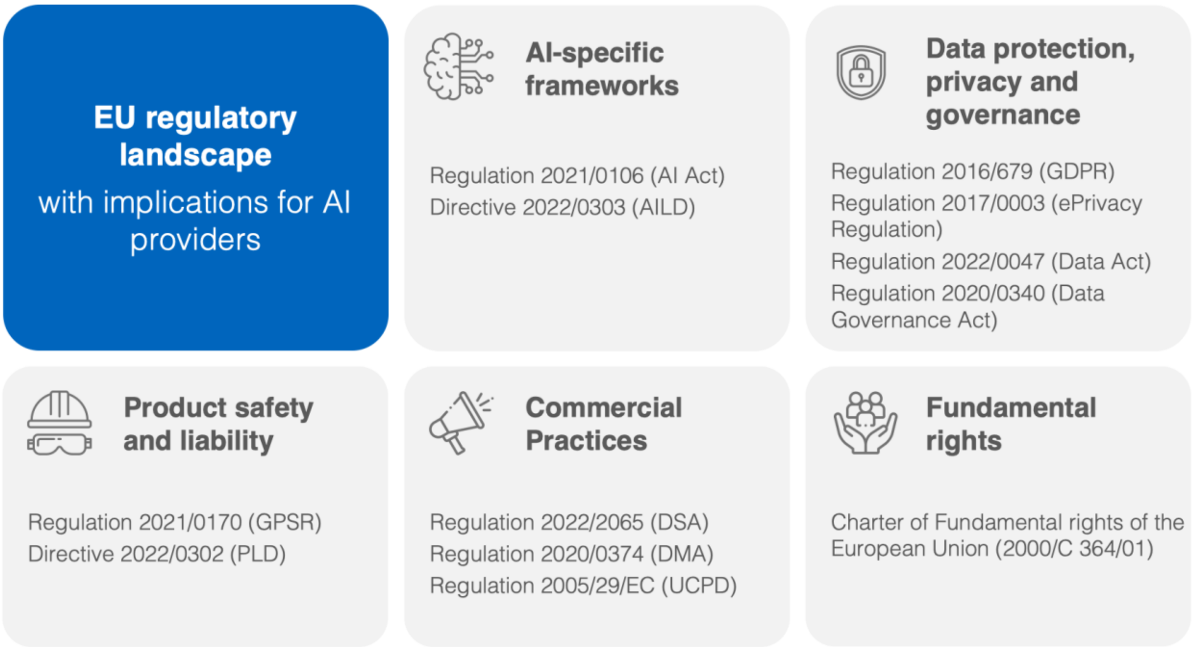
The implications of such a soft regulatory approach for AI providers, although not mandatory, are nevertheless significant. While governments recognize the need to tailor AI governance strategies to specific contexts by offering guidance, they leave the final decision-making in the hands of the AI provider. This approach grants the AI provider more flexibility in adapting its policies to its unique needs and circumstances, however, it also places the responsibility firmly on the shoulders of the AI provider. This implies the formulation of their own strategies as well as ensuring their effectiveness and compliance with ethical standards.

2.1.2 The EU Regulatory Landscape Linked to AI

The European Commission has set itself a leading role in the development of human-centered and trustworthy AI. This proactive stance reflects the EU's commitment to safeguarding individual rights, Union values and ethical considerations in the AI domain. In view of their forward-thinking initiatives, some even anticipate a potential "Brussels effect", akin to the General Data Protection Regulation, where considerations originating from the EU may influence forthcoming legislative measures in other jurisdictions as well (Voss, 2023). The EU’s stringent approach to AI regulation and the resulting significant influence on the global AI regulatory landscape provide valuable insights into the evolving standards and practices and therefore warrant a detailed investigation. Particularly, it can support the concretization of the implications of AI regulations on the AI governance activities of AI providers.

To achieve its goal of responsible or trustworthy AI, a wide range of legal documents and frameworks have been proposed or are being developed within the EU. Particularly, three inter-related legal initiatives are foreseen: (1) “a European legal framework for AI to address fundamental rights and safety risks specific to the AI systems” (European Commission, 2022b), the AI Act, (2) a civil liability framework, i.e., an updated Product Liability Directive to better adapt to new technologies including AI and (3) “a revision of sectoral safety legislation” (European Commission, 2022b), including e.g., the Machinery Regulation or the General Product Safety Directive (European Commission, 2022b).

While these provisions are currently partially under development, there is a range of existing regulations in place that AI-based systems, just like ‘regular’ products or services, must comply with. 5 primary fields that impact AI systems and lead to direct or indirect obligations for AI providers are summarized in Figure 2.



EU regulatory landscape with implications for AI providers

AI-specific frameworks

Regulation 2021/0106 (AI Act)
Directive 2022/0303 (AILD)

Data protection, privacy and governance

Regulation 2016/679 (GDPR)
Regulation 2017/0003 (ePrivacy Regulation)
Regulation 2022/0047 (Data Act)
Regulation 2020/0340 (Data Governance Act)

Product safety and liability

Regulation 2021/0170 (GPSR)
Directive 2022/0302 (PLD)

Commercial Practices

Regulation 2022/2065 (DSA)
Regulation 2020/0374 (DMA)
Regulation 2005/29/EC (UCPD)

Fundamental rights

Charter of Fundamental rights of the European Union (2000/C 364/01)

Figure 2: Overview of the identified primary fields of the EU regulatory landscape that impact the responsibilities of AI providers.

AI-specific frameworks

The Regulation of the European Parliament and of the Council for Laying Down Harmonized Rules on Artificial Intelligence and Amending Certain Union Legislative Acts, or short, the AI Act, has been proposed in an effort to promote legal clarity and harmonization on the development and use of AI in the EU as well as facilitate lawful, safe and trustworthy AI applications within its borders. To reach this, a four-level risk categorization into unacceptable, high, limited and minimal risk systems has been developed, foreseeing additional safeguards and measures for AI systems depending on their risk classification.

Unacceptable AI systems are set out in a conclusive list of prohibited practices. It includes AI systems that deploy subliminal manipulative techniques (AI Act, Art. 5(1)(a)), exploit certain vulnerabilities, e.g., due to a person's age or disability (AI Act, Art. 5(1)(b)) or are used as biometric categorization systems to deduce sensitive personal attributes, such as race, political opinion or religious beliefs (AI Act, Art. 5(1)(b)(a). Further, AI systems that are used for evaluation or classification of social behavior are prohibited, particularly if they lead to unfavorable treatment in social contexts that are unrelated to the initial data collection context or that are particularly disproportionate (AI Act, Art. 5(1)(c). In addition, "'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement" (AI Act, Art. 5(1)(d)) are generally prohibited unless for three legitimate objectives, and then with additional protection measures. Legitimate objectives include the targeted search for crime victims (AI Act, Art. 5(1)(d)(i)), prevention of life or physical safety, in particular, in connection with a terrorist attack (AI Act, Art. 5(1)(d)(ii)) and the localization of person suspected of having committed a criminal offence (AI Act, Art. 5(1)(d)(iii)) under certain conditions. Finally, systems shall be prohibited that are used to assess the risk of natural persons to commit a criminal offence (AI Act, Art. 5(1)(d)(a)), create facial recognition databases through scraping facial data (AI Act, Art. 5(1)(d)(b)), or infer emotions especially when they are used in the workplace or in education (AI Act, Art. 5(1)(d)(c)). These prohibited uses of AI will therefore no longer be permitted in whole or in part in the EU if the AI Act comes into force in its current form.

The most extensive category in terms of restrictive measures for AI providers are *high-risk* systems. Essentially, there are two types of classification as high-risk system. First, Art. 6(1) of the AI Act specifies that AI systems are classified as high-risk if they are a product or used as a safety component of a product which is required to undergo a third-party conformity assessment pursuant to Union harmonization legislation listed in Annex II. This includes machinery (Directive 2006/42/EC), toys (Directive 2009/48/EC) and medical devices (Regulation 2017/745/EU). Second, AI systems are also considered high-risk if they are used in one of the areas listed in Annex III. This listed areas include: (1) biometric identification, categorization and emotion recognition , (2) management and operation of critical infrastructure, (3) education and vocational training, (4) employment, workers management and access to self-employment, (5) access to and enjoyment of essential private services and public services and benefits, (6) law enforcement, (7) migration, asylum and border control management and (8) administration of justice and democratic processes (AI Act, Annex III). However, in the newest version of the AI Act, legislators have specifically included exceptions where AI systems falling in this second type of high-risk "do not pose a significant risk of harm, to the health, safety or fundamental rights of natural persons" (AI Act, Art. 6(2a)). This is the case if the AI system is used to narrow procedural tasks (AI Act, Art. 6(2a)(a)), improve a previously completed human activity (AI Act, Art. 6(2a)(b)), only used to detect decision-making patterns but not replace a human assessment (AI Act, Art. 6(2a)(c)), or only performs preparatory tasks under certain conditions (AI Act, Art. 6(2a)(d)). Furthermore, the provided list is non-conclusive, i.e., further cases might be removed or added through delegated acts by the Commission under certain conditions.

Additional safeguards are required for these high-risk systems if they are to be introduced or operated in the EU. Ensuring compliance with the requirements as well as additional tasks are shared among the different stakeholders (product manufacturers, authorized representatives, importers, distributors, users, or any other third party) to varying degrees, whereby the AI provider, bears the majority of obligations (Veale & Borgesius, 2021).

The level of sometimes called *limited-risk* AI systems was significantly adjusted during the development of the AI Act, particularly with the introduction of Large Language Models such as ChatGPT and the associated decision as to whether or not to respond to this evolution. In the newest version, Article 52 of the AI Act sets out "transparency obligations for providers and users of certain AI systems and GPAI models". Particularly, four use cases are mentioned. First, for applications where a natural person is interacting with an AI system, the AI provider is obliged to disclose this interaction if it is not yet obvious from the circumstances and context of use (AI Act, Art. 52(1)). This is sometimes referred to as 'bot disclosure' obligation (Veale & Borgesius, 2021). Second, providers of AI systems that generate synthetic audio, image, video or text content, including their generation by General Purpose AI (GPAI) systems, are required to mark the outcomes as artificially generated or manipulated (AI Act, Art. 52(1a)). Further technical obligations regarding the system's effectiveness, interoperability, robustness and reliability are demanded (AI Act, Art. 52(1)(a)). On top, the legislator saw a particular need to react and included Title VIII (a) forseeing additional obligations for providers of GPAI Models, in particular, where they are expected to pose systemic risk. A third use case within the limited-risk category are AI systems used for emotion recognition or biometric categorization. Their providers are obliged to inform natural persons who are exposed to them and to process their data in accordance with the relevant Union law (AI Act, Art. 52(2)). Finally, deepfakes are regulated in Art. 52(3) (AI Act) and are required to be marked as artificially generated or manipulated.

The fourth and final risk level, *minimal-risk* systems, is sometimes not even quoted as such, as it comprises all remaining uses of AI. Providers of minimal-risk AI systems are free of binding measures and are only recommended to provide and follow self-imposed codes of conduct to voluntarily commit to the same response measures as high-risk systems (AI Act, Art. 69).

Product safety and liability

Lately, there has been big upheaval in the regulation of product safety and liability in the EU. Both primary legislation documents, the General Product Safety Directive (GPSD) and Product Liability Directive (PLD), have been under revision in order to more accurately adapt them to emerging technologies of the digital age and circular economies (European Commission, 2022d). Serving as two complementary mechanisms for the enforcement of consumer claims for damages, the PLD establishes liability for claims arising from a defective or unsafe product, with product safety legislations setting out the conditions that a product must meet in order to be considered safe. This can be either specified in sector-specific legislation (e.g., for Machinery, Directive 2006/42/EC, or Toys, Directive 2009/48/EC), or, in the absence of such pertinent provisions, by general provisions set out in the GPSR. In the case of AI, the newly introduced AI Act serves as such a dedicated regulation and can therefore provide explicit guidance. However, in the case of systems for which the AI Act does not impose specific requirements, i.e. in particular for products where the AI component is considered to pose only a minimal risk, the GPSR explicitly “provides a safety net for products and risks to health and safety of consumers that do not enter into the scope of application of the AI proposal” (GPSR). AI-equipped products that are not subject to the more specific safety rules, e.g., as they do not fall within the high-risk category of the AI Act, hence, must comply with the provisions of the GPSR (Almada & Petit, 2022).

Followingly, safety regulations and in particular the GPSR shall clarify the general safety that can be expected for consumer products within the EU. In the new draft, a product is defined as “any item, interconnected or not to other items, supplied or made available, whether for consideration or not, in the course of a commercial activity including in the context of providing a service – which is intended for consumers or can, under reasonably foreseeable conditions, be used by consumers even if not intended for them” (GPSR, Art. 5, Nr. 1). Therefore, consumer products in the form of physical items that include an AI component, such as smart speakers equipped with virtual assistant technologies are in principle in scope of this definition. The general safety requirement, obliging economic operators to “place or make available on the Union market only safe products” (GPSR, Art. 5), thus applies. Fulfilling this requirement can either be presumed, for example, if conforming to relevant European standards (GPSR, Art. 6), or must be assessed along predefined aspects (GPSR, Art. 7). Obligations to ensure that the product is in accordance with the general safety requirement are imposed onto the different economic operators involved in the product development. While AI providers, or more generally component/service providers, are not explicitly mentioned, in cases where the AI provider is not the product manufacturer, implicit duties may follow from the manufacturer’s obligation to ensure that “products have been designed and manufactured in accordance with the general safety requirement” (GPSR, Art. 8, Nr. 1).

If the required safety standards are not met, a right to compensation for natural persons can follow from provisions set in the PLD for damages caused by the defective product (PLD, Art. 5). Damages in the new draft are considered material losses caused by death or personal injury, harm or destruction of property and loss or corruption of data (PLD, Art. 4(6)). A product’s defectiveness shall be considered along various aspects, including its cybersecurity and “any ability to continue to learn after deployment” (PLD, Art. 6). The new EU legislation therefore reacts to the many concerns that have been articulated regarding the old PLD version and its applicability in the context of AI. Borges (2021), for example, has raised the issue of enforcing claims for damages other than death or injury with the old PLD version, which is now acknowledged by expanding the definition of damage to data-related issues and psychological harms (if medically recognized). Similarly, the burden to prove the defectiveness of the product, the damage suffered and the causal link, which lies with the claimant and has been regarded as challenging in the old PLD version (Borges, 2021), has now been updated, e.g. with regard to a lightening of the burden of proof if the plaintiff has excessive evidentiary difficulties due to technical or scientific complexity (PLD, Art. 9(4)). This leads to the product manufacturer having the explicit duty to disclose relevant evidence at its disposal (PLD, Art. 9(2)(a)) and the implicit duty to ensure the outlined safety requirements in order to avoid the need for compensation payments.

It is worth emphasizing once again that the PLD regulates products and in particular worth highlighting that the definition of a 'product' has also been amended to now explicitly include software in its application scope. Nevertheless, the legislator saw a particular need for clarification in relation to AI and supports the PLD with an AI Liability Directive (AILD), that explicitly details obligations and liabilities for AI systems, such as provisions on access to information on high-risk systems or adaptations to the burden of proof (AILD, Art. 1(1)). The AILD establishes, under certain circumstances, a right for victims to disclosure of "relevant evidence at [the provider's] disposal about a specific high-risk AI system that is suspected of having caused damage" (AILD, Art. 3(1)). In addition, a presumption of causality between the defendant's fault and the AI system's outputs is established, under certain circumstances, such as a demonstrated non-compliance with a duty of care (AILD, Art. 4(1)(a)), to ease the victim's responsibility to explain exactly how a harm was caused (European Commission, 2022c). For high-risk AI systems, these duties of care are linked to requirements laid down in the AI Act, including, for example, data quality criteria or transparency obligations (AILD, Art. 4(2)). As a result, besides the obligation to provide access to information on high-risk AI systems upon request, the AILD may impose additional consequences for providers of high-risk AI systems in case of non-compliance with provisions set out in the AI Act.

Data protection, privacy and governance

Implications of the EU legal landscape of data protection and governance on AI systems can be considered in two categories of data, personal data and non-personal data. Personal data are "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (GDPR, Art. 4, Nr. 1). With non-personal data, this connection to a natural person is not given, and therefore it is "any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audio-visual recording" (Data Act, Art. 2, Nr. 1). For both categories of data, there are regulatory documents, for which implications for AI systems and their providers shall be outlined below.

Like other techniques for the "processing of personal data wholly or partly by automated means" (GDPR, Art. 2, Nr.1), AI systems that handle *personal data* generally fall within the scope of Regulation 2016/679, the General Data Protection Regulation (GDPR). Processing this type of data is only permitted under additional safety, privacy and transparency requirements along with granting special rights to the respective data subject to allow them to demand certain handling of their data from the data operator. One category is particularly singled out. The processing of 'special personal data categories' that relate to "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" (GDPR, Art. 9) is prohibited unless under certain legitimate conditions. To ensure compliance with the requirements and enable data subjects to exercise their granted rights, special obligations for protective measures result from the GDPR. Among other actors, most obligations are imposed on data controllers, i.e., the "natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (GDPR, Art. 4, Nr. 7) and data processors, i.e., the "natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller" (GDPR, Art. 4, Nr. 8). For AI providers this means the extent of power over the data, e.g., whether data collection is conducted by the AI provider or pre-collected data is used from third-party suppliers, determines which responsibilities arise.

As mentioned, the GDPR foresees particular provisions where processed data qualifies as personal. In addition, there are further EU regulations that generally concern the use of data, i.e., regardless of whether they are *personal or non-personal*, that affect the AI provider, even if to a limited extent. In the field of electronic communications, Regulation 2017/0003 on Privacy and Electronic Communications (ePrivacy Regulation) restricts interference with electronic communications data, e.g., in the form of listening, tapping, storing, or monitoring, to certain permitted use cases (ePrivacy Regulation, Art. 5). Two acts address obligations regarding use of data: Regulation 2022/0047 on harmonized rules on fair access to and use of data (Data Act) which sets out obligations for the provision of data generated by the use of (physical) products that collect and transmit data and Regulation 2020/0340 on European data governance (Data Governance Act) regulating the reuse and sharing of data between stakeholders in the EU to strengthen data availability and exchange. This legislation may have an impact on AI providers if the system falls under one of the covered use cases or if the AI provider is involved in further data collection or post-processing activities. However, their set out obligations are linked to the overall system and hence the direct impact on the AI component is rather limited.

Commercial practices

Similar to rules on safety and liability, the area of commercial practices has experienced significant updates with regard to EU regulation in recent years. Particularly, legislation for digital services has been extensively renewed to further strengthen consumer rights and trust. Regulation 2022/2065, the Digital Services Act (DSA), and Regulation 2020/0374, the Digital Markets Act (DMA), have been introduced to “create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses” (European Commission, 2022a). A major aim is the regulation of large, powerful platform providers in order to create more fairness and transparency in the digital market. Accordingly, the two legal instruments have specific impacts on certain AI applications.

Aimed at regulating activities of digital services providers, i.e., providers of ‘intermediary service’, such as conduit, caching or hosting services, the DSA touches AI-based use cases in many ways, predominantly in the form of online advertisement targeting or recommender systems. With the DSA, providers of online platforms are required to establish additional measures, such as the presentation of meaningful information on the parameters of online advertisements, to increase transparency. Online interfaces must be designed, organized and operated in a way that does not deceive or manipulate users, i.e., ‘dark patterns’ are at least restricted. Targeting minors is subject to further restrictions to protect their privacy, safety and security, and advertisement based on profiling of minors is generally prohibited. In addition, assessment and mitigation of systemic risks are required from very large platform providers as well as the provision of an option for the user not to be subject to profiling. Likewise, the DMA imposes additional obligations for 'gatekeepers', which are defined as large providers of core platform services, often mainly referring to big tech corporations like Amazon, Apple, Google, Meta or Microsoft. Mainly focusing on obligations in relation to restrictions of limiting access to data and services, again, similar to the DSA, AI-based applications are mainly concerned with regard to recommender systems. Art. 6(1)(d) (DMA), for example, constrains gatekeepers not to give preference to own products when offering ranking services.

Finally, besides the digital domain, responsibilities for the AI provider within the realm of commercial practices may also be found within the Directive 2005/29/EC, the Unfair Commercial Practices Directive (UCPD), which generally prohibits misleading and aggressive commercial practices. With respect to AI systems, this shows effect in the prohibition of the provision of false information or deception to consumers about, above all, the essential characteristics of the products, such as their benefits and risks, and compliance with advertised codes of conduct. As the AI Act recommends the establishment of codes of conduct for minimal risk AI systems, albeit on a voluntary basis, such documents can thus be implicitly binding, where the unfaithful communication and promotion is prohibited according to the UCPD.

Fundamental rights

Defined in the Charter of Fundamental Rights of the European Union (2000/C 364/01) and linked to dignity, freedoms, equality, solidarity, citizen rights and justice, human or fundamental rights are the foundational basis for most of the above-outlined legal frameworks in the EU. In the construct of Union law, they apply between the EU institutions/bodies and the citizens. As regards the obligations of AI providers, despite being one of the most frequently discussed risk for AI and hence reason for AI governance initiatives (Kriebitz & Lütge, 2020), this framework also implies that no direct obligations for AI providers can be derived from the Charter itself. It rather results in a direct obligation for the state to protect its citizens from restrictions on fundamental rights and thus an indirect obligation for the AI provider in particular to comply with the stipulated provisions and measures against imposing restrictions on these fundamental rights.

2.2 Industrial Standards for AI Governance

While regulation is an important ground for AI governance, particularly, for clarifying the fundamental principles and goals, the concrete guidance on the adoption through technical processes and requirements is often left to Standard Developing Organizations (SDOs). Standards thereby refer to “technical document[s] designed to be used as a rule, guideline or definition” (CEN/CENELEC, 2024). In this function, they significantly shape the understanding and implementation of AI governance practices, concretizing the foundational work that has been set out by policy.

The ongoing efforts on standardization of AI and AI governance mechanisms constitute a prominent and contemporary subject of discussion. Consequently, the development of standards within this domain remains a work in progress. As of September 2023, the ISO/IEC JTC 1/SC 42 Committee on Artificial Intelligence has released 20 standards, 32 are currently under development (ISO, 2023). Primarily, these standards encompass foundational aspects of AI, addressing its associated risks and societal ramifications, while more specific and concrete concepts are yet to be fully formulated. Nevertheless, there exists a substantial body of preliminary work that has been made available to the public. Consequently, in the following, I will present an overview of the current endeavors in technical standardization and their implications for the governance mechanisms employed by AI providers.

The landscape of SDOs that contribute to the field of AI standardization is vast. These organizations can be categorized based on their operational scope and level of influence. International SDOs, such as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are responsible for the development of globally aligned standards. They typically convene through collaborative committees comprising representatives from all member states. These international standards receive reinforcement from regional counterparts, such as the European Committee for Standardization (CEN) or the European Committee for Electrotechnical Standardization (CENELEC) within the European Union, as well as the National Institute of Standards and Technology (NIST) in the United States. Additionally, national bodies like the Deutsches Institut für Normung (DIN) in Germany establish standards at the national level. Although this indicates a wide variety of approaches, for brevity I will, in this review, focus only on the international approaches.

2.2.1 Foundational Standardization Work

Significant foundational work has been undertaken by SDOs on conceptualizing AI systems as well as governance frameworks for those. ISO/IEC 22989:2022, published by the ISO/IEC Joint Technical Committee 1/Subcommittee 42 (ISO/IEC JTC 1/SC 42), the committee responsible for Artificial Intelligence, provides a comprehensive overview of technical specifications, definitions and terminology delineation. Particularly, a concrete definition of the concept of AI has been highly demanded to unify communication, which is outlined in this standard as “research and development of mechanisms and applications of AI systems” (ISO/IEC 22989:2022, p. 1), i.e. “engineered system[s] that generate outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives” (ISO/IEC 22989:2022, p. 1). Furthermore, ISO/IEC standards linked to AI are provided in the context of Big Data (ISO/IEC 20546:2019), Machine Learning (ISO/IEC 23053:2022), or computational approaches for AI systems (ISO/IEC TR 24372:2021). Standards pertaining to system architectures concerning Big Data can be found in the ISO/IEC TR 20547 series, or linked to AI model inference, storage, distribution and management in IEEE Std 2941-2021. Finally, considerable groundwork has been laid concerning the fields of application for AI by ISO/IEC TR 24030:2021. In this document, the ISO and IEC have cataloged 132 AI use cases, thereby delineating the scope of AI applicability and laying the foundation for future standardization endeavors.

Besides conceptual and technical foundations, ethical and social concerns have also been addressed in the work of SDOs. ISO/IEC TR 24368:2022 provides a comprehensive overview of ethical and societal concerns related to AI including initial considerations for its practical implementation. More specifically, ISO/IEC TR 24028:2020 offers an extensive examination of trustworthiness considerations in the context of AI. It surveys various approaches aimed at fostering trust among stakeholders in AI systems and delineates potential strategies for enhancing trustworthiness while mitigating vulnerabilities in these systems. Ethical issues, particularly linked to trustworthiness, are included in the fundamental considerations of ISO/IEC 22989:2022. Within this standard, a dedicated section focuses on the trustworthiness of AI applications, outlining essential characteristics that empower stakeholders to assess whether AI systems align with their expectations. Particularly, the need and fundamentals of robustness, reliability, resilience, controllability, explainability, predictability, transparency and bias & fairness are elaborated. Moving further towards practical governance of such concerns, ISO/IEC 38507:2022 particularly deals with the governance implications regarding the use of AI by organizations. Important provisions as well as implications from policy are laid down to guide organizations in the governance of ethical and societal risks that come with the use of AI.

The next steps in foundational standardization work now involve transitioning from the broad identification of ethical issues to the more specific elicitation of implications resulting from the identified principles. This entails outlining the concrete requirements of key principles such as robustness, explainability, bias and fairness. The groundwork is currently under development in this regard. For instance, ISO/IEC DIS 5259-1 is dedicated to the standardization of terminology pertaining to data quality in the context of artificial intelligence, while ISO/IEC CD TS 6254 endeavors to elucidate objectives and methodologies for achieving explainability. While significant progress has been made in standardizing general ethical imperatives, the initiatives are now moving on to lay the groundwork for more concrete questions in order to standardize the basis for operationalization.

2.2.2 Current Efforts in Standardization of AI Governance Operationalization

Based on the defined foundations, SDOs have proposed more concretized standards that offer unification on technical specifications for certain technologies, AI governance mechanisms, particularly in the form of risk management approaches, and more detailed provisions on the determined AI ethics principles.

More concrete technical specification is provided for some use cases. For example, IEEE 2945-2023 states clear technical requirements for face recognition technologies, such as regarding the sample quality, or thresholds for the face detection rate. IEEE 2841-2022 proposes an index system to evaluate the reliability of deep learning methods, and IEEE 3652.1-2020 more concretely describes measures for performance evaluation of federated machine learning. Such standards can thus be a valuable source for the determination of AI governance measures in the context of the certain described use case.

Particular reference to comprehensive AI governance mechanisms is made in the standardization of AI (ethical) risk management approaches. Building off ISO 31000:2018 on general organizational risk management, ISO/IEC 23894:2023 specifies the principles, framework and processes for risk management in the context of AI. IEEE 7000 proposes processes to include considerations of ethical and societal risks in the design of AI systems. Particularly 5 processes are outlined that take place during the concept exploration and development stages: (1) concept of operations and context exploration process, (2) ethical values elicitation and prioritization process, (3) ethical requirements definition process, (4) ethical risk-based design process and (5) transparency management process (IEEE 7000). For each of the processes, purpose, outcomes, activities, inputs and outputs are specified, qualifying the standard as a potentially ready-for-practice tool. However, it also acknowledges that, while with the application of this standard, general ethical considerations can be implemented into system design, the specific guidance on algorithm design for ethical principles is out of the scope.

Considerations on such provisions are elaborated in other documents. We see standardization for specific ethical principles, for example, for data governance issues, bias, robustness, quality and performance or general human well-being. ISO/IEC 8183:2023, for instance, delineates a comprehensive data lifecycle framework for AI systems, including actions for data acquisition, creation, development, deployment, maintenance and decommissioning. A comprehensive overview of bias assessment techniques along the AI system lifecycle phases is provided by ISO/IEC TR 24027:2021. Assessment of robustness, particularly regarding existing methods for neural networks, is described in the ISO/IEC TR 24029 series. In the realm of performance and quality measurement, several standards have already been established. ISO/IEC TS 4213:2022 outlines methodologies for assessing classification performance of machine learning models. Moreover, a model for specifying, measuring and evaluating the quality of AI systems is provided by ISO/IEC 25059:2023 and methodologies and benchmarks for AI server system performance specified in IEEE 2937-2022. Finally, with IEEE 7010-2020, an impact assessment for Autonomous and Intelligent Systems regarding human well-being is provided.

While these efforts demonstrate significant progress in the field, with standards transitioning from establishing fundamental principles to elucidating them at a more operational level, it also underscores the ongoing nature of this advancement, with some issues yet to be fully addressed. Specifically, various detailed technical standards and specifications are still under development, for example, for transparency and explainability (ISO/IEC CD TS 6254, ISO/IEC CD 12792), bias and fairness (ISO/IEC DTS 12791, IEEE P3198), functional safety and its requirements (ISO/IEC DTR 5469, ISO/IEC AWI TS 22440) as well as controllability and human oversight (ISO/IEC CD TS 8200, ISO/IEC AWI 42105).

3 | Implications, Problems and a Way Forward for the Operationalization of AI Governance Concepts

In the evolving landscape of AI governance, regulatory approaches play a crucial role in providing guidance to organizations on how to design, implement and use AI systems in a responsible way. Both strong and soft regulatory approaches can support such endeavors. More specifically, legal frameworks can serve as a basis and support the harmonization of diverse approaches to clearly set out a uniform approach. AI providers seeking to develop appropriate AI governance strategies gain valuable support in identifying desired practices and methods to prevent risks. In the EU, for example, the risk classes and imposed specific measures within the AI Act provide clarity on the treatment of AI systems easing the creation of suitable business operations. Recent updates to the General Product Safety Regulation and the Product Liability Directive, along with the introduction of the Artificial Intelligence Liability Directive in the EU, shed light on how liability practices intersect with AI systems. They elucidate where AI systems should be treated equally or differently to conventional, i.e. non-AI-based products.

However, at the same time, these developments require AI providers to adapt their AI activities to regulatory constraints. This seems highly beneficial where the goal is to prevent harmful practices, however, comes with additional tension. Legislators in particular are faced with a conflict, as they must regulate and steer, while taking individual conditions into account, leaving enough room for innovation and flexibility and not inappropriately raising entry barriers for new market participants. Resolving this conflict is surely not straightforward and typically results in a less prescriptive legislation – one that remains overarching in its principles. Given that AI encompasses a spectrum of technologies and not a singular entity, legislation is unlikely to be able to offer a proper one-size-fits-all solution but instead must articulate risks and provisions from a broader perspective. Detailing these obligations is then delegated to further entities – a co-regulatory approach that has been determined highly effective (Clarke, 2019). In other words, with some legislators adopting a strong regulatory approach and others opting for a more flexible stance, the latter approach, while providing greater scope and flexibility for innovation, also places the responsibility for establishing their own governance measures and implementing appropriate AI governance practices on the companies. While this preserves the potential for innovation, questions remain about the potential for appropriate risk mitigation.

To resolve the tension between basic – but broad – principles and their detailed – potentially too narrow – implementation propositions, interest and efforts in creating unification in the field of governing AI have grown rapidly. Existing standards have already – and initiatives currently under development will soon – further unified fundamental aspects and offer steps to concretize defined responsible AI concepts. Taken together, these standards (will) provide a framework upon which AI practitioners can build, helping them navigate the often complex governance of AI development. Standardization, in particular, is usually seen as crucial to creating a level playing field that promotes consistency and optimization of development processes, which facilitates compliance requirements and strengthens consumer trust. Finally, adherence to standards, both within and outside the AI context, can enhance collaboration within the industry, creating opportunities for partnerships and innovation.

Given the importance of technical standardization in the industry, it is not surprising that AI practitioners often express a desire for standards as they can bridge the gap between theory and practice. Tartaro (2023), for example, has examined the landscape of international and European standardization in accordance with the AI Act to identify its potential, but also the challenges and limitations in adapting to regulatory requirements. Particularly the contextual adaptations to an organization’s specific environment and circumstances remain unresolved. As company goals, resources and unique challenges can vary significantly, standards may not cover the particular nuances of specific AI projects. Organizations can regard standards as a foundational basis, utilizing them as a starting point to tailor and contextualize according to their specific needs. This process nevertheless entails a critical assessment of how standards align with the organization's goals, values and constraints.

In summary, while we see that regulation and technical standardization play a central role in the implementation of responsible AI, best practices in practical contexts are currently rare due to the constrained possibility to customize provisioned frameworks to concrete practical use case needs. More specifically, practitioners often attribute these difficulties to a range of AI-specific characteristics and implications. Uncertainty regarding how to properly address risks caused by AI systems have been frequently highlighted (Hohma et al., 2023). While a variety of approaches, strategies and frameworks exists, the lack of unified standard procedures and the missing consensus on comprehensively agreed-upon mechanisms inhibits practitioners from effectively addressing risks linked to AI. Thereby, the problem of striking a balance between generalizing concepts to increase wide adaptability and specifying actions to ensure easy and fast adoption certainly intensifies the struggle to determine standard measures. Furthermore, resources are scarce and additional burdens to develop and implement suitable practices are often seen to outweigh the benefits (Hohma et al., 2023). A second major drawback in this regard is the often-mentioned black box nature of many AI systems that further inhibits the implementation of appropriate risk management mechanisms (Hohma et al., 2023). It particularly causes an issue for reusing concepts from non-AI-based contexts, as understanding is required for risk identification and therefore a prerequisite for addressing them.

In this thesis, I therefore aim to address issues for the implementation of regulation and standardization to ultimately bring responsible AI to practice, primarily through targeting two fundamental properties: clarity and concretization of existing approaches. Clarifying and contextualizing current concepts in responsible AI is crucial to provide a clear understanding of ethical guidelines and foster consistency across diverse approaches. For this purpose, the research in this thesis aims to detail the implications and related obligations in coping with these risks for AI providers building upon previous work. A major goal is thus to concretize possible measures ensuring a tangible and applicable approach to AI governance. Particularly the often abstract principles shall be translated into more concrete actions by studying them in detail and in a more contextualized format, with the underlying aspiration of supporting a clearer roadmap for AI providers to navigate the complex landscape of ethical AI development.

4 | Contribution to Advancing AI Governance to Business Practice

The dissertation addresses the outlined fundamental goal of transferring the developed foundations in corporate AI governance from theory to practice. It is based on three articles published in peer-reviewed international journals and conference proceedings that contribute to the clarification and contextualization of current AI governance approaches.

The first article serves as a comprehensive exploration of the implications of regulation and standardization on the development obligations of AI providers. In particular, its primary objective is to facilitate the translation of conceptual considerations surrounding AI governance into actionable business practices by deriving specific obligations and measures, hence offering the formulation of a trustworthy development process for AI systems. Acting as a foundational basis, this article lays the groundwork for the more detailed exploration of two of the demanded ethical principles and related measures in Articles 2 and 3. Article 2 is a deep dive into the principle of fairness, providing insights into the conceptualization of appropriate definitions for fairness within machine learning algorithms, particularly exploring recommendations for the appropriate choice of the underlying fairness model. Article 3 showcases the embedding of ethical principles into technical implementation through the development of a robust and accelerated version of the spectral clustering algorithm.

The full papers can be found in the appendix. In the following, the main purpose, results and outcomes are summarized as well as how they contribute to the overall thesis goals.

4.1 Extended Abstract Article 1 – From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems

Considerations on AI governance and responsible AI have found their way into corporate headquarters and strategy departments. Fundamental principles to guide this transition, such as the considerations on trustworthy AI by the European Commission's High-Level Expert Group on Artificial Intelligence, have been established and are widely acknowledged. To implement those, industry stakeholders have started to form initiatives or consortia to jointly discuss and prepare for the envisaged sound, robust development and targeted prevention of negative side effects of AI systems. Fundamental directions thus seem to have been determined and guiding principles to reach those agreed.

The practical implementation of such fundamentals, however, is currently seen as one of the greatest uncertainties and therefore drawbacks of responsible AI (Dafoe, 2018; Stix, 2021). There are many approaches to bringing the agreed principles to practice that focus on detailing high-level concepts with more concrete explanations regarding their implications for practical use (e.g., Ayling & Chapman, 2021; Georgieva et al., 2022; Hagendorff, 2020; Larsson, 2020; Li et al. 2021; Mittelstadt, 2019; Morley et al., 2021; Ryan & Stahl, 2020). Such considerations often center around the system properties that are required to align AI applications with agreed responsible AI principles. While defining such requirements is inevitable, particularly when implemented in practice, often, the system properties cannot be ensured directly but must rely on measures taken along the system development process to prove that all necessary actions have been considered to make the system as trustworthy as possible. To comprehensively and sustainably bring responsible AI mechanisms to practice, we, therefore, suggest a rethinking of the primary focus of trustworthy AI, broadening the focus from the resulting system to the perceived trustworthiness of the associated development process.

The ultimate goal of our research is to support the transition of responsible AI concepts to actionable mechanisms. We do this with a process-based approach by studying the requirements and elements of a trustworthy development process for AI systems. We analyze existing AI governance initiatives to retrieve practices to operationalize defined

AI ethics principles and map them to a trustworthy development process. Further, we build off existing research on procedural trustworthiness, e.g., from regular software development to identify key properties that characterize trustworthy processes. Finally, we study the suggested elements for a trustworthy development process built off previously proposed measures to determine its implications for trustworthiness perceptions.

Measures to develop a trustworthy development process for AI systems were retrieved through a semi-systematic literature analysis of AI governance efforts and EU-centered regulatory frameworks. These AI governance efforts were searched from non-regulative policy efforts, standards developing organizations (SDOs), academic and research institutes or consortiums and non-governmental organizations (NGOs) or civil society groups. We investigated 155 AI ethics stakeholders of which documents on the responsible development of AI were found from 45. Finally, 14 of these were found to have published practice-oriented reports on AI-related obligations, measures, or responsibilities that were publicly accessible. These documents were analyzed and coded in order to translate agreed AI ethics requirements to practical obligations and subsequently derive measures to fulfill them. The resulting measures were mapped to the AI development lifecycle to develop a trustworthy development process for AI systems.

Our results provided valuable contributions for practical implementation – regarding the analysis of measures and establishment of a trustworthy development process concept – as well as theoretical understanding – regarding the fulfillment of identified process trustworthiness requirements and thus perceived trustworthiness of currently proposed practices. The comparison with implications from legally binding and non-binding governance frameworks shows that some of the identified measures are already legally binding under certain circumstances, depending on the specific AI application or engaged target group. Particularly, severe risks are often addressed through regulatory approaches with the intention of limiting harm while offering room for technological advancement and innovation. While clarity is often cited as a major obstacle to the realization of these principles, breaking down obligations into specific measures can provide more concrete instructions for AI providers. However, some measures require further efforts to apply them to real-world scenarios, and therefore open questions for the research and practice of moving responsible AI forward have been pointed out. While recent research often focuses on the technical implementation of responsible AI, the determined AI governance mechanisms are largely recommended among non-technical methods, such as strategy-making, documentation and communication. This suggests that careful consideration needs to be given to when automated approaches are feasible and desirable or when non-technical processes, perhaps based on human intuition, are nevertheless required. A similar conclusion can be drawn for the analysis of the potential to facilitate procedural trustworthiness of the identified measures. While many factors of procedural trustworthiness were found to be fulfilled, limitations were primarily found due to the vagueness of proposed measures. Therefore, not only to enable implementation of trustworthiness but in addition to foster trustworthiness perceptions, a detailing of measures based on use cases and the system's context is required.

4.2 Extended Abstract Article 2 – Individuality and Fairness in Public Health Surveillance Technology: A Survey of User Perceptions in Contact Tracing Apps

AI is becoming increasingly prevalent, a trend that is particularly notable in the field of healthcare. Examples include, but go even beyond the use of AI-powered diagnostic tools, predictive analytics for patient care and personalized treatment recommendations based on the analysis of large datasets. The Covid-19 pandemic has further fueled the use of algorithmic capabilities to track and predict beyond human capabilities. Yet, such public health surveillance technologies, designed to process vast amounts of data, have faced criticism due to their potential to incorporate bias and discrimination, as they naturally are built around personal socio-economic information like race, gender, or previous health-related details. The proliferation of these technologies during the pandemic, combined with their ability to comprehensively capture sensitive data, has reinforced the need to develop public health surveillance applications in a fair manner. Underlying this goal of a fair development, however, is the need to determine an appropriate statistical model for the introduction of fairness and, consequently, identifying what notion of fairness represents an appropriate and thus desirable state. The goal of our research was therefore to gain insights on the perspectives that individuals hold regarding the definition of fairness. More specifically, we sought to understand what they require for a public health surveillance tool to be considered fair, with the concrete example of contact tracing applications. Besides its contribution to contextualizing ethical fundaments, our article hence also reinforces the initial assumption of required AI-specific investigations, as it nicely depicts their challenge – it is not, as with most traditional tools, a question of individual rule-based decisions, but rather a question of how to define predetermined and embedded principles, which must be set more cautiously due to, for example as in this case, their higher impact on the potential for discrimination.

To empirically assess different fairness definitions, a vignette study with 273 participants, 129 from the UK and 144 from Germany was conducted. The objective was to assess preferences in defining fairness within a treatment. For this, our experimental design focused on two extremes: highly individualized treatments based on the collection of personal data, such as the number of contacts people meet on a regular basis, and non-personalized treatments, where no further individual data was used for decision-making. In our contact tracing example, these approaches were translated as follows. Under the high-individuality approach, the algorithm's decision to recommend isolation to an individual depends upon a person's specific characteristics, such as regular interactions with a large number of people, indicating a higher risk of virus transmission. In contrast, the low-individuality approach would not depend on a person’s individual characteristics in the decision-making regarding isolation recommendations. Our study examined participants' perceptions of fairness as well as overall quality of each approach, alongside an exploration of participants’ privacy concerns about the two versions.

Our analysis revealed that although participants noted higher privacy concerns with the high-individuality approach, they rated it as slightly but significantly better overall in our survey and considered it fairer compared to the low-individuality approach. Further, a strong correlation between the participants’ perceived fairness of an approach and their overall impression of the tracking tool was determined. In other words, our results indicated that under certain circumstances, users value higher degrees of individuality in health surveillance-related decisions even though this might require accepting a certain degree of data release.

From a theoretical perspective, our study contributes to further contextualizing the personalization-privacy trade-off. Since participants in this study felt discriminated when judged solely based on demographics, but also unfairly treated when considered as an anonymous, homogeneous group, our results highlight the need to distinguish between parameters that are considered discriminatory or necessary for fair decision-making. Immutable traits like gender or race tend to be seen as discriminatory, whereas factors related to an individual's actions, such as their virus-spreading risk, are more accepted for disclosure. The study suggests that selecting the right attributes can increase people's willingness to accept limitations on personal freedoms while still perceiving the treatment as fair and highlights the ongoing challenge of balancing personal rights with societal well-being in healthcare and AI contexts.

The level of individuality is pivotal in determining the appropriate fairness model for AI-enabled technologies. However, making this decision remains complex and context-dependent. From a practical perspective, our study hence aims to assist developers by examining the impact of individuality on fairness and overall user perception. The results show that participants value some degree of individuality in decisions, even at the expense of stringent data privacy, emphasizing that data privacy is not always the top concern. Balancing these ethical principles is crucial, as enhanced fairness perceptions can improve user attitudes toward applications. The study underscores the context-dependent nature of fairness model selection. In public health surveillance, a preference for some individuality over complete homogeneity is observed. This suggests the importance of models like "Fairness through Awareness," which consider individual characteristics.

In summary, our findings reinforce the growing preference for personalization in healthcare, extending to health surveillance technologies. We suggest developers consider this trend when designing upcoming AI-powered public health surveillance tools. While our study suggests that greater individualization seems appealing to participants, it highlights the importance of the specific attributes used in decision-making, emphasizing the necessity for further research to differentiate between parameters viewed as fair vs. discriminatory.

4.3 Extended Abstract Article 3 – SCAR: Spectral Clustering Accelerated and Robustified

Clustering tasks are an important category of data mining and machine learning problems that focus on identifying groups in a given dataset such that data points are highly similar within the subgroups and rather dissimilar to points outside their subgroup. Clustering techniques are thus essential across various domains dealing with data and often serve as an unsupervised preprocessing step for numerous subsequent data analysis tasks. Among the many available clustering methods, spectral clustering stands out for its versatility, as it is applicable to non-numeric datasets and can identify clusters of complex, even non-convex shapes and varying densities.

However, while its theoretical potential is undenied, the practical application of spectral clustering algorithms is lacking, mainly due to two primary drawbacks: standard spectral clustering is slow and highly sensitive to noisy input data. On top, addressing one of these issues often exacerbates the other, as introducing additional data cleansing steps adds to the runtime, and techniques for speeding up the clustering tasks often entail losses of result quality. Particularly in real-world scenarios, standard spectral clustering approaches, therefore, pose challenges, as with newly developed methods for data collection, datasets have grown in both dimensionality and size in recent years (see e.g. in medicine, chemistry, or biology).

To address both primary issues of standard spectral clustering techniques at the same time, we propose SCAR, an accelerated and robustified spectral clustering method. To identify groups within a dataset, standard spectral clustering algorithms essentially follow 3 steps: (1) a similarity graph is constructed, (2) the Laplace matrix is computed from the similarity graph's matrix representation as well as respective eigenvectors calculated, and (3) the obtained eigenvectors are clustered using a basic clustering algorithm, e.g., k-means (Luxburg, 2007). While SCAR in principle follows the same steps, we achieve robustification by iteratively separating the constructed similarity graph into two latent components, the cleansed and the noisy data, and acceleration by speeding up the eigendecomposition – the most time-consuming step – through approximation with the Nyström method.

Our experiments show that SCAR can significantly reduce sensitivity to noisy input and runtime compared to standard spectral clustering. Using well-known, real-world benchmark datasets, we compare SCAR's clustering performance to state-of-the-art methods. Robustness was evaluated regarding noisy edges in the data's similarity graph representation as well as with respect to jitter in the original data – the two most challenging types of noise for clustering. SCAR consistently yielded low runtimes, while maintaining highly competitive clustering qualities on real and synthetic data.

Demonstrating its potential for delivering fast and high-quality results, SCAR's ability to outperform comparable algorithms in both speed and accuracy makes it particularly valuable in application scenarios such as image recognition and video segmentation, where a rapid analysis of large datasets with high precision is crucial – in particular combined with biometric data a highly interesting case also in light of EU AI Act regulation. Its potential for reducing vulnerability to noise of standard spectral clustering approaches in acceptable time frames thus supports improving the state-of-the-art of clustering approaches.

This article was based on work that has been initiated in the Bachelor Thesis "Accelerating Robust Spectral Clustering Using the Nyström Method" (Hohma, 2021).

5 | Discussion

The outlined three research articles contribute to the practical implementation of AI governance, in particular through improved clarification and contextualization of existing AI governance practices. The objective that was set for this thesis to operationalize the fundamental principles of AI governance, moving it from mere theoretical consideration towards enhanced and actionable industry practice, has hence been pushed in a variety of ways.

To lay the ground for the research endeavor, in the first paper, an analysis revealed an emerging common core of measures crucial for achieving trustworthy AI development. While the identified commonalities in existing responsible AI operationalization provide a foundational framework, the research also affirmed the multiplicity of existing approaches in the field each supporting its practical transition in their own way. The multitude of existing initiatives and players underscores the urgent need for standardization and summarization into cohesive and widely accepted approaches. The trustworthy process for AI development proposed in Paper 1 serves as a crucial starting point towards this unification. By delineating a systematic and comprehensive approach to ensure trustworthiness in the development process, it can offer a roadmap for practitioners and policymakers. The synthesis thus helps clarify existing initiatives by bringing together different approaches into a coherent basis to promote consistency.

Besides clarification, a second endeavor of this thesis is the advancement of contextualization of AI governance research to support its transition to practice. The research in Articles 2 and 3 support this objective particularly for two example cases: public health surveillance and spectral clustering techniques. As emphasized in Paper 2, the challenge of contextualization requires a nuanced understanding of the specific contexts in which AI governance operates. The contribution of this article is hence twofold. From a methodological perspective, it provides a proof of concept for theoretical foundations examined within a certain application scenario which can be adapted and extended across diverse contexts. From a more practical perspective it examines fairness considerations tailored to the specific demands of public health surveillance, providing useful insights on the unique interpretations of fairness definitions in this context. Similarly, the acceleration and implementation of robustness into spectral clustering techniques, explored in Paper 3, serves as a showcase for a technical realization of responsible AI principles for one specific context. From a more high-level perspective, this emphasizes the need to address risks related to the responsible development of AI with both organizational and technological approaches.

In summary, the implications of this work for the state of the art in AI governance are manifold. The thesis highlights and addresses the need for a unified and comprehensive approach towards responsible AI and, with the provision of a trustworthy development process that summarizes existing policies and principles as well as the exemplary examination of 2 of the identified mechanisms in specified contexts, it can support the practical conceptualization of the demanded foundations. By providing both conceptual and methodological groundwork, the research navigates the complexity of AI governance and makes it accessible and applicable in different contexts. The proposed contributions hence go beyond the theoretical framework and offer practical solutions and insights that directly address the identified challenges. Overall, this research advances the understanding and operationalization of AI governance by providing clarity, coherence and adaptability to contextual applications, thereby enriching the evolving transition of responsible AI from theory to practice.

Nevertheless, fulfilling the thesis objective to operationalize responsible AI towards a set of practical, actionable concepts comes with distinct limitations that primarily stem from the inherently emerging as well as unsteady nature of the field of AI governance. The rapid, dynamic evolution of new concepts and technologies naturally limits the exhaustive and comprehensive analysis of proposed approaches. Striving for comprehensiveness with this research, it is crucial to acknowledge this limitation as the outcomes of this thesis hence represent a snapshot of the current state of the field and may only capture emerging trends and concepts to a limited extent.

A second limitation lies in the scope of this research, particularly regarding the objective of contextualization. Contextualization inherently poses challenges, since, as the very nature of the term implies, it requires the consideration of multiple, diverse contexts.

Achieving comprehensiveness in contextualization is thus hardly achievable. However, as outlined above, the dual contribution of this work can further support the adaptation of developed concepts to various contexts, without the need to explicitly examine them in this regard. While the research findings provide insights into specific contexts, at the same time they can be read as a methodological proof of concept to be reused under further conditions. Thus, although this research may not comprehensively contextualize the existing multitude of AI principles and concepts, it establishes a foundation for reusing the developed methodologies across multiple scenarios.

In essence, the limitations once more highlight the challenges that current AI governance research must navigate, as, while striving for comprehensiveness and specificity, such conceptualizations must operate within the constraints of an evolving AI technology and AI governance landscape. This necessitates recognition and strategic adaptation, as progress towards viable approaches can only be achieved through continuous research and the advancement of new or existing concepts. Despite the current uncertainties that pose challenges to various actors and activities in the field, it remains imperative to persist in exploring diverse paths and consolidating insights, with the ultimate goal of attaining the desired outcomes.

The considerations on contributions and limitations of this thesis once more emphasize the challenges of the ever-evolving landscape of AI governance and particularly regarding its operationalization, thus opening up expansive avenues for future work. Given the dynamic landscape and the interrelatedness of scientific exploration and practical adoption in operationalizing AI governance, the potential for further research is substantial on both ends. As indicated in the context of limitations of this thesis, one straightforward trajectory for future theoretical exploration involves the clarification and contextualization of the proposed concepts beyond the studied domains of public health surveillance – as an example for a thematic context – and spectral clustering techniques – an example for technological context. Further, this would include an exploration of the two foundational principles examined in this research, fairness and robustness, under varied conditions, and in addition suggests analyzing proposed measures for the remaining responsible AI foundations, such as human oversight or transparency. A second crucial aspect of future research involves navigating the delicate balance between specialization and generalization. Inferring generally viable concepts along with blueprints for actionable mechanisms that go beyond the specific consideration of certain contexts is paramount for advancing the unified operationalization of AI governance across diverse applications.

From a practical perspective, developing established best practices that summarize and translate valuable experiences in operationalizing AI governance is imperative. On the one hand, this entails developing the needed actionable guidelines and operationalization recommendations. This thesis has pushed advancement in this direction through demonstrating that important steps have been taken prior to as well as within this work, and – even if eventually far – through continuous research and adaptation the destination is in sight. On the other hand, a research path that has not been entered in this thesis, yet, is no less important for the practical application of AI governance links to ensuring that such detailed and contextualized concepts are communicated to the right stakeholders and in the right format. Tailoring materials to the specific understanding of different stakeholders, including policy makers, industry practitioners and AI developers, is essential for effective adoption. Operationalizing AI governance in practice therefore requires a contextualized understanding of requirements not only of high-level field actors but must understand the very specific needs of stakeholders that are involved in the implementation of such measures. To achieve this and ensure the alignment of the provided material to the needs of stakeholders, the implementation of AI governance principles requires a collaborative and interdisciplinary research effort.

As final concluding words summarizing the contributions, limitations and proposed future work of this thesis from a more overarching perspective, this thesis has shown that a clarification and contextualization of AI Governance fundamentals is not only needed but also doable – in fact, the fundamentals developed thus far are close to unified practical alignment that an actionable development process could be retrieved at least from the current snapshot of activities within thesis. Their required contextualization although being quite advanced, however, seems to struggle more with the fast-paced evolution within the AI technology and governance field. Solving the challenge of generalization vs. specification will bring clarity on how far such a contextualization should and can go. In any case, this research has offered ways how to tangibly address responsible AI considerations under both thematic as well as technological contexts and has shown that and how this supports their actionable operationalization.

6 | Conclusion

The aim and contribution of this work was to effectively address various challenges related to the operationalization of AI governance which led to the inference of required subsequent investigations and suggestions for future research to further advance solutions in this field. These contributions were built off existing, agreed efforts aimed at managing risks associated with the responsible development of AI, including AI governance concepts grounded in regulation and standardization approaches. While this suggests that numerous fundamental problems in this field can find resolution through careful examination, cooperation and agreement, it is imperative to recognize the existence of contentious questions sparking heated debates as settling on agreed solutions for them does not seem straightforward. It is with an outlook on such challenges that I wish to conclude, to emphasize the need for ongoing research in navigating the multifaceted landscape of AI governance.

One such issue revolves around the fundamental approach on how to address (ethical) risks that arise with and around AI. In this thesis, I have described the simultaneous evolution of soft and strong regulatory paths. While we see governments around the world entering roads on pursuing each of them, there is an ongoing uncertainty on which to follow – or whether to follow one at all. In particular, it is heavily discussed as to whether a restrictive approach based on binding regulations is desirable. Regulation is seen as providing a structured framework for identifying and mitigating potential harms and risks inherent in AI systems which is essential for safeguarding individuals and society from adverse consequences, promoting safety, and ensuring accountability. By establishing clear lines of responsibility, regulations hold developers, manufacturers, and operators accountable for the performance and safety of AI systems, contributing to a more secure and trustworthy AI landscape that builds public trust by assuring users that AI technologies adhere to established safety standards, fostering wider adoption and engagement. In contrast, critics of a regulative approach emphasize potential drawbacks that may impede progress in the AI domain. It is argued that stringent regulations may stifle innovation by imposing rigid constraints on AI development. There is concern that an overly regulated environment could discourage experimentation and hinder the evolution of AI technologies. Additionally, the dynamic nature of AI development poses a challenge for regulatory frameworks, as rapid advancements may outpace the ability of regulations to adapt – an obstacle that has been seen with the development of the AI Act facing the rapid introduction of generative AI technologies and therefore the challenge of whether and how to react.

As a final concluding reflection – and to revisit considerations from the introduction – the existence of such fundamental controversies underlines that AI currently provokes a wide range of ambiguities. These raise the question of whether an urgent call for action is justified or whether AI can be treated as some sort of “next buzz technology" for which the currently established frameworks can be largely reused. While we may have embarked on the journey of AI evolution, it is crucial to recognize that we are still at the outset and therefore find ourselves at a crossroad, where decisions must be made about how to navigate this terrain. We are now in the exciting position to make choices on whether we want to view AI as an opportunity to rectify potentially identified past mistakes in dealing with powerful technologies or take the chance to harness its capabilities for high-performing innovation and growth – both paths sound tempting. Regardless, one thing is clear: we cannot simply ignore the advancement of AI technologies, for it is already underway, with all its excitement and challenges.

References

Access Partnership (2023). *Access Alert | Brazil's New AI Bill: A Comprehensive Framework for Ethical and Responsible Use of AI Systems.* Access Partnership. https://accesspartnership.com/access-alert-brazils-new-ai-bill-a-comprehensive-framework-for-ethical-and-responsible-use-of-ai-systems (last accessed: 11th Feb 2024).

Almada, M., & Petit, N. (2022). The EU AI Act: Between Product Safety and Fundamental Rights. *Available at SSRN*. https://doi.org/10.2139/ssrn.4308072

Alreemy, Z., Chang, V., Walters, R., & Wills, G. (2016). Critical success factors (CSFs) for information technology governance (ITG). *International Journal of Information Management*, *36*(6), 907-916.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, 405–429. https://doi.org/10.1007/s43681-021-00084-x

Bareis, J., & Katzenbach, C. (2022). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values*, *47*(5), 855-881.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An Introduction to Ethics in Robotics and AI.* Springer Nature. https://library.oapen.org/handle/20.500.12657/41303

Borges, G. (2021). AI systems and product liability. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL '21)*, 32–39. https://doi.org/10.1145/3462757.3466099

CEN/CENELEC (2024). *European Standards*. CEN/CENELEC. https://www.cencenelec.eu/european-standardization/european-standards/ (last accessed: 11th Feb 2024).

Clarke, R. (2019). Regulatory alternatives for AI. *Computer Law & Security Review*, *35*(4), 398-409.

Dafoe, A. (2018). AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, *1442*, 1443.

Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32005L0029

EIT Community. (2021). Creation of a Taxonomy for the European AI Ecosystem: A report of the Cross-KIC Activity "Innovation Impact Artificial Intelligence".

European Commission. (2022a). *The Digital Services Act package*. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

European Commission. (2022b). *A European approach to artificial intelligence*. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

European Commission. (2022c). *New liability rules on products and AI to protect consumers and foster innovation*. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807

European Commission. (2022d). *Product Liability Directive - Adapting liability rules to the digital age, circular economy and global value chains*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Product-Liability-Directive-Adapting-liability-rules-to-the-digital-age-circular-economy-and-global-value-chains_en

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., & Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (2022). From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*, 1-15. https://doi.org/10.1007/s43681-021-00127-3

Government of Canada (2023). *Artificial Intelligence and Data Act*. Government of Canada. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act#:~:text=Canada%20is%20one%20of%the,small%20and%20medium%2Dsized%20businesses. (last accessed: 11th Feb 2024).

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). *Ethics Guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hohma, E. (2021). *Accelerating Robust Spectral Clustering Using the Nyström Method.* [Bachelor Thesis, Ludwig-Maximilians-Universität München].

Hohma, E., Boch, A., Trauth, R., & Lütge, C. (2023). Investigating accountability for Artificial Intelligence through risk governance: A workshop-based exploratory study. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1073686

Holistic AI (2023). *The State of Global AI Regulations in 2023.* Holistic AI.

ISO/IEC. (2015). Information technology – Vocabulary. In.

International Standardization Organization (ISO) (2023). *ISO/IEC JTC 1/SC 42 Artificial intelligence*. ISO. https://www.iso.org/committee/6794475.html (last accessed: 11th Feb 2024).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kohn, B. & Pieper, F. (2023). *AI regulation around the world.* Taylor Wessing. https://www.taylorwessing.com/en/interface/2023/ai---are-we-getting-the-balance-between-regulation-and-innovation-right/ai-regulation-around-the-world (last accessed: 11th Feb 2024).

Kriebitz, A., & Lütge, C. (2020). Artificial intelligence and human rights: a business ethical assessment. *Business and Human Rights Journal*, *5*(1), 84-104. https://doi.org/10.1017/bhj.2019.28

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, *7*(3), 437-451. https://doi.org/10.1017/als.2020.19

Latham & Watkins (2023). China's New AI Regulations. Latham & Watkins. https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf (last accessed: 11th Feb 2024).

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2021). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, *55*(9), 1–46. https://doi.org/10.1145/3555803

Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17, 395-416.

Lynch, S. (2023). 2023 State of AI in 14 Charts. *Human-Centered Artificial Intelligence, Stanford University.* *https://hai.stanford.edu/news/2023-state-ai-14-charts* (last accessed: 11th Feb 2024).

McCarthy, J. (2007). What is artificial intelligence? *Stanford University*.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mohamed, N., & Kaur a/p Gian Singh, J. (2012). A conceptual framework for information technology governance effectiveness in private organizations. *Information Management & Computer Security*, *20*(2), 88-106.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 153-183). Springer. https://doi.org/10.1007/s11948-019-00165-5

Nussey, S. & Kelly, T. (2023). *Japan leaning toward softer AI rules than EU, official close to deliberations says*. Reuters. https://www.reuters.com/technology/japan-leaning-toward-softer-ai-rules-than-eu-source-2023-07-03/ (last accessed: 11th Feb 2024).

Organisation for Economic Co-operation and Development (OECD) (2019). *Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449. https://oecd.ai/en/ai-principles

Organisation for Economic Co-operation and Development (OECD) (2024). *Technology Governance.* https://www.oecd.org/sti/science-technology-innovation-outlook/technology-governance/ (last accessed: 11th Feb 2024).

Peterson, R. (2004). Crafting information technology governance. *Information systems management, 21(4)*, 7-22.

Radu, R. (2021). Steering the governance of artificial intelligence: national strategies in perspective. *Policy and society*, *40*(2), 178-193.

Regulation 2016/679. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Regulation 2017/0003. Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on

Privacy and Electronic Communications). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32002L0058

Regulation 2020/0340. Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52020PC0767

Regulation 2020/0374. Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A842%3AFIN

Regulation 2021/0106. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artifical Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Regulation 2022/0047. Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=COM:2020:0825:FIN

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61-86. https://doi.org/10.1108/JICES-12-2019-0138

Samoili, S., Cobo, M. L., Gomez, E., De Prato, G., Martinez-Plumed, F., & Delipetrev, B. (2020). AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence. Technical Report. Joint Research Centre (Seville site). https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

Smits, D., & Hillegersberg, J. v. (2013). The continuing mismatch between IT governance theory and practice: Results from a Delphi study with CIO's.

Stix, C. (2021). Actionable principles for artificial intelligence policy: three pathways. *Science and Engineering Ethics*, *27*(1), 15. https://doi.org/10.1007/s11948-020-00277-3

Tartaro, A. (2023). Towards European Standards Supporting the AI Act: Alignment Challenges on the Path to Trustworthy AI. Proceedings of the AISB Convention,

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137 (last accessed: 11th Feb 2024).

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97-112. https://doi.org/10.9785/cri-2021-220402

Voss, A. (2023). *„Der Brüssel Effekt"*. C.H.Beck Verlag. https://rsw.beck.de/zeitschriften/rdi/single/2023/08/29/der-br%C3%BCssel-effekt (last accessed: 25th Feb 2024).

White House Office of Science and Technology Policy (OSTP). (2022). *Blueprint for an AI Bill of Rights – Making Automated Systems work for the American People*. Washington: White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last accessed: 11th Feb 2024).

Wilson, C. (2022). Public engagement and AI: A values analysis of national strategies. *Government Information Quarterly*, *39*(1), 101652.

A.1 Supporting Material to Article 1

In the following, supporting material for Article 1 is provided. It contains (1) copyright information, (2) the statement of contribution and (3) the article as printed in the respective journal.

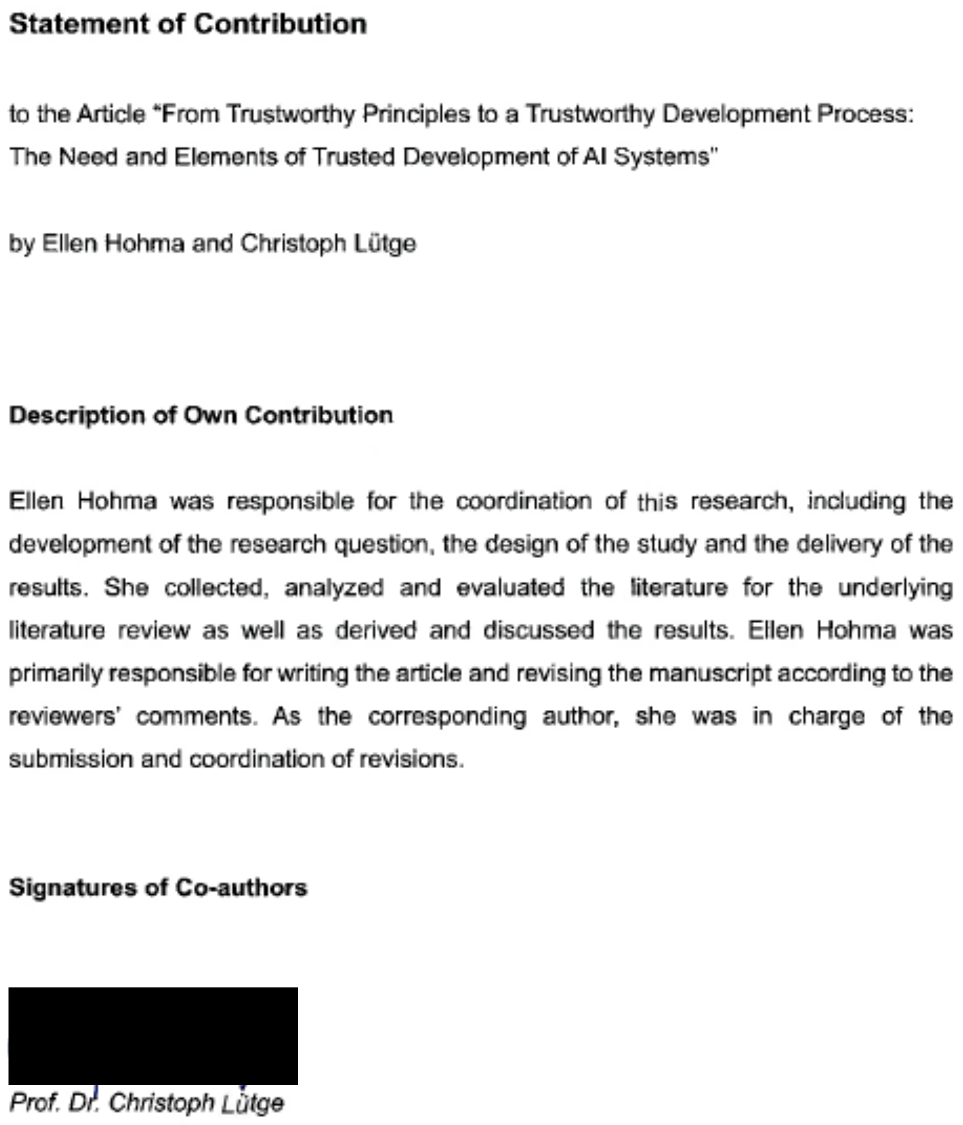
A.1.1 Copyright Information

This paper was published as

Hohma, E. & Lütge, C. (2023). From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI 4(4)*, 904-925. https://doi.org/10.3390/ai4040046

under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

A.1.2 Statement of Contribution



Statement of Contribution

to the Article "From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems"

by Ellen Hohma and Christoph Lütge

Description of Own Contribution

Ellen Hohma was responsible for the coordination of this research, including the development of the research question, the design of the study and the delivery of the results. She collected, analyzed and evaluated the literature for the underlying literature review as well as derived and discussed the results. Ellen Hohma was primarily responsible for writing the article and revising the manuscript according to the reviewers' comments. As the corresponding author, she was in charge of the submission and coordination of revisions.

Signatures of Co-authors

[redacted]

Prof. Dr. Christoph Lütge

A.1.3 Full Article

Article

From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems

Ellen Hohma * and Christoph Lütge

Insitute for Ethics in Artificial Intelligence, School of Social Sciences & Technology, Technical University of Munich, 80333 Munich, Germany; luetge@tum.de
* Correspondence: ellen.hohma@tum.de

Abstract: The current endeavor of moving AI ethics from theory to practice can frequently be observed in academia and industry and indicates a major achievement in the theoretical understanding of responsible AI. Its practical application, however, currently poses challenges, as mechanisms for translating the proposed principles into easily feasible actions are often considered unclear and not ready for practice. In particular, a lack of uniform, standardized approaches that are aligned with regulatory provisions is often highlighted by practitioners as a major drawback to the practical realization of AI governance. To address these challenges, we propose a stronger shift in focus from solely the trustworthiness of AI products to the perceived trustworthiness of the development process by introducing a concept for a trustworthy development process for AI systems. We derive this process from a semi-systematic literature analysis of common AI governance documents to identify the most prominent measures for operationalizing responsible AI and compare them to implications for AI providers from EU-centered regulatory frameworks. Assessing the resulting process along derived characteristics of trustworthy processes shows that, while clarity is often mentioned as a major drawback, and many AI providers tend to wait for finalized regulations before reacting, the summarized landscape of proposed AI governance mechanisms can already cover many of the binding and non-binding demands circulating similar activities to address fundamental risks. Furthermore, while many factors of procedural trustworthiness are already fulfilled, limitations are seen particularly due to the vagueness of currently proposed measures, calling for a detailing of measures based on use cases and the system's context.

Keywords: artificial intelligence governance framework; ethical duties; legal duties; AI ethics principle operationalization; responsible AI; semi-systematic review

check for updates

Citation: Hohma, E.; Lütge, C. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI* **2023**, *4*, 904–925. https://doi.org/10.3390/ai4040046

Academic Editors: Pablo Rivas and Gissella Bejarano

Received: 29 August 2023
Revised: 29 September 2023
Accepted: 10 October 2023
Published: 13 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Numerous international governmental or non-governmental stakeholders have proposed fundamental principles for responsible AI that are supported by many organizations using and providing AI applications. A consensus on the fundamental values that shall build the foundation for responsible AI conceptualizations has been found around principles like transparency, justice and fairness, non-maleficence, responsibility, and privacy [1]. Ensuring that AI systems adhere to such fundamental system properties is expected to foster their perceived trustworthiness and increase stakeholder trust in AI technologies [2,3].

To incorporate these ethical principles in practice, a popular endeavor is to move responsible AI from principle to practice by operationalizing the derived characteristics for trustworthy AI applications. Mittelstadt [4], for example, confirms a lack of proven methods to translate AI ethics principles into practice, and Ryan and Stahl [5] argue that a mapping between higher-level principles and concrete methods is required to adopt them. Larsson [6] even more specifically concludes a "need for moving from principle to process in the governance of AI" (p. 437). Particularly, duties that arise for the AI-providing organization are rarely concretely defined, although, of course, AI providers bear a central role in the AI actors' ecosystem and thus the move from principles for AI ethics to responsible AI in practice [7]. Based on the high-level fundamentals of trustworthy AI, AI governance research, therefore, has started to focus on elaborating implications of the defined principles to determine characteristics of responsible AI systems (e.g., [5,8]) as well as mechanisms to implement those (e.g., [9–11]).

This shows that the need for responsible AI is widely recognized, and the operationalization of abstract principles to concrete actions is often identified as an appropriate way to bring them into practice. However, the implementation of agreed-upon practices within the industry has not been as comprehensive as one might hope. This discrepancy can, among others, predominantly be attributed to two prominent obstacles frequently cited by practitioners. First, many organizations hesitate to take substantial action until comprehensive regulations are put in place [12]. Often, waiting for clear regulatory guidelines is preferred over additional efforts and burdens if one's own initiatives do not align with future provisions. Second, a lack of uniform, standardized approaches to translate the agreed guidelines to easily implementable and effective actions has been noted within the AI community [4,13,14], which is seen as a significant hurdle to their practical adoption. The diverse landscape of measures and guidelines proposed by various entities creates confusion and makes it challenging for organizations to discern the most appropriate path to follow. In addition, there is uncertainty about which specific measures are the most effective in ensuring trustworthy AI [12]. This ambiguity can inhibit decision-makers and make them reluctant to commit to any particular strategy.

Our article shall support practitioners in finding a solution to these problems. We approach the issue by moving the focus away from solely the trustworthiness of the product itself to a stronger focus on the perceived trustworthiness of the development process. While trustworthy AI is often defined as system requirements, in order to tangibly operationalize it, we need to understand its link to the measures along the development process. Therefore, we propose a concept for a trustworthy development process for AI systems. Our suggested framework is built off a semi-systematic literature analysis of AI governance efforts to derive obligations and measures to fulfill agreed AI ethics requirements and map them onto the AI development lifecycle. The results are compared to the implications for practical AI governance from the landscape of EU-centered regulatory frameworks.

Our research can support AI practitioners, particularly regarding the two major problems mentioned above. The review-based methodology shows a growing consensus regarding prominent measures of corporate AI governance. Incorporating well-known and diverse, action-oriented governance frameworks presents a unified summary and can provide clarity on which measures are generally proposed. The comparison with the EU-focused regulatory landscape shows a high degree of consistency between the proposed binding and non-binding measures and points to the core elements that can already be addressed without the final regulations. Finally, our proposed conceptual extension of trustworthy AI from solely a system configuration to the associated development process can be the first step towards a heuristic for determining the efficacy of a measure—the stronger it is linked to process trustworthiness characteristics, the stronger its potential for fostering stakeholders' trustworthiness perceptions, the underlying goal.

2. Theoretical Foundations of Trustworthy Processes for Operationalizing AI Governance

A fundamental goal for AI providers in operationalizing AI governance is to foster trust among their stakeholders. Societal actors are often predominantly impacted by the outcomes of AI, although they can only indirectly influence their system design. This results in a need for these stakeholders to trust in the responsible development of AI systems by the AI provider and other contributors and, thus, in return, their induced obligation to indicate their trustworthiness back to society.

To display the resulting necessity of establishing concrete trustworthy processes for AI development activities as well as requirements and first steps towards this, we outline the theoretical foundations of trustworthy processes in the context of AI governance research in the following. We start with an overview of the AI ecosystem, including its stakeholders and their power over AI development processes, to display the need for trust and, therefore, examine the underlying goal of trustworthy AI development. Building off traditional trustworthy software development concepts, we subsequently review opportunities for signaling trust in the development process and derive characteristics that can guide process trustworthiness assessment. Finally, we examine related research on the translation of AI governance mechanisms into trustworthy development processes to examine its current status and reasons for the problems regarding operationalization in practice.

2.1. The Need for Trust in the AI Ecosystem

A major authority when establishing trust lies with providers of AI systems, as they are responsible for understanding the user's requirements and translating them into technical applications. With their two key roles in realizing AI projects, deciding over and developing the system along with its main characteristics, the AI provider naturally holds a large share of power in the development process [7,15,16].

However, of course, it is intertwined and shared with a variety of different stakeholders that contribute to the development of AI systems at different stages. The different phases that are needed to evolve from the first problem statement to the final AI system deployment and post-processing are outlined in the AI development lifecycle. It typically includes (1) the problem understanding and design phase, where the problem, its characteristics, and requirements are determined and a solution drafted; (2) the data collection and handling phase, where relevant data are obtained, preprocessed, analyzed, and managed; (3) the model building phase, involving the actual model development and testing; and (4) the deployment and monitoring phase, where the system is deployed to the user and monitored over time [17–20]. Figure 1 summarizes the four stages of the AI lifecycle and their various required tasks and introduces the major stakeholders involved in each step.

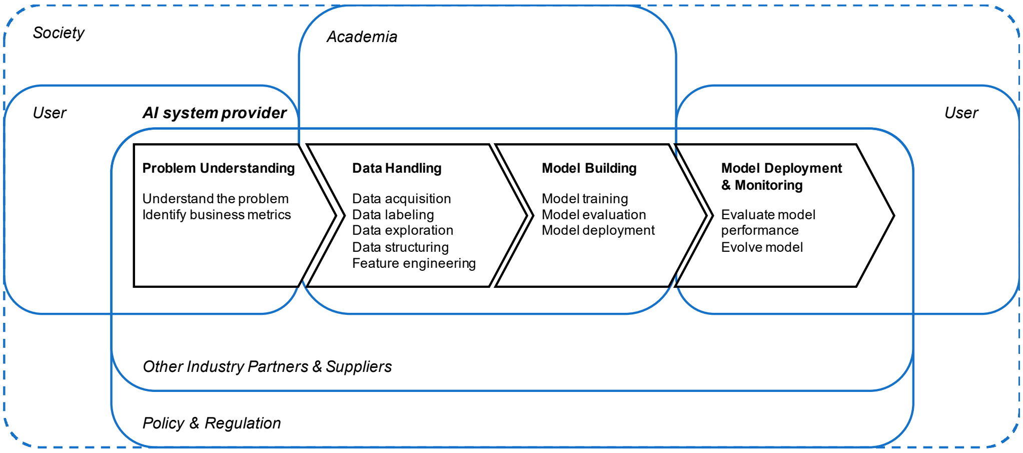


Figure 1. Stakeholder influence along the AI lifecycle.

The different stakeholder groups can influence different lifecycle stages to varying degrees. The user, representing the organization that operates the AI system after deployment, sets system requirements, supports problem understanding in the beginning, and engages and collaborates in deployment and after-monitoring at the end of an AI development project [10,21]. Apart from the system provider, this stakeholder group is also the one that can most actively control AI development. Other stakeholder groups, such as policy or academia, can consult, guide, or govern AI system development, e.g., through research, legislation, standards, and regulation; however, implementation of these guidelines is left to the AI system developers. Finally, the arguably most passively engaged stakeholder group relates to broader society. They often require representatives, such as civil society communities that consult policy and industry, to enforce their demands in the AI development process [15].

Figure 1 shows that the AI ecosystem is manifold, and stakeholders can express and assert their interests with varying influence. In particular, the conflict that broader society is highly influenced by AI systems, however, can most passively engage in their development is still unresolved. This is confirmed by the many guidelines for responsible or trustworthy AI that place societal values at the center of considerations. Legislative efforts have further highlighted the high importance of fundamental and human as well as civil rights and democratic values (e.g., the AI Bill of Rights or the EU AI Act), prioritizing lawful, safe, and trustable AI applications [2]. However, societal engagement's outlined rather passive nature requires them to trust that these values are responsibly integrated into the design and development processes, which opens room for exploring how this trust can be promoted and fulfilled.

2.2. Characteristics of Trustworthy AI Development Processes

While concrete requirements and best practices for trustworthy development processes are continuously discussed in the more specific field of AI, their identification can draw from the common ground determined for general software development. With its main aim to reach trustworthy products, i.e., software that can satisfy objectives of trustworthiness based on predefined requirements [22,23], a trustworthy development process is the procedure through which such trustworthy products are created [23], i.e., the procedure by which the requirements for considering the outcome trustworthy are ensured. For software in general, characteristics of trustworthy products have been agreed upon and are often reported among security, privacy, reliability, or business integrity [24]. The development of trustworthy AI applications can draw on these characteristics; however, its enhanced capabilities require further adaptations. The consensus on the ethical fundamentals of trustworthy AI has led to the definition of requirements for trustworthy AI systems, often around the concepts of human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability [2]. Most scholars from theory and practice argue that AI applications that fulfill these requirements can be deemed trustworthy.

This indicates that the foundations of system characteristics that trustworthy AI applications shall fulfill are already conceptualized. However, as pointed out earlier, ensuring trustworthiness only partially depends on the resulting system and, in addition, must take into consideration the process by which the system has been developed. Therefore, in order to evoke stakeholder trust, not only are the system characteristics that make an AI application trustworthy decisive, but also the characteristics that classify an AI development process as trustworthy. We can also benefit from the overlaps between AI development processes and regular software development. Although not finally settled, trustworthy development processes have been suggested and discussed for regular software. Standards like [25] on system life cycle processes for software engineering or [26] on a system security engineering approach have been assessed regarding their potential for introducing trustworthiness [27]. Further, standards like [28] or IEEE's approach to ethically aligned design [29] can give guidance on the central backbone of trustworthy development processes. The core goal in this endeavor is to enhance the predictability and controllability of development processes [23] and hence reduce the perceived dependability and uncertainty for the trustor. In particular, five characteristics are mentioned for development processes to be perceived as trustable. Figure 2 presents an overview of these characteristics.

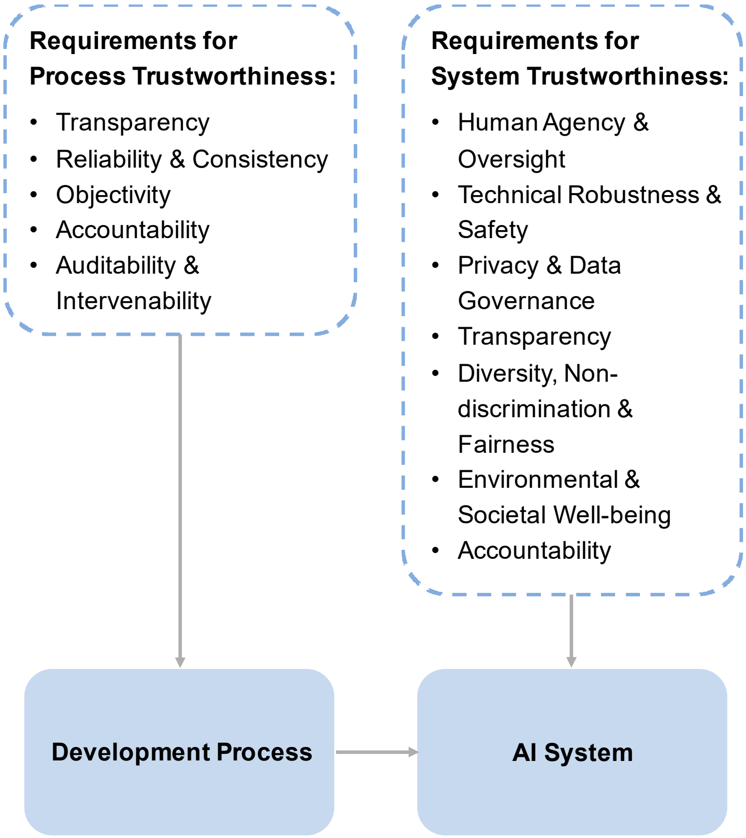

Requirements for Process Trustworthiness:
Transparency
Reliability & Consistency
Objectivity
Accountability
Auditability & Intervenability
Requirements for System Trustworthiness:
Human Agency & Oversight
Technical Robustness & Safety
Privacy & Data Governance
Transparency
Diversity, Non-discrimination & Fairness
Environmental & Societal Well-being
Accountability
Development Process
AI System

Figure 2. Requirements are presented that can be used to examine the trustworthiness of AI applications as outlined by the AI HLEG, as well as to examine the trustworthiness of the AI development process.

As a major pillar of trustworthiness, transparency requires clear and understandable communication about the steps taken during the AI development process, particularly making the process comprehensible and accessible to a wider audience, ensuring that people can understand how requirements for trustworthy systems are intended to be met. The steps taken within the development process to reach trustworthiness requirements must thus be clear, comprehensible, and communicable with a broader audience. The effects of transparency on trust have frequently been studied within and outside the information systems domain. Particularly within an AI context, transparency is often mentioned alongside explainability or interpretability, requiring that predictions or actions of AI systems are justifiable and traceable. The motivation behind transparency, however, differs according to the stakeholder and their perspectives. Weller [30], for example, lists eight types of goals that should be reached with transparency, ranging from transparency as a means for the developer to understand and debug a system to the facilitation of monitoring and auditing. Such explanations can foster the acceptance of the system and its design [31]. In a similar manner, transparency can facilitate acceptance regarding the planned development practices and is, therefore, a central factor for trustworthiness perceptions. On top, it lays the foundation for outside stakeholders to judge whether they perceive the envisioned steps as suitable for ensuring the responsible development of the system and thus have contributed to the prevention of undesired development practices.

Reliability and consistency are particularly important to enhance the process predictability, an essential component of fostering trust [32]. Hence, the development process must be reasonable, predictable, and standardized. A core goal is to prove to outside actors that processes follow a predefined plan and are not subject to arbitrary decisions or actions. At the same time, reliable and consistent development indicates that the goal of reaching responsible AI is pursued in a conscious and stringent way. A consistent process that is, at best, transparently communicated can reduce uncertainty for the trustor. Obligatoriness can thereby further enhance process predictability. Mandatory activities in the process can ensure the user that certain minimum requirements have been fulfilled and clarify upcoming steps. Moreover, process credibility and legitimacy are supported if mandatory steps are based on a legal foundation.

Objectivity in this context refers to the absence of favoritism when processing data or deriving outcomes. It requires executing every step of the process in a neutral and standardized manner without adapting steps to certain preferred implications or outputs. All steps in the development process must be executed in an objective, unbiased, and fair way.

Further, concrete accountabilities determined for each step of the development process can foster user trust as they define responsibility and create the opportunity for more clearly tracing system malfunctions to obligations within the development stage. Clear obligations and accountabilities must, therefore, be discernible for the steps within the development process. Accountability is a core requirement for considering trustworthy AI or within legal frameworks. The EU AI Act mandates the definition of accountability frameworks for high-risk AI systems, and the HLEG defines it as one of the key requirements for trustworthy AI [2]. In a similar manner, it is a core characteristic of trustworthy processes. Paulus, et al. [22], for example, mention it as a means of fostering security in trustworthy software development.

Finally, the perceived trustworthiness of development processes can be enhanced through enabling external monitoring and control. Auditability and intervenability, i.e., monitoring, checks, and options for intervention from external stakeholders, are, therefore, crucial process requirements. The potential of using auditing practices to ensure ethical AI design has lately been widely discussed [33]. Thereby, auditing refers to the process of examining the consistency between a "set of auditable artifacts that record decisions, systems, and processes" [9] and stated principles, regulations, norms, standard metrics, or benchmarks [9,34]. Such checks are common for other high-risk technologies, such as in aerospace or finance, and have been found promising in the context of AI [35]. The fundamental definition, however, also shows that practices for auditing responsible AI cannot limit themselves to system properties but must consider the steps taken to ensure ethical design along the development. Therefore, options for outside checks and intervention are important characteristics for both the system and its development process when aiming to enhance trustworthiness.

2.3. Moving towards Trustworthy AI Development

Works from different actors guide activities in the move of responsible AI to practice and thus the development of corporate AI governance strategies, e.g., from policymakers, standards development organizations (SDOs), or research and academia. In particular, regulatory efforts are often demanded and seen as an appropriate way to unify currently proposed approaches [12]. Regulations such as the EU AI Act, with its proposed risk categories and mandatory risk-dependent countermeasures, can clarify important groundwork for AI providers and provide consistent guidance on what measures to take or which red flags to avoid. However, regulation can and should specify AI governance mechanisms only on a conceptual level. Going down to more specific, process-based provisions is clarified by accompanying industry standards. For example, the US National Institute of Standards and Technology AI Risk Management Framework defines important actions along the development of AI systems to govern, map, measure, and manage AI-related risks [36]. IEEE Standard 7000, on a standard model process for addressing ethical concerns during system design on a broader scale, outlines a process for system engineers to incorporate ethical values into their design practices [28]. The ISO/IEC JTC 1/SC 42 working group on artificial intelligence has (up to now) 20 published standards and 32 under development [37]. Among the published ones are fundamental groundworks on the trustworthiness of AI (e.g., [38]) or AI risk management (e.g., [39]). More concrete guidance on, for example, how to integrate safety or transparency is currently under development.

Important work on operationalizing AI governance mechanisms also comes from research and the academic field. Here we see efforts within two, often interrelated streams: research proposing proactive approaches to set up responsible AI strategies or measures to actively implement those in the development process and reactive approaches referring to mechanisms to check—internally or externally—the responsibility of resulting systems or processes such as through auditing or impact assessment. Much research focuses on identifying and developing appropriate tools to integrate AI ethics into the design and development stage, particularly in proactive approaches. For example, overviews can be found on the current landscape of technical [8,10,11] and organizational [9,40] tools and methods. Important summaries on methods for responsible design, development, and deployment also come from the field of ethical AI assurance. Overviews and frameworks are proposed (e.g., [41,42]) and are slowly transitioning from mere theoretical considerations to actionable tools [43]. In contrast, reactive approaches focus less on the active adaptation of the development process and more on the evaluation of the resulting system. Nevertheless, they have important implications for the operationalization of responsible AI. Suggested impact assessments (e.g., [44,45]) can offer valuable guidance on the required system or process characteristics to implement. Auditing processes require the definition of "set[s] of auditable artifacts that record decisions, systems, and processes" ([9], p. 4) to allow checking them against predefined principles, metrics, or norms.

Therefore, while previous work already makes significant contributions to the operationalization of AI governance mechanisms, further efforts are needed to make them ready for practical application. Proposed tools can, at least from a theoretical perspective, solve many of the challenges; however, they are often considered unsuitable for their application in practice and are, therefore, rarely used [11,46,47]. In the field of AI assurance, Burr and Leslie ([41], p. 96), for example, specifically call for "practical systems and standards that can help teams and organizations" as a next research step. As a foundation for their adaptation to practical needs, a comprehensive understanding of the prominent measures is required to identify how they can support the development process.

With our article, we want to contribute to this transition to practice. The benefit and contribution of our research lie in integrating these approaches from the various stakeholder groups. Our primary objective is to provide a more comprehensive overview of the prominent measures and propel them into a tangible, actionable process to offer a unified, actionable summary—an asset that practitioners have pointed out as missing. This required a careful analysis of the multiple elements and their integration into a coherent framework. Finally, to measure its effectiveness, we examine the developed concept regarding its potential to facilitate the underlying goal: reflecting trustworthiness and, hence, fostering trust with stakeholders.

3. Methodology

Underlying our research is the assumption that in order to make an AI system trustworthy, we cannot only focus on the trustworthiness of the application itself but also take into account the trustworthiness of the development process. In this paper, our key objective is to derive the elements for a trustworthy process for responsible AI development from existing AI governance frameworks and discuss its potential for satisfying characteristics for process trustworthiness. We conceptualize the process from an analysis of regulatory and organizational frameworks that represent both legally binding and non-binding measures.

We have chosen a semi-systematic qualitative literature analysis methodology. This research method is used for heterogeneous topics that are conceptualized and studied by a wide variety of research disciplines [48]. In particular, it is needed where fully systematic reviews are impossible due to the complexity or variety of research approaches, topics, and types [49]. We follow this methodology, as we aimed for practice-oriented frameworks, including standards, civil society comments, or policy efforts, and thus, a systematic keyword search on scientific search engines did not provide the anticipated results. Non-binding measures were collected from a variety of global AI governance frameworks and compared to legally binding measures that are currently enforced or planned in the EU.

3.1. Data Collection

Ethical and robustness-related obligations were obtained from four main fields of sources: (1) non-regulative policy efforts, (2) standards developing organizations (SDOs), (3) academic and research institutes or consortiums, and (4) non-governmental organizations (NGOs) or civil society groups. Actors to be investigated were retrieved from the aiethicist.org-repository [50]. The tabs "AI Principles" and "AI Governance" were systematically searched for stakeholders active in the field who had published documents giving insights on obligations and responsibilities for AI providers. As the primary focus of the study is how to operationalize trustworthy AI in practice, only documents that went beyond defining AI principles in general but included practice-orientated recommendations were included. A specific link to AI systems was required, i.e., documents where the primary focus lay on data governance were excluded. Further, these documents needed to be referenced on the respective stakeholder's website or found through regular online searches.

This systematic search resulted in a total of 155 considered stakeholders, of which documents on the responsible development of AI were found from 45 actors. Finally, 14 of these were found to have published practice-oriented reports on AI-related obligations, measures, or responsibilities, meeting the outlined inclusion criteria. Four documents were from non-regulative policy efforts [2,51–53], three were from SDOs [29,36,54], five documents were published by research institutes or academic consortiums [55–59], and two by NGOs or other civil society interest communities [60,61]. An overview of all retrieved stakeholders and those of which publications were included in the analysis can be found as Supplementary Material.

The main aim of the legal analysis was to provide an overview of thematic fields from which legal obligations for AI providers can arise when developing or deploying AI systems. The EU AI Act was used as a first point of reference to identify those. While it provides the backbone of AI-specific obligations, the EU AI Act's Explanatory Memorandum (particularly its sections 1.2 and 1.3) was used to determine related regulative fields. A context-independent search confirmed and extended a first draft of the categorization into obligation topics resulting from the AI Act's Recitals. For this, the EUR-lex summary repository [62] was used. Administered by the EU, it provides overviews of the main EU legal acts. The listed 32 policy fields were searched for EU decisions that could have an impact and result in obligations for AI providers. Finally, a non-systematic literature review on legal obligations for AI systems via Google Scholar confirmed the resulting five obligation fields presented in Section 4.1.

3.2. Data Analysis

The resulting documents were analyzed to retrieve the elements of the trustworthy AI development process. As shown in Figure 3, the identified policy and governance recommendations were used to derive obligations for AI providers from principles for trustworthy AI. These obligations were mapped to related measures to fulfill the identified duties. The resulting process for trustworthy AI development was compared to legally binding obligations from current EU law.

The non-binding obligations and related measures from policy and governance documents were retrieved using a thematic analysis methodology. The principles for trustworthy AI, more specifically, the seven key requirements for trustworthy AI as proposed by the AI-HLEG, were used as guidance for the analysis. For each principle, related obligations were determined iteratively by repeatedly working through all documents and retrieving codes. Similar to the process described by Braun and Clarke [63], these codes were then translated to more overarching themes, which can be found in the Supplementary Material.

The resulting themes were reviewed and refined until they represented distinctive AI provider obligations.

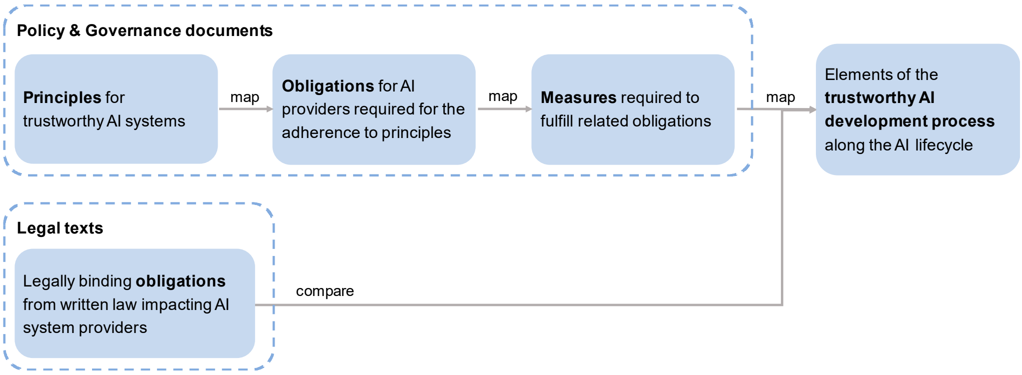

Policy & Governance documents
Principles for trustworthy AI systems
map
Obligations for AI providers required for the adherence to principles
map
Measures required to fulfill related obligations
map
Elements of the trustworthy AI development process along the AI lifecycle
Legal texts
Legally binding obligations from written law impacting AI system providers
compare

Figure 3. Procedure of mapping principles for trustworthy AI to a trustworthy AI development process.

In a similar manner, the documents were re-evaluated to retrieve the measures required to fulfill the identified obligations. For each identified obligation, related measures were derived through thematic analysis, retrieving and summarizing measure suggestions, and iteratively converging to a concise set of measures [64]. The resulting landscape of measures and the documents in which they were found can be reviewed in the Supplementary Material. Summarizing the resulting measures and mapping them along the AI lifecycle revealed the final elements of the trustworthy AI development process.

Finally, the resulting process was compared to obligations and measures suggested in binding legal texts. The identified regulatory fields and related EU-level legal documents were summarized regarding their imposed measures for the AI provider (see Supplementary Material). The resulting legally binding measures were compared to the developed trustworthy process, and elements of the process that were found legally enforced were marked accordingly. From this analysis, no new measures were added to the process, as no fundamentally new measures were found in the legal texts that directly impacted the providers' processes linked to the system's AI component.

4. From Trustworthy Principles to a Trustworthy Development Process

To investigate the elements as well as the state and implications of process trustworthiness in AI development procedures, we outline the derived framework for a trustworthy process of AI development from the existing fundamentals of trustworthy AI in the following. In particular, we elaborate on its foundation in the identification of AI provider obligations and the determination of related measures to address them.

4.1. From Principles to Obligations

The conducted analysis of obligations linked to responsible AI development supports the consensus on fundamental values that have been mentioned in the previous literature. Identified obligations are seen along the previously determined underlying principles. In Table 1, we report them along the seven key requirements for trustworthy AI due to their comprehensiveness and widespread acknowledgment. Within the columns, obligations are sorted according to the frequency with which they were found in the studied documents.

Table 1. Identified obligations for the AI provider along with AI HLEG's seven key requirements for trustworthy AI.

Human Agency and Oversight	**Technical Robustness and Safety**	**Privacy and Data Governance**	**Transparency**	**Diversity, Non-Discrimination, and Fairness**	**Societal and Environmental Well-Being**	**Accountability**
Ensure human autonomy/agency/determination Respect and protect fundamental/human rights Ensure human oversight Enable system termination Promote human augmentation	Ensure safety Ensure accuracy Ensure security Ensure reliability Ensure robustness Ensure validity Ensure reproducibility Ensure resilience to attack Ensure traceability Establish a fallback plan Ensure system quality Ensure verification	Ensure privacy Ensure data protection Ensure data quality Control data access Ensure lawful data processing Prevent data misuse/overuse Ensure data security Ensure data integrity Foster data risk awareness	Enable explainability of technical processes Communicate system capabilities and limitations Explain related human decisions/reasoning Ensure traceability of datasets and processes Inform about AI interaction Promote AI education Allow access for auditing Communicate intended use Ensure explicability Allow for intervention Ensure independence Ensure transparency on responsibilities Ensure truthfulness	Avoid/Correct/Monitor unfair bias Ensure non-discrimination Ensure diversity and inclusion Ensure equity, equality, and solidarity Ensure accessibility Ensure lawful development Enable multi-stakeholder engagement Enable compensation and remedy in case of discrimination Ensure peace and justice Define fairness Enable opportunity for correction	Prevent and reduce harm Monitor social impact Do more good than harm Ensure environmental friendliness Ensure proportionality to legitimate aim Ensure sustainability Monitor democratic impact Prevent misuse Establish multi-stakeholder dialog Ensure right foundation Ensure scientific foundation	Ensure auditability Provide documentation and information Assess general impacts Determine/assign responsibilities Allow for redress Establish appropriate oversight Establish ethics overseeing internal/external entity Establish measurement mechanisms Ensure public engagement Control access Foster accountability by design Create codes of conduct Collect feedback Ensure harm compensation

4.2. From Obligations to Measures

While the outlined obligations present the foundation for trustworthy AI development, moving closer towards implementation, measures to ensure their fulfillment can be derived. These measures were consolidated from existing AI governance frameworks and legal considerations with a focus on EU policy. The resulting list of binding and non-binding measures is presented in Table 2.

Table 2. Measures AI providers can adopt to fulfill their obligations according to the studied AI governance documents.

	Non-Binding	**Binding ***
Plan	Create **codes of conduct**	Develop AI **governance strategies** regarding: - trustworthy AI measurement and evaluation - data protection and access - quality management - risk management - human intervention - displacement and business change - privacy and accountability (by design) - education and awareness raising regarding harms and system misuse Determine/assign **responsibilities and accountabilities.** Set **requirements and thresholds** for: - system safety - accuracy, reliability - quality of data preparation and training - supporting hardware, software (incl. cloud applications) - industrial and consumer use case **Redress and compensation** for harms (incl. due to discrimination)

Table 2. *Cont.*

	Non-Binding	Binding *
Create and establish	Establish **participatory development processes** through: - pull mechanisms: offer public feedback opportunities, adoption of open standards, and interoperability to facilitate collaboration - push mechanisms: clarification of public concerns and questions, consultation of directly or indirectly affected stakeholders Create **ethics overseeing** internal/external entity Ensure **team diversity** regarding backgrounds, cultures, disciplines Establish **risk prevention/management** regarding: - wrong, unintended, or forbidden use of data - data modification or abuse - fairness-related harm - adversarial patch tricking - human errors - intentionally or unintentionally included biases Avoid, correct, and monitor **unfair bias** through: - removing identifiable discriminatory bias where possible - testing and monitoring mechanisms - evaluating how potential biases might arise Enable and ensure **human control** over data and processes **Educate** relevant personnel Ensure **oversight and control** regarding: - system purpose, constraints, requirements, decisions - shut down or modify misbehaving systems Enable **auditing** through: - developing audit trail requirements - provide access for internal or external auditing	Apply **systematic risk management** (incl. a fallback plan) Enable **human oversight** (human-on-the-loop) or **human control** (human-in-the-loop) by the user to: - assess and rectify incorrect predictions - avoid human subordination - avoid basing decisions with significant impact solely on automated processing - enable attribution of ethical and legal responsibility - terminate the system if human control of the system is no longer possible Provide options for **public intervention and participation** regarding: - choosing which digital services to use or to avoid using them - correcting false information - questioning and changing unfair, biased, or discriminatory systems - right to a final determination made by a person - consider bias and safety bug-bounty programs Ensure **explainability of technical processes**, e.g., through using tools, regarding: - system outcomes (incl. why similar-looking circumstances generate different outcomes) - logic or algorithm behind the outcome - main factors in a decision - data quality, accuracy
Assess and evaluate	Ex ante **impact assessments** regarding: - fundamental and human rights - privacy - society and societal norms - sustainability and environment - democracy - system's legitimate, proportionate, and scientific foundation Evaluate opportunities for quality labels and **certifications** Evaluate **independence of (critical) infrastructure** Ex post **impact assessment** regarding: - system accessibility - unfair denial of resources, rights, goods, participation	Assess compliance with applicable international and domestic **legislation, standards, and practices** **Test data quality** regarding: - accuracy - actuality - integrity - representativeness Assess and ensure **lawfulness** of data processing regarding: - protection of data and metadata - data access and control - user's freedom of intrusion - limiting observations **Test** system regarding: - accuracy/reliability (through model selection, measurement metrics, mitigation of model over-/underfitting) - robustness (through sensitivity analysis) - security (regarding data poisoning, model leakage, unexpected, adversarial or malicious use, cybersecurity threats) - discrimination (regarding use of protected classes and dataset representativeness) - validity - verification - domain-specific requirements (through simulation, in-domain testing, software/hardware requirements) Assess and ensure **lawful** development Assess and ensure **truthfulness** regarding statements to customers and consumers
Document and communicate	Support **AI education** through: - supporting educational curricula and public awareness activities - engaging with civil society to understand best practice for education	**Documentation** and record-keeping of: - datasets - data provenance, gathering, and labeling - data testing processes - data access **Document** for traceability and reproducibility: - provide replication files - use tools (e.g., to abstract computational graphs and archive data at each step of transformation pipelines) - adopt open standards **Disclose during use**: - interaction with AI system - which consumer actions can negatively impact scores/decisions **Communicate** with relevant stakeholders, e.g., in the form of use manuals: - definitions of key concepts and measures - system purpose, reach, (intended) use, capabilities, and limitations - design decisions, including characteristics of training data and model structure - data protection processes - responsible internal and external actors Provide **documentation** of: - system goals - design choices including definitions, standards, testing, measurements, and assessments of performance, privacy, and fairness - identified biases and their potential impacts - stakeholders involved and their responsibilities - risk management structures, including monitoring, feedback, error correction, human control, and cybersecurity

* Binding measures were found in the studied EU regulations or directives as mandatory for some AI technologies or under certain conditions.

4.2.1. Measures According to AI Governance Documents

Measures to address the obligations of AI providers were found, as similarly suggested in previous work by Mäntymäki, et al. [65], among activities for planning, assessment, and ensuring creation and communication.

Planning activities include mechanisms for establishing the strategic alignment of system development and associated risks regarding the AI provider's obligations, including the creation of governance strategies and codes of conduct, frameworks for risk management and accountability attribution, as well as determination and documentation of system requirements and thresholds. A fundamental goal to foster is the prevention of harms related to violation of rights, privacy, safety, security, sustainability, or, generally, the public good [52]. Considerations for strategy development should be based on conditions of normal use and potentially unanticipated but foreseeable use or misuse [51]. In addition, the diversity of cultural norms within the user groups should be taken into account, whereby the inclusion of different stakeholders during the creation of strategic directions can help [29].

Mechanisms to assess the fulfillment of certain AI provider obligations and ensure appropriate action-taking if needed include standardized impact assessments (including technical testing) before and after development on certain system properties and implications, as well as evaluation of system dependencies, lawfulness of data processing, and development, team characteristics and capabilities, options for auditing (including quality labels and certifications), truthfulness, and ensuring of compensation. The rationale behind this group of measures is to ensure that AI systems should only be deployed after a proper assessment of their purpose, objectives, benefits, and risks [61], bearing in mind that these assessments must be proportionate, rational, and methodologically sound [2].

The group of mechanisms to establish certain conditions comprises activities linked to obligations that require or impact the active (re-)design of the system or organizational processes linked to it. This is mainly required regarding the creation of participatory development and public intervention mechanisms, measures to ensure human oversights and control (including a focus on monitoring ethical aspects), as well as provision of certain documentation and enabling of explanation to ensure that processes can be appropriately steered. It is important for the implementation of such measures to take into account "shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals" [52].

Finally, communication activities are a prerequisite for multi-stakeholder engagement. Such measures are linked to the disclosure of certain information, communication of system definitions, purpose, limitations, risks, and use and education of staff, users, and the general public. Particular proactive engagement from AI providers has been demanded regarding the evaluation of augmenting human capabilities, advancing inclusion of underrepresented populations, reducing economic, social, gender, and other inequalities, and protecting natural environments [51]. For due and effective communication measures, the information provided needs to be understood and accurate and disclosed to the general public or the responsible human in charge [60].

4.2.2. Comparison to Legal Perspective

Requirements from legal frameworks have been found among five thematic fields and depend on the type and use of an AI-based application. Figure 4 presents an overview of the identified and analyzed regulatory texts.

To account for risks that arise specifically from the development and use of AI, the Regulation of the European Parliament and the Council for Laying Down Harmonized Rules on Artificial Intelligence and Amending Certain Union Legislative Acts, or short, the AI Act [66], suggests a four-level risk categorization into unacceptable, high, limited, and minimal risk systems. AI systems bearing unacceptable risks will be prohibited partially or in their entirety from use on the EU market; high-risk systems will be subject to conformity assessments, and thus, further restrictive measures will be implemented, mostly for the AI provider. While systems posing limited risks will need to ensure certain transparency requirements, minimal-risk systems are free of binding measures, only recommended to provide and follow self-imposed codes of conduct to voluntarily commit to the same response measures as high-risk systems.

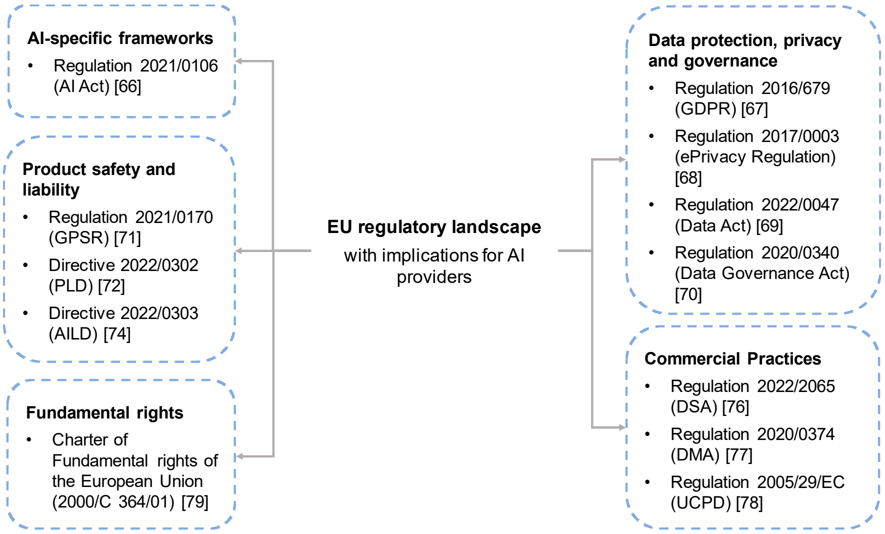

AI-specific frameworks
• Regulation 2021/0106 (AI Act) [66]
Product safety and liability
• Regulation 2021/0170 (GPSR) [71]
• Directive 2022/0302 (PLD) [72]
• Directive 2022/0303 (AILD) [74]
Fundamental rights
• Charter of Fundamental rights of the European Union (2000/C 364/01) [79]
EU regulatory landscape
with implications for AI providers
Data protection, privacy and governance
• Regulation 2016/679 (GDPR) [67]
• Regulation 2017/0003 (ePrivacy Regulation) [68]
• Regulation 2022/0047 (Data Act) [69]
• Regulation 2020/0340 (Data Governance Act) [70]
Commercial Practices
• Regulation 2022/2065 (DSA) [76]
• Regulation 2020/0374 (DMA) [77]
• Regulation 2005/29/EC (UCPD) [78]

Figure 4. Overview of the identified primary fields of the EU regulatory landscape that have impacted the responsibilities of AI providers.

As AI systems rely on data processing, they are subject to the restrictions set out by the General Data Protection Regulation (GDPR) [67], particularly when it comes to processing personal data or even special categories thereof. In such cases, the processing must meet additional safety and privacy requirements. In addition, special rights are granted to the data subject, which allow them to demand certain handling of their data from the data operator. While the GDPR, thus, directly entails mandatory measures for the AI provider, there are further EU regulations concerning data, both personal and non-personal, that affect the AI provider, however, only to a limited extent. Regulation 2017/0003 [68] on Privacy and Electronic Communications (ePrivacy Regulation), provides more concrete specifications for electronic communications data and thus complements GDPR, where such data qualify as personal. It generally restricts interference with electronic communications data, e.g., in the form of listening, tapping, storing, or monitoring, to certain permitted use cases (ePrivacy Regulation, Art. 5). The proposal for Regulation 2022/0047 [69] on harmonized rules on fair access to and use of data (Data Act) sets out obligations for the provision of data generated by the use of (physical) products that collect and transmit data. Regulation 2020/0340 [70] on European data governance (Data Governance Act) regulates the reuse and sharing of data between stakeholders in the EU to strengthen data availability and exchange. While these legislations may impact the AI-providing organization, e.g., if the system falls within one of the covered use cases or if the AI provider is involved in data collection and post-processing, their set out obligations, however, are linked to the overall system and have less direct implications for the AI component.

If AI components are treated as or built into products, requirements from product safety and liability can impose obligations and, therefore, require protective measures from the AI component provider. In this regard, the two constructs of the General Product Safety Directive (GPSD) [71] and Product Liability Directive (PLD) [72] provide two complementary mechanisms for enforcement of damage-related consumer claims, where the PLD outlines the liability specifications to assert claims that result from a defect or unsafe product, and product safety regulations lay down the specifications that a product must adhere to in order to be considered safe. While the AI Act outlines some safety-related AI requirements, the new GPSR "provides a safety net for products and risks to health and safety of consumers that do not enter into the scope of application of the AI proposal" (GPSR), and therefore AI-equipped products that are not subject to the more specific safety rules of the AI Act, i.e., products where the AI component is considered to pose only minimal risk, must comply with the provisions of the GPSR [73]. Nevertheless, the legislator sees a particular need for action regarding AI and supports the PLD with the proposition of an AI Liability Directive (AILD) [74] that explicitly addresses claims in relation to AI-based systems. Here, especially access to information on high-risk systems and the burden of prove shall be adapted to the specific circumstances of AI (AILD, Art. 1(1)).

In order to "create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses" [75], the EU has adapted and extended some of its existing regulations in the area of commercial practices to further strengthen consumer rights and trust. This particularly includes the development of Regulation 2022/2065 [76], the Digital Services Act (DSA), and Regulation 2020/0374 [77], the Digital Markets Act (DMA). The DSA sets restrictions for digital services providers, defined as providers of "intermediary service", such as conduit, caching, or hosting services. AI-based use cases are affected in many ways, predominantly in the form of online advertisement targeting and recommender systems. The DMA similarly imposes additional obligations for "gatekeepers", i.e., large providers of core platform services, such as Amazon, Apple, Google, Meta, or Microsoft. While most obligations relate to restrictions of limiting access to data and services, AI-based applications are particularly addressed when it comes to recommender systems. In addition to the specific rules of the digital domain, existing legal frameworks on commercial practices also carry responsibilities for the AI provider. Directive 2005/29/EC [78], the Unfair Commercial Practices Directive (UCPD), prohibits misleading and aggressive commercial practices. Particularly, restriction of misleading practices can impact AI applications, as the AI provider must not provide false information or deceive consumers regarding, among others, the main characteristics of products, such as their benefits and risks, and compliance with promoted codes of conduct.

Finally, human or fundamental rights, founded in the Charter of Fundamental rights of the European Union (2000/C 364/01) [79], are the foundational basis for most of the above-outlined legal frameworks. In the implementation of Union law, they apply between EU institutions and bodies and the people; therefore, in the context of AI, no direct obligations for AI providers can be inferred from the legal text. Nevertheless, they are frequently mentioned in the context of AI, as risks and challenges for the respect of human rights are often identified with the introduction of AI systems [80]. This results in a direct obligation for the state to protect its citizens from restrictions on fundamental rights and an indirect obligation for the AI provider to comply with the stipulated provisions and measures against imposing restrictions on fundamental rights.

In summary, the policy takes a similar view to AI governance and has already incorporated some of the determined measures into regulatory and legal guidance. The conditions under which they are mandatory depend on the particular AI application, e.g., a systematic risk management procedure is required only for AI systems classified as high-risk under the EU AI Act, or the engaged target group, e.g., truthful statements regarding a system's properties and capabilities are particularly mandatory for communication with consumers according to UCPD. Therefore, we would like to note that our indications in Table 2 regarding binding practices should not be read as legal advice on which mechanisms to implement but constitute an analysis of which methods are regarded as both relevant and generalizable enough by legislators to be mandated across multiple use cases. This objective can provide interesting insights when examining which measures are suggested as mandatory or not and is useful for the development of a standardized, trustworthy AI development process.

4.3. From Measures to Process

The determined measures result in a framework for the trustworthy development of AI systems. Figure 5 outlines the derived activities and outputs.

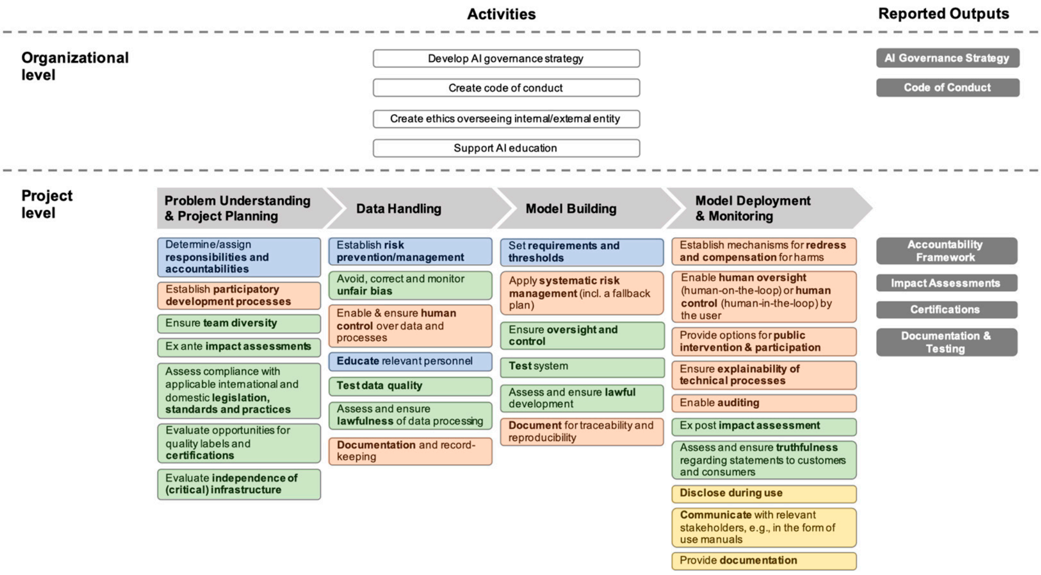

Activities
Reported Outputs
Organizational level
Develop AI governance strategy
Create code of conduct
Create ethics overseeing internal/external entity
Support AI education
AI Governance Strategy
Code of Conduct
Project level
Problem Understanding & Project Planning
Data Handling
Model Building
Model Deployment & Monitoring
Determine/assign responsibilities and accountabilities
Establish participatory development processes
Ensure team diversity
Ex ante impact assessments
Assess compliance with applicable international and domestic legislation, standards and practices
Evaluate opportunities for quality labels and certifications
Evaluate independence of (critical) infrastructure
Establish risk prevention/management
Avoid, correct and monitor unfair bias
Enable & ensure human control over data and processes
Educate relevant personnel
Test data quality
Assess and ensure lawfulness of data processing
Documentation and record-keeping
Set requirements and thresholds
Apply systematic risk management (incl. a fallback plan)
Ensure oversight and control
Test system
Assess and ensure lawful development
Document for traceability and reproducibility
Establish mechanisms for redress and compensation for harms
Enable human oversight (human-on-the-loop) or human control (human-in-the-loop) by the user
Provide options for public intervention & participation
Ensure explainability of technical processes
Enable auditing
Ex post impact assessment
Assess and ensure truthfulness regarding statements to customers and consumers
Disclose during use
Communicate with relevant stakeholders, e.g., in the form of use manuals
Provide documentation
Accountability Framework
Impact Assessments
Certifications
Documentation & Testing

Figure 5. Core framework of a trustworthy development process for AI applications. Colors indicate measures to plan (blue), assess and ensure (green), create (orange), and communicate (yellow).

Measures within the process were found on two levels: activities that relate to strategic decision-making and, therefore, must be clarified on an organizational level and activities that depend on the use case or context-specific circumstances and, therefore, are specified per project. Organizational-level activities include strategy-setting tasks, such as developing a corporate AI governance roadmap or planning AI education and training offerings within and outside the organization. Project-level activities are structured along the AI development lifecycle, outlining the tasks that are required in each stage according to our analysis. Reporting certain outputs within or outside the organization is recommended on both levels.

5. Discussion

The proposed conceptual framework of a trustworthy development process for AI applications outlines the measures that have been identified to ensure the trustworthiness of AI applications regarding their responsible development and use. The identification of prominent AI governance measures from regulatory, policy, standardization, and research activities can support AI providers in setting up appropriate corporate AI governance strategies. In the following, we discuss its potential for fostering outlined process trustworthiness characteristics as well as determine the next steps in the practical implementation of such activities and processes.

5.1. Process Trustworthiness of Current Responsible AI Development

Previous research has indicated that trustworthiness can be enhanced by facilitating stakeholders' perception of certain key process characteristics. Transparency, reliability and consistency, objectivity, accountability, and audibility and intervenability were seen among the most crucial ones for procedural trustworthiness. To foster those, currently proposed measures for trustworthy AI and, more specifically, the derived process for trustworthy development of AI systems can show promising potential, above all by two important properties of the derived development process.

First, the clear structure it introduces can offer stakeholders insight into the measures that will be taken to prevent ethical problems or irresponsible AI. The implementation of a clear and structured framework to address responsible and trustworthy AI has become essential for stakeholders in the field; however, clarity regarding required measures is often mentioned as a major drawback [12]. The derived development process outlines a series of concrete steps and actions that can be taken at different levels to prevent and mitigate potential ethical challenges and the proposed outcomes to be communicated more broadly. This can have an impact on transparency on several levels, firstly through the fact that there is a concrete process, which in the best case is also openly communicated, and secondly, more obviously, through the proposed publicly reported outputs. In that, it promotes two important functions of transparency. The clear structure of measures helps stakeholders understand a system's strengths and limitations, and the clear communication of processes and outputs allows for checking whether a system works appropriately and, in this case, ethically [30]. In addition, its second function, the disclosure of processes and thus comprehensibility of mitigating measures, influences the perception of auditability. As previously outlined, such auditing processes require artifacts and evidence to be checked against the established ethical principles [9,34]. The determined development measures make it easier to meet this requirement, as the artifacts needed for auditing can be derived from the concrete actions taken throughout the process. Trustworthiness can further be enhanced through the introduction of legitimacy. The proposed measures result from the analysis of existing framework conceptions; most have been subject to a consultation period involving relevant stakeholders in the framework's development. This ensures that the decisions made are well-founded and align with the values and expectations of the broader community. Finally, obligatoriness is signaled in the process. The derived development process identifies certain steps as mandatory, reinforcing its commitment to ethical practices and responsible AI. This emphasizes the seriousness of the endeavor and underscores the importance of adhering to these guidelines in AI-related initiatives.

A second key property of the derived development measures that can influence perceived procedural trustworthiness is the breakdown of steps into specific tangible actions. Such a step-by-step approach can foster accountability by breaking an AI provider's ethical obligations down to related countermeasures, enabling the definition of responsibilities on this level. This, in return, further enhances an essential function of explanation and, thus, a key requirement for trustworthy AI, enabling the meaningful challenge of an AI system's outputs [30].

However, while such an approach can thus ensure much improvement of procedural trustworthiness, there are also certain characteristics in the current form of the process that demand further reiteration to fully support trustworthiness perceptions.

One significant issue that pertains is the vagueness of some steps within the process. While many steps are well-defined and offer a clear roadmap for addressing ethical concerns, some remain ambiguous. This lack of preciseness can hinder perceptions of the reliability of the overall process. If certain steps are open to interpretation, it might be left to individual organizations or decision-makers to decide how they are implemented. This subjectivity could lead to inconsistent practices and potentially compromise the overall effectiveness of the process. Moreover, the binding nature of the procedure is not always clear. Even in the investigated legally mandated measures, there is some leeway on the interpretation and requirements for implementation. The level of obligatoriness might vary depending on the specific use case, context, or area of implementation. Such inconsistencies could raise questions about the enforceability and effectiveness of the process. Stakeholders might wonder whether the process's guidelines are universally binding or if they can be overlooked or circumvented in certain circumstances, leading to potential compromises in the perceived trustworthiness of the process.

A final point of consideration when evaluating the process's effect on trustworthiness is its ability to actually ensure the system characteristics that are seen as required for trustworthy AI. The question can be raised whether merely following the prescribed steps is sufficient to ensure a trustworthy AI system. For example, while we have seen that the clarified steps can promote some functions of transparency, others, such as the need to "overcome the reasonable fear of the unknown" [30], are not necessarily addressed. The "unknown" is often related to malfunctioning or misuse of the system, but it is not necessarily clear whether the proposed measures are sufficient to avoid this in the best possible way. The consensus identified on certain measures to ensure the trustworthiness of AI systems and the high level of attention currently being paid to this issue may indicate that at least all the obvious measures have been defined. However, whether these measures can finally lead to enhanced trustworthiness of the systems and more stakeholder trust in the development procedure is neither clear for procedural trustworthiness nor is it solved in general.

5.2. Next Steps in Trustworthy AI Development

Our analysis indicates a high-level consensus around measures that can be taken to mitigate ethical issues of AI systems along the development process. Clarity is often mentioned by practitioners as a major drawback to the realization of AI ethics principles. While the entirety of developed concepts can bring clarification on what is generally required from AI providers to ensure responsible AI, the vagueness of some proposed measures is also apparent in our analysis. Particularly when breaking obligations down into respective measures, we see that while some measures seem to be quite "ready-to-use", others require further efforts to apply them to real-world scenarios. For example, while the task of determining and assigning responsibilities seems straightforward, recent studies have indicated that accountabilities for AI systems are often ambiguous in reality, calling for the creation of detailed accountability frameworks [12,81,82]. In implementing such development processes, it is important to identify with practice which of the measures are already implementable and which need more detailing. The context will certainly play a major role here, as different industry use cases have different characteristics and requirements regarding the demand for and magnitude of obligations [83]. We see that legislation already accounts for this, and, for example, the current proposal of the AI Act classifies AI systems according to their use case industry into risk levels. The scope of imposed, binding measures followingly depends on the risk that results from a system's use. While our analysis gives an overview of which measures exist in general, we make no assumptions about which of them are relevant or more important than others given certain contexts. Clearly, a simultaneous implementation of all the measures described does not seem realistic or desirable for all AI systems. Future research can, therefore, focus on how the summarized obligations can be applied to real case studies and what this implies for the trustworthiness of AI systems.

Further, while recent research often focuses on the technical implementation of responsible AI, for example, through tools or "by-design" concepts, our analysis shows that AI governance mechanisms currently are largely recommended among non-technical methods, for example, regarding strategy-making, documentation, and communication. This either reemphasizes the inability of available technical tools to meet the needs of practice or suggests that careful consideration must be given when automated or "by-design" approaches are feasible and desirable and when a non-technical assessment of activities, perhaps based on human intuition, is required. For instance, although technical tools to detect bias in training data might be helpful, a thorough interpretation of the results regarding whether the identified biases result from unfair assumptions will surely be needed. The notion that AI governance measures are largely non-technical methods further impacts the current view on accountability for AI systems. In comparison, system developers and designers are often named as responsible entities for ensuring responsible AI [81]. Most of the identified measures are not directly linked to system implementation and would require inputs or action from further departments, such as those related to strategy, communication, and management. This suggests that a more concrete examination of what actions should be taken to fulfill which obligations, taking into account the context of application and the characteristics of the expected mechanisms, can further facilitate the identification of the bodies or roles that can be held accountable for the outcomes of an AI system.

Finally, our analysis presents a comprehensive list of obligations and measures that current, practice-oriented AI governance frameworks offer to ensure the trustworthy behavior of AI. It, however, cannot answer the question of whether implementing these measures, to a reasonable extent and with reasonable effort, will finally increase stakeholder trust and lead to a proper realization of the goal of responsible AI. Trust is a property of an individual trustor, who, based on personal perceptions and experiences, beliefs that an outcome is beneficial enough to engage in an unknown situation [22,84]. Being a psychological state of the trustor based on their subjective perceptions and decisions [85], trust can only be influenced; however, it cannot be directly controlled by the trusted organization. Both theoretical and empirical research suggest that certain measures to implement ethical considerations into system design and hence to signal trustworthiness positively influence the stakeholder perceptions—for example, measures to enhance fairness and individuality can promote user satisfaction with applications [86] or certain system design choices such as human-in-the-loop architectures can help reduce algorithmic aversion [87]. Nevertheless, the final decision remains to the respective stakeholders. Therefore, in practice, appropriate techniques for measuring the impact achieved, as well as continuous monitoring and re-evaluation of the measures implemented, will be key.

5.3. Limitations

Our derived framework of a trustworthy AI development process provides a unified overview of corporate AI governance mechanisms as proposed by various stakeholder groups, and the resulting clarity can thus support AI providers in the development of responsible AI governance strategies. However, there are also limitations to the scope of our results.

The chosen methodology was found suitable and required for determining practice-oriented recommendations from a variety of stakeholders. However, due to the semi-systematic approach, it is possible that other similarly relevant documents were not considered. In addition, given the rapid growth of this field, further practical guidelines may emerge which have not been included in this research. However, given the comprehensive approach to the identification of stakeholders, the diversity of considered stakeholder groups, and the similarity of identified measures as determined in the analysis, we do not see this as a weakness of our study. A second limitation stems from the required geographical focus when analyzing the implications of regulation. It was necessary to limit the scope either in breadth or in depth, which is why we opted for a granular review only of the EU regulatory landscape. However, given the EU's leadership in responsible AI governance and their advanced regulation in this field, we regard this as a minor limitation to our analysis and, in contrast, see valuable insights for other geographical areas.

6. Conclusions

The need to detail principles of ethical AI and adapt them for operationalization in practice is repeatedly emphasized in various fields. Roads to this goal are seen in assessing the current governance landscape, further clarification, and detailed conceptualizations [4,13]. Particularly from a practitioner's perspective, unification and clarification regarding responsibilities and related measures are needed to support them in establishing appropriate corporate AI governance strategies.

Our research aims to support these required objectives by advancing research on trustworthy development processes for AI applications. We explored the essential characteristics that define a process as trustworthy, drawing upon traditional concepts from trustworthy software development. By investigating the fulfillment of these trustworthy development propositions within frequently proposed trustworthy AI measures, we assessed whether they can effectively ensure responsible AI applications. Through a semi-systematic literature analysis of AI governance efforts and EU-centered regulatory frameworks, we translated agreed AI ethics requirements into practical obligations and derived the measures suggested to fulfill them. By mapping these measures onto the AI development lifecycle, we conceptualized the framework of a trustworthy AI development process.

Our research can, therefore, provide important insights for the practical implementation of AI governance measures. Obligations of AI providers to comply with the agreed ethical principles of AI were determined, and the corresponding measures that can be implemented to fulfill these were systematically retrieved. The resulting concept of a process for the trustworthy development of AI systems can help support clarity on the state of the art of demanded mechanisms. Finally, the discussion on the degree of process trustworthiness that can be fulfilled with such a process sheds light on the overall state of trust in AI applications.

While our analysis can thus provide much clarification regarding the steps toward principle operationalization, it also sheds light on the open questions that will be clarified next. While a general catalog of measures applicable across various application scenarios seems useful to obtain a standardized overview, a case-based consideration is needed to identify which obligations and related measures are seen as particularly relevant for certain uses and, more generally, how to determine a heuristic to discover these variances. Further, a clearer distinction in which use cases or contexts automated technical approaches are feasible and desirable, and thus, developing technical tools to implement them is needed. Finally, our results provide a comprehensive overview of the tasks required to fulfill certain AI provider obligations. Whether these tasks are reasonable in practice and particularly whether they are enough to thoroughly consider an AI system as appropriately responsible might require further measurement mechanisms or independent assessments.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai4040046/s1, Table S1: Data Sources; Table S2: Obligations and Measures retrieved from the Governance Documents Analysis; Table S3: Requirements and Obligations retrieved from the Legal Analysis.

Author Contributions: Conceptualization, E.H.; Methodology, E.H.; Formal analysis, E.H.; Investigation, E.H.; Resources, E.H.; Writing—original draft, E.H.; Writing—review & editing, E.H.; Visualization, E.H.; Supervision, C.L.; Project administration, E.H. and C.L.; Funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Fujitsu Limited and the Technical University of Munich's Institute for Ethics in Artificial Intelligence (IEAI).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
2. High-Level Expert Group on Artificial Intelligence (AI HLEG). *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
3. Bartneck, C.; Lütge, C.; Wagner, A.; Welsh, S. *An Introduction to Ethics in Robotics and AI*; Springer Nature: Berlin, Germany, 2021.
4. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
5. Ryan, M.; Stahl, B.C. Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *J. Inf. Commun. Ethics Soc.* **2020**, *19*, 61–86. [CrossRef]
6. Larsson, S. On the governance of artificial intelligence through ethics guidelines. *Asian J. Law Soc.* **2020**, *7*, 437–451. [CrossRef]
7. Deshpande, A.; Sharp, H. Responsible AI Systems: Who are the Stakeholders? In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 22), New York, NY, USA, 7–8 February 2022; pp. 227–236.
8. Georgieva, I.; Lazo, C.; Timan, T.; van Veenstra, A.F. From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics* **2022**, *2*, 697–711. [CrossRef]
9. Ayling, J.; Chapman, A. Putting AI ethics to work: Are the tools fit for purpose? *AI Ethics* **2021**, *2*, 405–429. [CrossRef]
10. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **2021**, *55*, 1–46. [CrossRef]
11. Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 153–183. [CrossRef]
12. Hohma, E.; Boch, A.; Trauth, R.; Lütge, C. Investigating accountability for Artificial Intelligence through risk governance: A workshop-based exploratory study. *Front. Psychol.* **2023**, *14*, 1–17. [CrossRef] [PubMed]
13. Stix, C. Actionable principles for artificial intelligence policy: Three pathways. *Sci. Eng. Ethics* **2021**, *27*, 15. [CrossRef]

14. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018; Volume 1442.
15. Miller, G.J. Stakeholder roles in artificial intelligence projects. *Proj. Leadersh. Soc.* **2022**, *3*, 100068. [CrossRef]
16. Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 20), New York, NY, USA, 27–30 January 2020; pp. 1–18.
17. De Silva, D.; Alahakoon, D. An artificial intelligence life cycle: From conception to production. *Patterns* **2022**, *3*, 100489. [CrossRef]
18. Haakman, M.; Cruz, L.; Huijgens, H.; van Deursen, A. AI lifecycle models need to be revised. *Empir. Softw. Eng.* **2021**, *26*, 95. [CrossRef]
19. Suresh, H.; Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ’21), New York, NY, USA, 5–9 October 2021; pp. 1–9. [CrossRef]
20. de Souza Nascimento, E.; Ahmed, I.; Oliveira, E.; Palheta, M.P.; Steinmacher, I.; Conte, T. Understanding development process of machine learning systems: Challenges and solutions. In Proceedings of the 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Porto de Galinhas, Brazil, 19–20 September 2019; pp. 1–6. [CrossRef]
21. Rybalko, D.; Portilla, I.; Kozhaya, J.; Ishizaki, K.; Hall, K.; Madan, N. *AI Model Lifecycle Management: What is ModelOps? A Technical Perspective*; IBM Point of View: Armonk, NY, USA, 2020.
22. Paulus, S.; Mohammadi, N.G.; Weyer, T. Trustworthy software development. In Proceedings of the Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany, 25–26 September 2013. pp. 233–247.
23. Yang, Y.; Wang, Q.; Li, M. Process trustworthiness as a capability indicator for measuring and improving software trustworthiness. In Proceedings of the Trustworthy Software Development Processes: International Conference on Software Process, ICSP 2009, Vancouver, BC, Canada, 16–17 May 2009; pp. 389–401.
24. Safonov, V.O. *Using Aspect-Oriented Programming for Trustworthy Software Development*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
25. Systems and Software Engineering—System Life Cycle Processes, ISO/IEC/IEEE. 2015. Available online: https://www.iso.org/standard/81702.html (accessed on 29 September 2023).
26. Developing Cyber-Resilient Systems: A Systems Security Engineering Approach, NIST. 2021. Available online: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v2r1.pdf (accessed on 29 September 2023).
27. Shiang-Jiun, C.; Yu-Chun, P.; Yi-Wei, M.; Cheng-Mou, C.; Chi-Chin, T. Trustworthy Software Development—Practical view of security processes through MVP methodology. In Proceedings of the 2022 24th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 13–16 February 2022; pp. 412–416.
28. IEEE Standards Association. *Addressing Ethical Concerns During Systems Design*; IEEE Standards Association: Piscataway, NJ, USA, 2021; Volume 7000.
29. IEEE Standards Association. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*; IEEE Standards Association: Piscataway, NJ, USA, 2019.
30. Weller, A. Transparency: Motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–40.
31. Wanner, J.; Herm, L.-V.; Heinrich, K.; Janiesch, C. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electron. Mark.* **2022**, *32*, 2079–2102. [CrossRef]
32. Tyler, T.R. Why Do People Rely on Others? Social Identity and Social Aspects of Trust. In *Trust in Society*; Cook, K.S., Ed.; Russell Sage Foundation: New York, NY, USA, 2001; pp. 285–306.
33. Mökander, J.; Floridi, L. Operationalising AI governance through ethics-based auditing: An industry case study. *AI Ethics* **2023**, *3*, 451–468. [CrossRef] [PubMed]
34. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv* **2020**, arXiv:2004.07213.
35. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 20), New York, NY, USA, 27–30 January 2020; pp. 33–44.
36. NIST. *AI Risk Management Framework: Initial Draft*; NIST: Gaithersburg, MD, USA, 2022.
37. ISO. ISO/IEC JTC 1/SC 42 Artificial Intelligence. Available online: https://www.iso.org/committee/6794475/x/catalogue/p/1/u/0/w/0/d/0 (accessed on 19 December 2022).
38. ISO/IEC. *Information Technology—Artificial Intelligence—Overview of Trustworthiness in Artificial Intelligence*; ISO: Geneva, Switzerland, 2020. Available online: https://www.iso.org/standard/77608.html (accessed on 29 September 2023).
39. ISO/IEC. *Information Technology—Artificial Intelligence—Guidance on Risk Management*; ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/77304.html (accessed on 29 September 2023).
40. Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Abrahamsson, P. The current state of industrial practice in artificial intelligence ethics. *IEEE Softw.* **2020**, *37*, 50–57. [CrossRef]

41. Burr, C.; Leslie, D. Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. *AI Ethics* **2023**, *3*, 73–98. [CrossRef]
42. Ashmore, R.; Calinescu, R.; Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 111. [CrossRef]
43. AI Assurance Guide. Available online: https://cdeiuk.github.io/ai-assurance-guide/ (accessed on 21 September 2023).
44. Ada Lovelace Institute. *NMIP Algorithmic Impact Assessment User Guide*; Ada Lovelace Institute: Londond, UK, 2022.
45. High-Level Expert Group on Artificial Intelligence (AI HLEG). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*; European Commission: Brussels, Belgium, 2020.
46. Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Siponen, M.; Abrahamsson, P. Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv* **2019**, arXiv:1906.07946.
47. Greenstein, B.; Rao, A. PwC 2022 AI Business Survey. Available online: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html (accessed on 29 September 2023).
48. Wong, G.; Greenhalgh, T.; Westhorp, G.; Buckingham, J.; Pawson, R. RAMESES publication standards: Meta-narrative reviews. *J. Adv. Nurs.* **2013**, *69*, 987–1004. [CrossRef]
49. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
50. Aiethicist.org. Artificial Intelligence Resources; Aiethicist.org: 2022. Available online: https://www.aiethicist.org (accessed on 21 September 2023).
51. OECD. *Recommendation of the Council on Artificial Intelligence*; OECD/LEGAL/0449; OECD: Paris, France, 2019.
52. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021.
53. US Federal Trade Commission (FTC). *Aiming for Truth, Fairness, and Equity in Your Company's Use of AI*; US Federal Trade Commission (FTC): Washington, DC, USA, 2021.
54. CEN-CENELEC Focus Group. *Road Map on Artificial Intelligence (AI)*; CEN-CENELEC: Brussels, Belgium, 2020.
55. Elam, M.; Reich, R. *Stanford HAI Artificial Intelligence Bill of Rights: A White Paper for Standford's Institute for Human-Centered Artificial Intelligence*; Stanford Human-Centered Artificial Intelligence: Stanford, CA, USA, 2022.
56. Felländer, A.; Rebane, J.; Larsson, S.; Wiggberg, M.; Heintz, F. Achieving a Data-driven Risk Assessment Methodology for Ethical AI. *Digit. Soc.* **2021**, *1*, 1–13. [CrossRef]
57. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
58. Reisman, D.; Schultz, J.; Crawford, K.; Whittaker, M. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. *AI Now* **2018**. Available online: https://www.nist.gov/system/files/documents/2021/10/04/aiareport2018.pdf (accessed on 21 September 2023).
59. The Responsible Machine Learning Principles: A Practical Framework to Develop AI Responsibl. The Institute for Ethical AI & Machine Learning: London, UK. Available online: https://ethical.institute/index.html (accessed on 21 September 2023).
60. Loi, M.; Matzener, A.; Muller, A.; Spielkamp, M. *Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities*; AW AlgorithmWatch gGmbH: Berlin, Germany, 2021.
61. The Public Voice. *Universal Guidelines for Artificial Intelligence*; The Public Voice: Burlington, VT, USA, 2018.
62. European Union. Summaries of EU Legislation; European Union: 2022. Available online: https://eur-lex.europa.eu/browse/summaries.html (accessed on 21 September 2023).
63. Braun, V.; Clarke, V. *Thematic Analysis*; American Psychological Association: Worcester, MA, USA, 2012.
64. Vaismoradi, M.; Snelgrove, S. Theme in qualitative content analysis and thematic analysis. *Forum Qual. Sozialforschung/Forum: Qual. Soc. Res.* **2019**, *20*. [CrossRef]
65. Mäntymäki, M.; Minkkinen, M.; Birkstedt, T.; Viljanen, M. Defining organizational AI governance. *AI Ethics* **2022**, *2*, 603–609. [CrossRef]
66. *Regulation 2021/0106*; Proposal for a Regulation of the Euopean Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Union: Brussels, Belgium, 2021.
67. *Regulation 2016/679*; Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). European Union: Brussels, Belgium, 2016.
68. *Regulation 2017/0003*; Proposal for a Regulation of the European Parliament and of the Council Concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications). European Union: Brussels, Belgium, 2017.
69. *Regulation 2022/0047*; Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act). European Union: Brussels, Belgium, 2022.
70. *Regulation 2020/0340*; Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act). European Union: Brussels, Belgium, 2020.
71. *Regulation 2021/0170*; Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and Repealing Council Directive

87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council. European Union: Brussels, Belgium, 2021.

72. *Directive 2022/0302*; Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products. European Union: Brussels, Belgium, 2022.
73. Almada, M.; Petit, N. The EU AI Act: Between Product Safety and Fundamental Rights. *Available SSRN* **2022**. [CrossRef]
74. *Directive 2022/0303*; Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive). European Union: Brussels, Belgium, 2022.
75. European Commission. The Digital Services Act Package. Available online: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (accessed on 29 September 2023).
76. *Regulation (EU) 2022/2065*; European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act) . European Union: Brussels, Belgium, 2022.
77. *Regulation 2020/0374*; Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act). European Union: Brussels, Belgium, 2020.
78. *Directive 2005/29/EC*; European Parliament and of the Council of 11 May 2005 Concerning Unfair Business-to-Consumer Commercial Practices in the Internal Market and Amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'). European Union: Brussels, Belgium, 2005.
79. *Charter 2000/C 364/01*; Charter of Fundamental Rights of the European Union (2000/C 364/01). European Union: Brussels, Belgium, 2000.
80. Kriebitz, A.; Lütge, C. Artificial intelligence and human rights: A business ethical assessment. *Bus. Hum. Rights J.* **2020**, *5*, 84–104. [CrossRef]
81. Seppälä, A.; Birkstedt, T.; Mäntymäki, M. From ethical AI principles to governed AI. In Proceedings of the 42nd International Conference on Information Systems (ICIS2021), Austin, TX, USA, 12–15 December 2021; pp. 1–17.
82. Hohma, E.; Boch, A.; Trauth, R. *Towards an Accountability Framework for Artificial Intelligence Systems*; TUM IEAI Whitepaper; TUM Institute for Ethics in Artificial Intelligence: Munich, Germany, 2022.
83. Anagnostou, M.; Karvounidou, O.; Katritzidaki, C.; Kechagia, C.; Melidou, K.; Mpeza, E.; Konstantinidis, I.; Kapantai, E.; Berberidis, C.; Magnisalis, I. Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics Inf. Technol.* **2022**, *24*, 37. [CrossRef]
84. Gefen, D. E-commerce: The role of familiarity and trust. *Omega* **2000**, *28*, 725–737. [CrossRef]
85. Stern, M.J.; Coleman, K.J. The multidimensionality of trust: Applications in collaborative natural resource management. *Soc. Nat. Resour.* **2015**, *28*, 117–132. [CrossRef]
86. Hohma, E.; Burnell, R.; Corrigan, C.C.; Luetge, C. Individuality and fairness in public health surveillance technology: A survey of user perceptions in contact tracing apps. *IEEE Trans. Technol. Soc.* **2022**, *3*, 300–306. [CrossRef]
87. Burton, J.W.; Stein, M.K.; Jensen, T.B. A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **2020**, *33*, 220–239. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

A.2 Supporting Material to Article 2

In the following, supporting material for Article 2 is provided. It contains (1) copyright information, (2) the statement of contribution and (3) the article as printed in the respective journal.

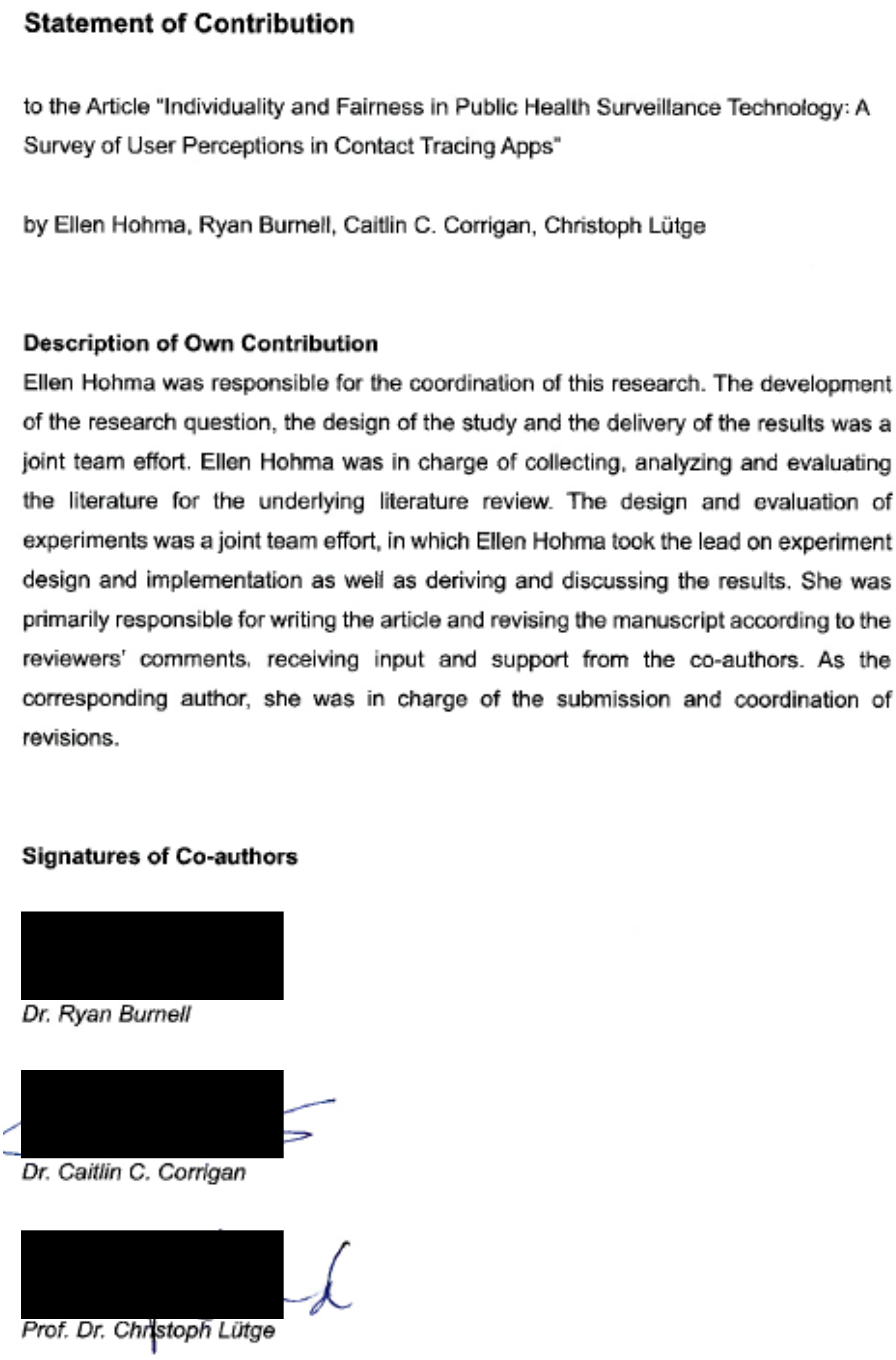
A.2.1 Copyright Information

This paper was published as

Hohma, E., Burnell, R., Corrigan, C. C. & Luetge, C. (2022). Individuality and Fairness in Public Health Surveillance Technology: A Survey of User Perceptions in Contact Tracing Apps. *IEEE Transactions on Technology and Society, 3(4),* 300-306. https://doi.org/10.1109/TTS.2022.3211073.

under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

A.2.2 Statement of Contribution



Statement of Contribution

to the Article "Individuality and Fairness in Public Health Surveillance Technology: A Survey of User Perceptions in Contact Tracing Apps"

by Ellen Hohma, Ryan Burnell, Caitlin C. Corrigan, Christoph Lütge

Description of Own Contribution

Ellen Hohma was responsible for the coordination of this research. The development of the research question, the design of the study and the delivery of the results was a joint team effort. Ellen Hohma was in charge of collecting, analyzing and evaluating the literature for the underlying literature review. The design and evaluation of experiments was a joint team effort, in which Ellen Hohma took the lead on experiment design and implementation as well as deriving and discussing the results. She was primarily responsible for writing the article and revising the manuscript according to the reviewers' comments, receiving input and support from the co-authors. As the corresponding author, she was in charge of the submission and coordination of revisions.

Signatures of Co-authors

[redacted]

Dr. Ryan Burnell

[redacted]

Dr. Caitlin C. Corrigan

[redacted]

Prof. Dr. Christoph Lütge

A.2.3 Full Article

Individuality and Fairness in Public Health Surveillance Technology: A Survey of User Perceptions in Contact Tracing Apps

Ellen Hohma, Ryan Burnell, Caitlin C. Corrigan, and Christoph Luetge

Abstract—**Machine learning algorithms are playing an increasingly important role in public health measures, accelerated by the Covid-19 pandemic. It is therefore vital that machine learning algorithms are applied in ways that are generally considered fair. However, the question of how to define fairness in a public health context is still an open one. In this study, we investigated people's attitudes towards two ways of defining fairness in the context of Covid-19 contact tracing apps. In the first, 'high-individuality' approach, the likelihood of an algorithm asking a person to self-isolate would depend on the person's individual characteristics, such as their risk of spreading the virus through regular contacts. In the second 'low individuality' approach, these individual characteristics would not be used to come to a decision. For each approach, participants rated its fairness, overall quality, and their privacy concerns, and answered questions about basic psychological need satisfaction. Participants rated the high-individuality approach as fairer and better overall compared to the low-individuality approach, despite having greater privacy concerns. Further, we found a strong correlation between the participants' fairness perceptions and their overall impression of the tracking tool. Together, these findings suggest that people prefer individualised approaches in some contexts and perceive them as fairer. However, policy makers should consider the privacy trade-off of employing such measures.**

Index Terms—**Algorithmic decision-making, contact tracing, data privacy, fairness, individuality, machine learning, public health surveillance.**

I. Introduction

THE OUTBREAK of the Covid-19 pandemic has highlighted the importance of continuous development and advancement of technologies to oppose health crises and to maintain public security and well-being. Multiple innovative approaches have quickly been developed in an effort to contain the spread of the virus, many equipped with Artificial Intelligence (AI) and Machine Learning (ML). In particular, the public health surveillance sector has seen a flood of tools that employ ML techniques to support public health monitoring, such as video surveillance for mask regulation compliance and fever or quarantine verification checks [1]. However, with the rapid development and deployment of these tools, ethical concerns about them have spread quickly, including concerns about whether they use individuals' data in ways that are fair.

In addition to the obvious ethical issues surrounding algorithmic fairness, the extent to which ML algorithms are fair might have practical consequences if it affects people's willingness to use AI-enabled tools or their acceptance of the outcomes recommended by the algorithms. For example, in the context of organisations, there is evidence for an interrelation between fairness concerns and overall satisfaction. Martinez-Tur et al. [2], for instance, showed that perceived distributive justice of gastronomy services (i.e., the perceived fairness of the outcome) was the primary determinant of customer satisfaction. Studying the effect of post-complaint behaviour, Blodgett et al. [3] concluded that although, in their case, distributive justice did not have an impact on complainant's satisfaction, the way in which the outcome was communicated could compensate unfair treatments. Similar evidence was further found in the relationship between organisations and employees. Sudin [4], for example, observed that distributive justice has a significant impact on overall employee satisfaction when studying performance appraisal processes. In the case of public health measures, it is therefore vital to ensure that people view ML algorithms as fair to improve the overall acceptability of such measures.

However, there are many challenges to ensuring that ML algorithms are considered fair. For instance, as ML algorithms are designed to predict outcomes based on input data, any biases in the input will lead to biased outcomes [5], [6]. In addition, the algorithmic design itself can produce biases because the underlying model chosen for a ML based system is a crucial factor for determining the outputs [7]. A further problem stems from the difficulty in defining "fairness" in different contexts. ML researchers have proposed a variety of fairness definitions that could be used to guide the design and evaluation of algorithms. But deciding which definition is best is not always easy. In particular, one critical issue for determining the right notion of fairness is the intangibility of the concept itself, even in anthropological and psychological studies [8]. Especially problematic is the fact that people's beliefs about what is fair differ depending on the context [9].

Manuscript received 31 January 2022; revised 17 May 2022 and 26 September 2022; accepted 26 September 2022. Date of publication 30 September 2022; date of current version 15 December 2022. This work was supported in part by the Technical University of Munich's Institute for Ethics in Artificial Intelligence (IEAI) and in part by the Leverhulme Centre for the Future of Intelligence. *(Corresponding author: Ellen Hohma.)*

Ellen Hohma, Caitlin C. Corrigan, and Christoph Luetge are with the Institute for Ethics in Artificial Intelligence, School of Social Sciences and Technology, Technical University of Munich, 80333 Munich, Germany (e-mail: ellen.hohma@tum.de; c.corrigan@tum.de; luetge@tum.de).

Ryan Burnell was with the Dyson School of Design Engineering, Imperial College London, SW7 2BX London, U.K. He is now with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, CB2 1TN Cambridge, U.K. (e-mail: rb967@cam.ac.uk).

Digital Object Identifier 10.1109/TTS.2022.3211073

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

One key part of algorithmic design that is closely related to people's perceptions of fairness is how an algorithm draws on personal information about individuals. Essentially, there are two contrary extremes. In 'high-individuality' approaches, algorithms make use of personal information about people to treat them according to their personal needs. By contrast, 'low-individuality' approaches consider people as a homogeneous group, treating everyone in that group the same. This has led ML researchers to categorise previously developed fairness concepts into Individual Fairness models aiming for similar predictions for similar individuals, Group Fairness models treating different groups equally, and Subgroup Fairness models—a combination of the former two—that categorize individuals based on their personal features into subgroups and ensuring group fairness constraints for those subgroups [6]. In many situations individual fairness models seem the most appropriate and fair. The pandemic in particular has brought forward many examples of how people would like to see more actions tailored specifically to their circumstances, such as vaccination status for contact restrictions. At the same time, concerns have been raised as to how much data can be justifiably requested, particularly when it contains sensitive information (e.g., debates around compulsory vaccination at the workplace). This shows that balancing different ethical principles can be challenging, but it is vital because these factors are linked to the users' uptake and acceptance of tools. Of course, in order for 'high-individuality' algorithms to treat people according to their needs, they sometimes require more extensive data on users. Using more data in the decision-making process might be seen as an invasion of privacy in certain situations [10], [11]. Some researchers, e.g., [12], even argue that fairness always comes at the cost of privacy. This trade-off between individuality and data privacy, often referred to as the personalisation-privacy paradox, has been frequently identified and studied in literature, e.g., [13], [14], [15], [16]. Still, even holding data collection equal, we might expect that the ways in which data are used might affect perceptions of fairness.

The distinction between high and low individuality approaches is highly linked to the debate on equity vs. equality in distribution decisions. The concept of equity is based on the equity theory by Adams [17] and refers to treating individuals according to their needs in a way to ensure equal outcomes (e.g., using affirmative action to assist disadvantaged groups). Equality, by contrast, involves treating all individuals the same–for example by giving everyone equal amounts of resources, even if this ultimately leads to inequality of outcomes. Views on which approach is fairer differ and depend greatly on the context, e.g., [18], [19], [20]. It is therefore vital that we understand whether people think high- or low-individuality approaches are fairer across different contexts.

In the context of public health measures, there are some data that speak to this issue. For example, Srivastava et al. [21] found that demographic parity (i.e., having the same probability for a positive outcome regardless of an individual's group membership) was most appealing to participants, suggesting a preference for low-individuality approaches. On the other hand, there is an emerging trend towards increasingly individualised medical treatments in the healthcare sector. Recent advancements in Individualised Medicine have made it possible to classify patients into subgroups based on their clinical characteristics instead of treating them as one homogenous group [22]. Taking this even further, Personalised Medicine–which involves practices such as analysing the patient's genome and resulting predictions on the patient's future health risks–has become a realistic possibility [23]. These approaches highlight the ongoing trend towards more individuality in data collection and processing in healthcare, as well as the resulting shift towards a rising focus on need-based treatment. In general, patients appear to support and value these personalised approaches, as they emphasize the uniqueness of each medical case and the corresponding individuality required to provide appropriate treatment [24].

But it remains unclear whether people feel the same when it comes to health control measures such as those being implemented in response to the Covid-19 pandemic. For example, many countries were using contact-tracing apps that rely on algorithms to decide who was required to self-isolate based on their contact with Covid-positive individuals. For such apps to work effectively, they need mass adoption among the population [25]. However, adherence to these apps tended to be low. Among others, Walrave et al. [26] have studied factors that influence people's adoption intention for using contact tracing apps. They found that fewer than 50% of their surveyed participants intended to use these apps, with app-related privacy concerns being one factor that negatively influenced users' intentions [26]. However, it remains unclear whether people consider it fair to use highly individualised approaches to decide who needs to isolate, or if they prefer low-individuality approaches that use fewer personal data and treat people more uniformly. We address this gap in the study reported here.

In addition, it is important to consider why people think particular application approaches are fairer than others. One factor that might play a role can be found in self-determination and its link to basic psychological need satisfaction. According to the Self-Determination Theory, an individual's motivation and engagement can be stimulated through three basic psychological needs: autonomy, competence and relatedness [27]. Autonomy reflects the extent to which individuals feel they are acting according to their own volition, willingness and choice. Competence reflects feelings of effectiveness and the capability of achieving important goals. Finally, Relatedness captures the feeling of being connected to and cared for by others. These basic needs have been associated with perceptions of fairness in organisational contexts. For instance, Olafsen et al. [28], found that employees' basic need satisfaction ratings were related to their judgments of the extent to which companies' payment distribution procedures are fair. Haar and Spell [29] found evidence for job autonomy to directly influence job satisfaction and, at the same time, moderate the relation between distributive justice and job satisfaction. Similar results were reported by Aryee et al. [30], who found a significant influence of justice on need satisfaction, which in turn was positively associated with intrinsic motivation. These findings indicate that fairness can promote basic psychological need satisfaction. In the context of public health surveillance, satisfying the three basic needs, and hence, stimulating intrinsic motivation, could encourage uptake of such tools and, ultimately, increase their overall effectiveness. Therefore, we sought to determine the relationship between perceptions of fairness and psychological need satisfaction.

To do so, we focused on public health instruments that were developed during the Covid-19 crisis, with the specific use case of contact tracing apps. We investigated how incorporating more individuality to a decision-making process–in this case, the risk of spreading the virus a person poses to others–affects people's perceptions of fairness and quality of such tools. We also investigated how these fairness perceptions related to basic psychological need satisfaction and frustration.

II. Research Method

An online vignette study using a within-subject design was conducted. To test our hypotheses on a prominent and real-world example from public health surveillance, contact tracing technology was chosen as a use case. To test the two extreme approaches, 'high-individuality' and 'low-individuality', we derived two policies on how contact tracing applications could use personalised data to determine who should be asked to self-isolate. Since contact-tracing apps were developed in various ways in different countries during the pandemic, people have divergent previous experiences with such tools according to their origin and place of residence. Because of this, we collected data from two countries: the U.K. and Germany. The two cases were selected based on the feasible access of the researchers to survey subjects in these countries, the similar design and policies surrounding the two countries' national contact tracing apps.

This research received ethical approval from Imperial College London's Research Governance and Integrity Team.

A. Participants

Participants were recruited through the Prolific survey platform–any participants who lived in the U.K. or Germany and who speak English were invited to participate. Not knowing how big the effect of our manipulation would be, we aimed to collect data from 150 participants from each country. Participants were only excluded if they failed one or both of the attention checks in the survey (n = 31). In total, 273 participants were considered in the analysis, 129 from the U.K. and 144 from Germany. The mean age of the participants was 28.01 (SD = 8.43). In terms of their highest level of education, 123 participants had finished secondary school, 92 held a Bachelor's, 38 held a Master's and 20 held a Doctoral degree.[1]

B. Procedure

The survey was conducted in English for all participants. Participants read and rated, in a random order, two different approaches for how contact tracing apps could determine who should be required to self-isolate following contact with infected individuals. In the "high-individuality" approach, the algorithm would consider the risk of a person spreading the virus to others in deciding whether to send that person to self-isolation. By contrast, in the "low-individuality" approach, the algorithm would not consider the risk of spreading the virus in its decision.

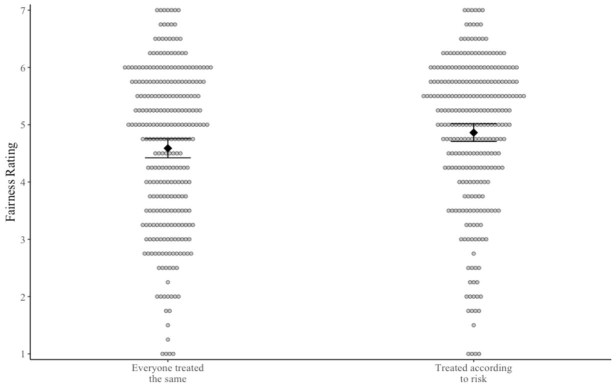


Fig. 1. Distribution of averaged fairness perception per participant for the low individuality approach (left) and high individuality approach (right).

Because the primary purpose of this study was to investigate how the *usage* of data affects perceived fairness, we held constant the *amount* of data collected across conditions. Therefore, both approaches mentioned that the app collects on location and contacts with others—the only difference was whether these data would be used to make decisions about self-isolation. For each algorithmic approach, participants were asked to rate how fair they perceived it, to what extent they had privacy concerns about the approach, whether they would be in general satisfied with such an approach, and how much the policy supported or frustrated their feelings of autonomy, competence, and relatedness. Measurement items for perceived fairness in the algorithmic decision-making process were based on Wang et al. [31], perceived privacy on Roca et al. [32], and adapted to this study's context. Need satisfaction and frustration was measured by adapting items from Peters et al. [33]. The full wording of the items is displayed in Table I. After completing these questions, participants were asked to provide demographic data, as well as answer some control variables as proposed by Wang et al. [31]. The survey ended with questions on the real Covid-19 contact tracing apps of the respective countries (NHS Covid-19 Tracing App for U.K., Corona-Warn-App for Germany), but because these data are not central to our research question, we do not report them here.

III. Results

To test whether the level of individuality in decision-making affects fairness perceptions, we compared the mean fairness ratings for the two proposed app approaches. Fig. 1 shows the distribution of the averaged fairness ratings for each participant for the low individuality approach (left) and high individuality approach (right). We found that participants rated the high individuality approach as significantly slightly fairer than the

[1] We found no relationship between people's age or education and their ratings of the approaches.

TABLE I
OVERVIEW OF THE ITEMS USED TO MEASURE STUDIED PERCEPTIONS

	Items
Fairness	In general, the above approach is fair The above approach is fair to people who are asked to self-isolate The above approach is fair to people who are at risk of serious consequences of COVID-19 The above approach will lead the app to make fair decisions about who to send to self-isolation
Privacy	I am concerned that an app following the above approach will use my personal information for other purposes without my authorization I think that too much of my personal information will be collected by an app following the above approach I am concerned about the privacy of my personal information when using an app following the above approach I am concerned that personal information about me collected by an app following the above approach will be shared with other entities without my authorization
Autonomy	
Satisfaction	The above approach would give people a sense of choice and freedom in the things they undertake The above approach would allow people's decisions to reflect what they really want
Frustration	The above approach would get in the way of people's ability to make their own choices The above approach would be controlling
Competence	
Satisfaction	The above approach is easy to understand The above approach would be easy to implement
Frustration	The above approach would get in the way of people's ability to achieve their goals The above approach would reduce people's confidence in their abilities
Relatedness	
Satisfaction	The above approach would help people feel close to people they care about The above approach would help people feel connected to other people
Frustration	The above approach would make it harder for people to connect in a meaningful way The above approach would make people feel excluded from groups they want to belong to

TABLE II
DIFFERENCES IN NEED SATISFACTION AND FRUSTRATION BY CONDITION

		M_{diff}	95% CI
Autonomy	Satisfaction	0.13	[-0.05, 0.30]
	Frustration	-0.01	[-0.16, 0.15]
Competence	Satisfaction	-0.44***	[-0.27, -0.61]
	Frustration	-0.13*	[-0.01, -0.25]
Relatedness	Satisfaction	0.06	[-0.09, 0.20]
	Frustration	0.07	[-0.07, 0.21]

Note: ***indicates $p < 0.001$. *indicates $p < 0.01$. Positive mean differences represent higher ratings for the high individuality group.

TABLE III
BETA WEIGHTS FROM REGRESSIONS PREDICTING OUTCOMES

		Fairness Perception	Overall Impression
Autonomy	Satisfaction	0.18	0.2***
	Frustration	0.08	-0.01
Competence	Satisfaction	0.25***	0.30***
	Frustration	-0.22***	-0.22***
Relatedness	Satisfaction	0.18***	0.19***
	Frustration	0.03	-0.02
Privacy Concerns		-0.02	-0.10***

Note: ***indicates $p < 0.001$.

low individuality approach, $M_{diff} = 0.27$, 95% CI [0.04, 0.51], which supports the hypothesis that increasing the degree of information individuality improves perceptions of fairness.

Moreover, individuality impacted participants' overall evaluations of the approaches—they rated the high individuality approach as significantly better overall than the low individuality approach, $M_{diff} = 0.23$, 95% CI [0.03, 0.43].

We also found that, within each condition, there was a strong correlation between participants' ratings of fairness and their overall evaluations of the approach, $r_{LowIndividuality}(271) = 0.74$, 95% CI [0.68, 0.79], $r_{HighIndividuality}(271) = 0.75$, 95% CI [0.69, 0.79], supporting our hypothesis that positive perceptions of overall fairness can increase satisfaction with public health technologies.

Next, we examined the impact of our manipulation on basic psychological need satisfaction and frustration. As Table II shows, we found no significant differences between the approaches in terms of autonomy satisfaction or frustration, nor in relatedness satisfaction or frustration. However, we found that the low individuality approach was rated significantly higher on both competence satisfaction and frustration.

Next, we conducted a linear regression with perceived fairness and overall impression as the dependent measures to examine their relationship with basic psychological needs. Table III presents the results of this regression. We found similar patterns for both dependant variables. For basic need satisfaction, competence satisfaction and relatedness satisfaction were both significantly, positively related to fairness and overall user impression. Autonomy satisfaction was only significantly, positively related to the user's overall impression. Investigating its counterpart basic need frustration revealed that competence frustration was negatively related to fairness as well as overall user impression. No evidence was found that autonomy frustration or relatedness frustration are related to perceptions of fairness or overall ratings. In other words, our results support the hypothesis that positive perceptions of fairness can stimulate basic psychological need satisfaction, although only for competence and relatedness.

Finally, we studied the effect of perceived data privacy on the participants' perceptions of fairness and overall satisfaction with the application. We found that people reported slightly greater privacy concerns for the high individuality approach compared to the low individuality approach, $M_{diff} = 0.15$, 95% CI [0.06, 0.24]. We also found a significant, negative effect of privacy concerns on the overall user impression in both conditions. However, we found no evidence that privacy concerns are related to perceptions of fairness in our study.

These results suggest that perceived data privacy is related to evaluations of the proposed tools.

We found no differences in perceptions of fairness, overall rating, or privacy concerns between the U.K. and Germany (*p*s $>.41$). Therefore, we combined the data from these two countries for the main analyses.

IV. Discussion

The study's aim was to examine the extent to which people think two different approaches to ML-based public health surveillance technologies are fair. In our study, we found evidence that participants preferred high-individuality approaches to contact-tracing—participants rated the high-individuality approach as both fairer and better overall. Moreover, we found a strong correlation between participants' fairness perceptions and their overall impression, suggesting that perceptions of fairness are tightly linked with people's evaluation of public health tools. However, we did not find evidence that need satisfaction can explain these effects.

A. Ethical Implications

Issues of justice and fairness have been emphasized repeatedly in ethical frameworks for healthcare and AI-based tools [34]. Accelerated by the spread of the Covid-19 pandemic, recent literature has identified numerous instances of bias and unfairness in public health surveillance due to the fact that these technologies collect and analyse large amounts of data, often including socio-economic information such as race, ethnicity, gender or political affiliation [35]. For example, biased data collection strategies can result in subgroups not being visible or being stigmatised as they lack the needed technical devices [36], mobile communication or Internet access [37], [38]. The inevitable trade-off between individual freedom and civil obligations necessitates a delicate balance between collecting all the information needed to best protect public benefits while avoiding discrimination of certain populations [35]. Essentially, this leads to the dilemma of how far we can limit personal freedoms for the public benefit that has driven many controversies during the Covid-19 pandemic.

Although we did not explicitly study individuality with regard to demographic characteristics, our findings suggest that users, at least in the U.K. and Germany, value a more personal treatment based on their individual characteristics in health surveillance applications. While people feel discriminated when judgement is based on their demographic attributes, they seem to likewise feel treated in an unfair way if they are regarded as a fully anonymous, homogeneous group. Clearly, more work is therefore needed to determine the individualized uses of data that people see as discriminatory and those that contribute to positive perceptions of fairness. While in general, it is likely that individual and unchangeable traits, such as gender or race, might be counted among the discriminatory ones, our findings suggest that personal parameters resulting from an individual's actions instead of traits—such as in the context of this study the risk of spreading the virus that a person poses to others—might be among those features where disclosure is accepted to enable a fairer decision. However, the further and more concrete identification of such a distinction of features can foster the pursuit of a solution to the question of how to balance personal rights against the well-being of the broader society.

In this study, we found that people preferred the highly individualized approach despite reporting greater privacy concerns regarding the use of data with that approach. This finding indicates that, in certain contexts, people might consider some invasions of privacy or limitations of freedoms fair and acceptable, at least to the extent that they are important for public health. Of course, in this regard, context is crucial. Nissenbaum [39] argues that contextual integrity is the benchmark of privacy, and consent to the use of data is only given in relation to its respective circumstances. Empirical field studies and scenario-based surveys, such as [40] or [41], support this notion. Perhaps, then, our participants were willing to sacrifice some data privacy because they viewed the high-individualization approach as fairer. The circumstances under which privacy is seen as an acceptable trade-off for fairness is worth of further investigation.

It is also worth thinking about how individual differences might affect people's preferences and perceptions of fairness. For example, individualistic persons or cultures put a higher focus on personal autonomy and self-fulfilment and base identity on themselves as well as their personal achievements [42]. By contrast, collectivistic persons or cultures value group belonging and loyalty and derive beliefs from group decisions and the social system [42]. Studies that measured public acceptance of digital contact tracing applications during Covid-19 have found the acceptance rate to be nearly twice as high in collectivist countries such as China than in individualistic countries such as Germany [43]. We might also expect cultural and social norms to affect people's evaluations of fairness and preferences for individualized approaches. In particular, people who value individualism might be more likely than users who value collectivism to prefer high-individuality approaches to satisfy their 'personalization' demand.

B. Practical Implications

Trying to solve the issue on how to incorporate fairness in ML algorithms, researchers have already gathered and developed numerous definitions of fairness, e.g., [6], [44], [45] and translated them into several distinct fairness models, e.g., [5], [46], [47], [48], [49]. The ultimate goal is to translate intangible notions such as fairness to statistically measurable features and probability equations. To achieve this goal, we need theoretical and empirical work that investigates what people consider fair in different contexts. More broadly, we need methods to concretely define, optimise, and evaluate fairness algorithms. In an effort to ease the model selection, researchers have categorised the identified definitions along their degree of personalisation, into individual, subgroup and group fairness models, e.g., [6], [7], [45]. Individual fairness models compare features of individuals under investigation to ensure that individuals with similar feature scores obtain similar predictions, whereas group fairness models cluster individuals into groups and ensure certain statistical paradigms between the groups.

Subgroup fairness models form a combination of the former two categories, categorizing individuals based on their personal features into subgroups and ensuring group fairness constraints for those subgroups [6]. Taking this classification as a basis, the degree of individuality is crucial for examining the various notions of fairness and needs to be weighed to determine which fairness model should be chosen for a specific AI-enabled technology. However, deciding which fairness model category to draw from almost always require an in-depth understanding of the specific context. In the context of public-health surveillance, our data suggest that people indeed valued some degree of individuality in the decision even at the expense of data privacy. This means that, in the field of public health surveillance and the context of our study design, our findings suggest that some individuality is desired over complete homogeneity, pointing towards models like "Fairness through Awareness" [50] that allow for a greater consideration of individual personal characteristics.

Of course, this leads to the question of how such a balance between data privacy and fairness perceptions can be ensured. We suggest that it can be targeted with a clearer classification of attributes into those that users consider as purely privacy-intrusive or those that are perceived to contribute to enhancing fairness. For this study, we chose to examine a scenario in which it was not clear which approach people would view as fairer. Preferences regarding individualised decisions would probably look different if we selected inherent traits as personalisation factors, such as gender or social status. The fact that the chosen attributes are derived from people's actions or decisions (i.e., characteristics that can be more consciously and more easily influenced) might make these more acceptable factors for individualization. Therefore, when separating parameters into those that are discriminatory and those that are acceptable for algorithmic decision-making, it is important to consider the type of individualisation and the attribute's specific nature.

C. Limitations and Future Research

Although we found empirical evidence for a preference towards more individuality in public health surveillance tools, there was considerable overlap in the distributions of people's responses across the two approaches. One explanation for this small effect is that the distinction between the individuality approaches was not stressed precisely enough in our experiment. Another possibility is that the selected feature, a person's risk of spreading the virus, was not perceived as sufficiently individual to substantially impact fairness perceptions. Future work should examine how other factors affect fairness perceptions in public health contexts.

Although the user's perceived data privacy did not predict perceived fairness, we found evidence that it might still affect the user's satisfaction with the application. While we interpreted this as indication that the way data are used in decision-making can be important for perceptions of data privacy, it is possible participants did not fully understand that the data collected was the same across the two approaches. Future research should take this into account when developing similar experiments, as studying the relative and cumulative effects of data collection and data use could help inform policy decisions.

Furthermore, widening the focus of this study in future research to include people from a broader range of cultural backgrounds and to examine other public health measures could complement the picture to a more holistic overview.

V. Conclusion

In this study, we investigated the relation between the individuality of a ML-based public health surveillance method and the perceived fairness as well as overall impression of that tool on the example of contact tracing applications.

Our findings suggest that users (in the U.K. and Germany) value higher degrees of individuality in health surveillance related decisions and perceive 'high-individuality' contact tracing app versions as fairer and more satisfactory overall. This pattern held despite the fact that people viewed higher levels of individuality as more privacy intrusive. Moreover, our findings suggest that perceptions of fairness are important for people's evaluations of public health surveillance tools and could affect people's adoption and acceptance of those applications.

Our results support the general trend towards more personalisation in healthcare also in health surveillance technologies and inform the design of future ML-enabled public health surveillance tools. While more individuality seemed more appealing for participants in our study, the nature of attributes that are used within a decision seems to be crucial for fairness perceptions, pointing towards a greater need for research to distinguish the parameters considered as fair or discriminating.

References

[1] F. Piccialli, V. S. di Cola, F. Giampaolo, and S. Cuomo, "The role of artificial intelligence in fighting the COVID-19 pandemic," *Inf. Syst. Front.*, vol. 23, pp. 1467–1497, Apr. 2021.

[2] V. Martinez-Tur, J. M. Peiro, J. Ramos, and C. Moliner, "Justice perceptions as predictors of customer satisfaction: The impact of distributive, procedural, and interactional justice," *J. Appl. Soc. Psychol.*, vol. 36, no. 1, pp. 100–119, 2006.

[3] J. G. Blodgett, D. J. Hill, and S. S. Tax, "The effects of distributive, procedural, and interactional justice on postcomplaint behavior," *J. Retailing*, vol. 73, no. 2, pp. 185–210, 1997.

[4] S. Sudin, "Fairness of and satisfaction with performance appraisal process," *J. Global Manag.*, vol. 2, no. 1, pp. 66–83, 2011.

[5] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," 2018, *arXiv:1810.08810*.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," 2019, *arXiv:1908.09635*.

[7] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI: Challenges and opportunities," *Bus. Inf. Syst. Eng.*, vol. 62, no. 4, pp. 379–384, 2020.

[8] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proc. World Wide Web Conf.*, 2018, pp. 903–912.

[9] D. Shin, "User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability," *J. Broadcast. Electron. Media*, vol. 64, no. 4, pp. 541–565, 2020.

[10] C. Top and B. J. Ali, "Customer satisfaction in online meeting platforms: Impact of efficiency, fulfillment, system availability, and privacy," *Amazonia Investiga*, vol. 10, no. 38, pp. 70–81, 2021.

[11] N. D. Nayeri and M. Aghajani, "Patients' privacy and satisfaction in the emergency department: A descriptive analytical study," *Nursing Ethics*, vol. 17, no. 2, pp. 167–177, 2010.

[12] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *Proc. IEEE Eur. Symp. Security Privacy (EuroSP)*, 2021, pp. 292–303.

[13] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon," *Comput. Security*, vol. 64, pp. 122–134, Jan. 2017.

[14] N. Gerber, P. Gerber, and M. Volkamer, "Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior," *Comput. Security*, vol. 77, pp. 226–261, Aug. 2018.

[15] S. Barth and M. D. T. De Jong, "The privacy paradox—Investigating discrepancies between expressed privacy concerns and actual online behavior—A systematic literature review," *Telematics Inform.*, vol. 34, no. 7, pp. 1038–1058, 2017.

[16] E. Aguirre, A. L. Roggeveen, D. Grewal, and M. Wetzels, "The personalization-privacy paradox: Implications for new media," *J. Consum. Market.*, vol. 33, no. 2, pp. 98–110, 2016.

[17] J. S. Adams, "Inequity in social exchange," in *Advances in Experimental Social Psychology*, vol. 2. Amsterdam, The Netherlands: Elsevier, 1965, pp. 267–299.

[18] K. S. Cook and K. A. Hegtvedt, "Distributive justice, equity, and equality," *Annu. Rev. Sociol.*, vol. 9, no. 1, pp. 217–241, 1983.

[19] B. Kabanoff, "Equity, equality, power, and conflict," *Acad. Manag. Rev.*, vol. 16, no. 2, pp. 416–441, 1991.

[20] M. Deutsch, "Equity, equality, and need: What determines which value will be used as the basis of distributive justice?" *J. Soc. Issues*, vol. 31, no. 3, pp. 137–149, 1975.

[21] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2019, pp. 2459–2468.

[22] B. Hüsing, "Individualisierte Medizin–Potenziale und Handlungsbedarf," *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 104, no. 10, pp. 727–731, 2010.

[23] A. Agusti, "The path to personalised medicine in COPD," *Thorax*, vol. 69, no. 9, pp. 857–864, 2014.

[24] J. P. Richardson *et al.*, "Patient apprehensions about the use of artificial intelligence in healthcare," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–6, 2021.

[25] S. Trang, M. Trenz, W. H. Weiger, M. Tarafdar, and C. M. K. Cheung, "One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps," *Eur. J. Inf. Syst.*, vol. 29, no. 4, pp. 415–428, 2020.

[26] M. Walrave, C. Waeterloos, and K. Ponnet, "Ready or not for contact tracing? Investigating the adoption intention of COVID-19 contact-tracing technology using an extended unified theory of acceptance and use of technology model," *Cyberpsychol. Behav. Soc. Netw.*, vol. 24, no. 6, pp. 377–383, 2021.

[27] E. Deci and R. Ryan, *Intrinsic Motivation and Self Determination in Human Behavior*. New York, NY, USA: Plenum, 1985.

[28] A. H. Olafsen, H. Halvari, J. Forest, and E. L. Deci, "Show them the money? The role of pay, managerial need support, and justice in a self-determination theory model of intrinsic work motivation," *Scandinavian J. Psychol.*, vol. 56, no. 4, pp. 447–457, 2015.

[29] J. M. Haar and C. S. Spell, "How does distributive justice affect work attitudes? The moderating effects of autonomy," *Int. J. Human Resour. Manag.*, vol. 20, no. 8, pp. 1827–1842, 2009.

[30] S. Aryee, F. O. Walumbwa, R. Mondejar, and C. W. L. Chu, "Accounting for the influence of overall justice on job performance: Integrating self-determination and social exchange theories," *J. Manag. Stud.*, vol. 52, no. 2, pp. 231–252, 2015.

[31] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–14.

[32] J. C. Roca, J. J. García, and J. J. De La Vega, "The importance of perceived trust, security and privacy in online trading systems," *Inf. Manag. Comput. Security*, vol. 17, no. 2, pp. 96–113, 2009.

[33] D. Peters, R. A. Calvo, and R. M. Ryan, "Designing for motivation, engagement and wellbeing in digital experience," *Front. Psychol.*, vol. 9, p. 797, May 2018.

[34] L. Floridi *et al.*, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018.

[35] U. Gasser, M. Ienca, J. Scheibner, J. Sleigh, and E. Vayena, "Digital tools against COVID-19: Taxonomy, ethical challenges, and navigation aid," *Lancet Digit. Health*, vol. 2, no. 8, pp. E425–E434, 2020.

[36] M. M. Mello and C. J. Wang, "Ethics and governance for digital disease surveillance," *Science*, vol. 368, no. 6494, pp. 951–954, 2020.

[37] C. Klingler, D. S. Silva, C. Schuermann, A. A. Reis, A. Saxena, and D. Strech, "Ethical issues in public health surveillance: A systematic qualitative review," *BMC Public Health*, vol. 17, no. 1, p. 295, 2017.

[38] J. Budd *et al.*, "Digital technologies in the public-health response to COVID-19," *Nat. Med.*, vol. 26, pp. 1183–1192, Aug. 2020.

[39] H. Nissenbaum, "A contextual approach to privacy online," *Daedalus*, vol. 140, no. 4, pp. 32–48, 2011.

[40] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov, "Android permissions remystified: A field study on contextual integrity," in *Proc. 24th USENIX Security Symp. (USENIX Security)*, 2015, pp. 499–514.

[41] Y. Wang, H. Xia, and Y. Huang, "Examining American and Chinese Internet users' contextual privacy preferences of behavioral advertising," in *Proc. 19th ACM Conf. Comput. Supported Cooperative Work Soc. Comput.*, 2016, pp. 539–552.

[42] G. Hofstede, "Motivation, leadership, and organization: Do American theories apply abroad?" *Org. Dyn.*, vol. 9, no. 1, pp. 42–63, 1980.

[43] M. V. Zetterholm, Y. Lin, and P. Jokela, "Digital contact tracing applications during COVID-19: A scoping review about public acceptance," *Informatics*, vol. 8, no. 3, p. 48, 2021.

[44] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proc. Conf. Fairness Accountability Transparency*, 2018, pp. 149–159.

[45] B. Hutchinson and M. Mitchell, "50 years of test (Un)fairness: Lessons for machine learning," in *Proc. Conf. Fairness Accountability Transparency*, 2019, pp. 49–58.

[46] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2017, pp. 797–806.

[47] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions," 2018, *arXiv:1811.07867*.

[48] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How do fairness definitions fare?: Examining public attitudes towards algorithmic definitions of fairness," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 99–106.

[49] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM Int. Workshop Softw. Fairness (FairWare)*, 2018, pp. 1–7.

[50] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.

A.3 Supporting Material to Article 3

In the following, supporting material for Article 3 is provided. It contains (1) copyright information, (2) the statement of contribution and (3) the article as printed in the respective journal.

A.3.1 Copyright Information

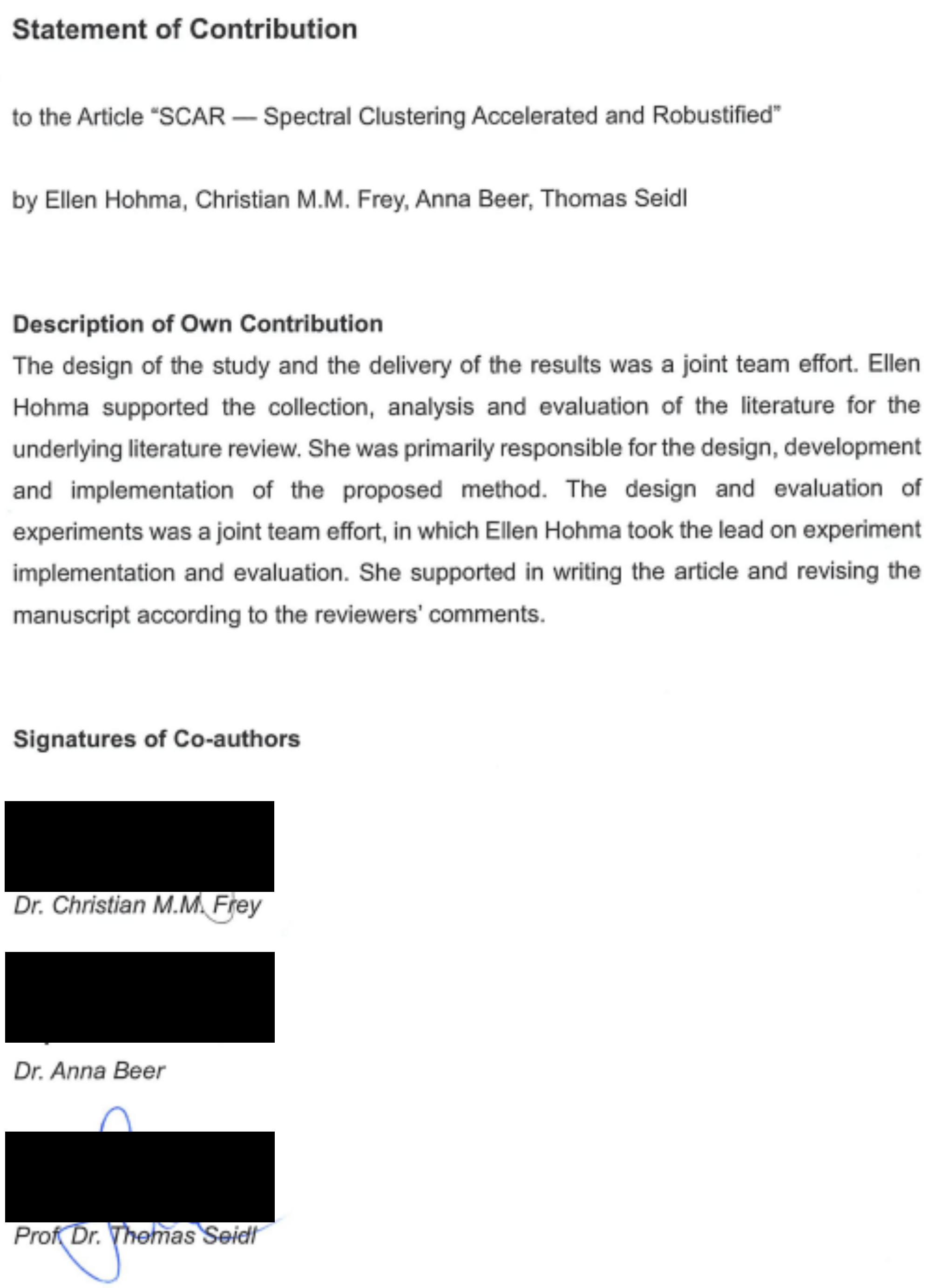
This paper was published as

Hohma, E., Frey, C. M.M., Beer, A. & Seidl, T. (2022). SCAR – Spectral Clustering Accelerated and Robustified. *PVLDB, 15(11)*, 3031-3044. https://doi.org/10.14778/3551793.3551850

under the terms and conditions of the Creative Commons BY-NC-ND 4.0 International license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

The included copy has not been changed or modified (e.g. remixed, transformed or built upon).

A.3.2 Statement of Contribution



Statement of Contribution

to the Article "SCAR — Spectral Clustering Accelerated and Robustified"

by Ellen Hohma, Christian M.M. Frey, Anna Beer, Thomas Seidl

Description of Own Contribution

The design of the study and the delivery of the results was a joint team effort. Ellen Hohma supported the collection, analysis and evaluation of the literature for the underlying literature review. She was primarily responsible for the design, development and implementation of the proposed method. The design and evaluation of experiments was a joint team effort, in which Ellen Hohma took the lead on experiment implementation and evaluation. She supported in writing the article and revising the manuscript according to the reviewers' comments.

Signatures of Co-authors

[redacted]

Dr. Christian M.M. Frey

[redacted]

Dr. Anna Beer

[redacted]

Prof. Dr. Thomas Seidl

A.3.3 Full Article

Artifacts Available
acm
V1.1

SCAR — Spectral Clustering Accelerated and Robustified

Ellen Hohma*
Technical University of Munich
Munich, Germany
ellen.hohma@tum.de

Christian M.M. Frey*
Christian-Albrecht University of Kiel
Kiel, Germany
cfr@informatik.uni-kiel.de

Anna Beer*
Aarhus University
Aarhus, Denmark
beer@cs.au.dk

Thomas Seidl
LMU Munich
Munich, Germany
seidl@dbs.ifi.lmu.de

ABSTRACT

Spectral clustering is one of the most advantageous clustering approaches. However, standard Spectral Clustering is sensitive to noisy input data and has a high runtime complexity. Tackling one of these problems often exacerbates the other. As real-world datasets are often large *and* compromised by noise, we need to improve both robustness and runtime at once. Thus, we propose **S**pectral **C**lustering - **A**ccelerated and **R**obust (SCAR), an accelerated, robustified spectral clustering method. In an iterative approach, we achieve robustness by separating the data into two latent components: cleansed and noisy data. We accelerate the eigendecomposition – the most time-consuming step – based on the Nyström method. We compare SCAR to related recent state-of-the-art algorithms in extensive experiments. SCAR surpasses its competitors in terms of speed and clustering quality on highly noisy data.

PVLDB Reference Format:
Ellen Hohma, Christian M.M. Frey, Anna Beer, and Thomas Seidl. SCAR — Spectral Clustering Accelerated and Robustified. PVLDB, 15(11): 3031 - 3044, 2022.
doi:10.14778/3551793.3551850

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/SpectralClusteringAcceleratedRobust/SCAR.

1 INTRODUCTION

Clustering is a fundamental data mining task needed in virtually all areas working with data and also serves as an unsupervised preprocessing step for a plethora of subsequent tasks. One of the most favorable clustering methods is spectral clustering: it is applicable to non-numeric datasets, can find clusters of complex shapes and different densities, and optimizes a mathematically well-defined problem [52]. However, real-world datasets are challenging for several reasons: with newly developed data gathering methods (e.g., in medicine, chemistry, or biology), in recent years datasets grew in dimensionality as well as in size. The runtime complexity of spectral clustering methods is only linear in the number of dimensions, as it works on an affinity graph of the data, making it superior to more traditional clustering methods when working on high-dimensional data. However, the runtime complexity (for the naive implementation) of $O(n^3)$ w.r.t. the number of data points is comparably large. Furthermore, real-world data often contains noise that is neither handled well by standard spectral clustering methods nor by other clustering methods. Clustering noisy data is in fact a very challenging task, as Fig. 1 illustrates. It shows a very noisy version of the well-known synthetic moons dataset as clustered by diverse algorithms. It may not come as a surprise that standard Spectral Clustering fails to detect the moons correctly. But also state-of-the-art algorithms that are designed specifically to be robust against noise can only handle noise up to a certain degree and were not able to detect the two clusters correctly. Our competitors in Fig. 1 (as well as in our experiments in Sec. 5) are recent clustering algorithms published at high-quality conferences: RSC [7], DCF [50], and SpectACl [21]. The authors of all methods performed extensive experiments showing their superiority against a variety of other clustering methods regarding noise robustness. RSC and DCF also successfully tackle the efficiency problems of clustering. Nevertheless, with our newly developed method SCAR (**S**pectral **C**lustering - **A**ccelerated and **R**obust), we found a way to even further improve both, clustering quality on highly noisy data *and* efficiency on high-dimensional data.

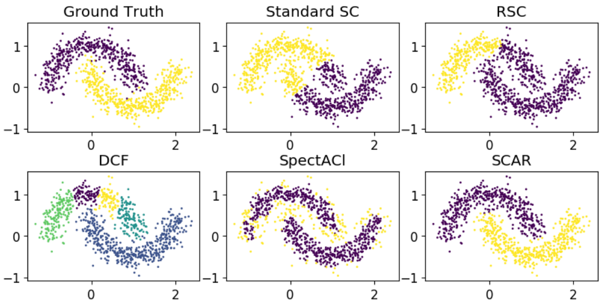


Figure 1: Our method SCAR vs. state-of-the art related clustering algorithms on the moons dataset with $noise = 0.15$.

SCAR uses weighted kNN graphs to capture highly complex structures in the data implying clusters of non-convex shapes. For a good segmentation of the graph, normalized cuts have proven to be

[*] Authors contributed equally to this work.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 15, No. 11 ISSN 2150-8097.
doi:10.14778/3551793.3551850

desirable [10, 12], suggesting a spectral approach. Based on the concept of RSC [7], we divide the data into a subset containing noise and a subset containing the relevant information for clustering. However, RSC involves the frequent calculation of eigendecompositions in an iterative approach, which we accelerate with the Nyström method. With an elaborated combination of synergistic methods and changes we manage to achieve highly competitive results regarding the clustering quality and robustness. In extensive and reproducible experiments we examine and compare our clustering results w.r.t. quality and runtime. SCAR shows the desired behavior for highly noisy datasets, where it outperformed recent state-of-the-art algorithms in quality, noise robustness, and runtime. We evaluated diverse types of noise and used well-known benchmark datasets. Our main contributions are as follows:

- We introduce SCAR, our novel spectral clustering method tackling both, robustness *and* speed.
- We incorporate the Nyström method to accelerate the eigendecomposition in robust spectral clustering.
- We further enhance quality and stability of clusterings
- We evaluate our method thoroughly, fairly, and reproduciblyand compare our method to recent state-of-the-art methods on the established real-world benchmark datasets.

Outline. In Sec. 2 we give an overview on related methods. In Sec. 3 we explain the basics for our new method. In Sec. 4 we introduce our new fast and robust spectral clustering method, called SCAR. In Sec. 5 we evaluate SCAR thoroughly, objectively, and reproducibly. Sec. 6 concludes this paper.

2 RELATED WORK

Spectral clustering refers to a set of clustering algorithms that partition a given dataset based on the spectrum of the datapoints' affinity matrix. They essentially follow three steps [52]: (1) construct a similarity graph $\mathcal{G}$, (2) compute the Laplacian of $\mathcal{G}$ and its eigendecomposition, and (3) cluster its eigenvectors with a standard clustering method, e.g., k-Means [36, 37]. Spectral clustering surpasses traditional clustering techniques in several aspects: e.g., they find arbitrarily shaped clusters, are applicable on categorical data, solve a clearly defined mathematical goal [52], and can handle varying densities. However, spectral clustering is noise-sensitive [7, 21] and has a relatively high runtime. In the following, we provide an overview of related works in the research field.

2.1 Improving Runtime

Most recent advances improving any of the steps of spectral clustering can be found in [51]. In the following, we focus on approaches accelerating the most time-intensive step of spectral clustering, the eigendecomposition. The acceleration is usually achieved with one of two strategies: iteration or sampling.

Iterative approaches. The probably most common method to accelerate the computation of eigenvectors and eigenvalues of a matrix is the *power iteration.* By iteratively multiplying the matrix with a randomly initialized vector (or an estimation of the dominant eigenvector), the eigenvector belonging to the largest eigenvalue is approximated. Generally, the frequent matrix multiplications are expensive, and only the dominant eigenvector can be approximated with the original power method – for spectral clustering, the eigenvectors belonging to the *smallest* eigenvalues are of interest. Note, that the behavior of convergence of iterative approaches usually depends on the distribution and gaps between the eigenvalues [51]. There is a wealth of extensions based on the power iteration aimed at alleviating its downsides for spectral clustering. Using Krylov subspaces allows approximating several eigenvectors at once: E.g., the *Arnoldi iteration* [1] orthogonalizes the vectors spanning the Krylov subspace by applying the Gram-Schmidt process. For Hermitian matrices like symmetric Laplacians, which are used in the process of spectral clustering, the *Lanczos* method has been proposed [29]. The Lanczos method approximates the largest k eigenvectors in $O(|\mathcal{E}|k + |\mathcal{V}|k^2)$ for a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ [51]. It is used for spectral clustering in [46]. While the Lanczos method performs well even for sparse matrices, it is often prone to numerical instability [9]. The *Implicitly Restarted Lanczos Method* (IRLM) as used, e.g., in ARPACK [31], can reduce numerical instability. Further adaptions involve among others using the inverse matrix to get the smallest eigenvalues and respective eigenvectors [15]. The Krylov-Schur algorithm [47] alleviates additional problems emerging with very large Hermitian or non-Hermitian matrices. As the convergence of symmetric cases depends on the gap ratio of the eigenvalues [41], both ends of the spectrum are approximated, e.g., with *IRLM-BE.*

Sampling-based approaches. Sampling based approaches work on (1) a subset of the edges in the similarity graph or on (2) a subset of the nodes: (1) implies a sparser Laplacian than the original one, while (2) implies a Laplacian of lower dimensionality.

Approach (1) accelerates the eigendecomposition by leveraging matrix operations that are optimized towards sparse input matrices. Working on matrices defined by kNN graphs, choosing a small k leads to sparse matrices that still hold relevant information on the structure of the data. For general graphs, *spectral sparsification* can be applied. It approximates the graph Laplacian with a matrix of same size containing fewer sampled entries. The sampling process ensures that certain pre-defined properties are respected (cf. [51]).

Approach (2) includes *graph coarsening* methods (e.g., [20, 27, 51]) that reduce the original similarity graph to a coarser graph, leading to an adjacency matrix of significantly lower dimensionality. Computing the eigendecomposition on the lower-dimensional matrix and refining it afterwards leads to a significant acceleration.

In our approach introduced in Sec. 4, we use the sampling-based *Nyström method*, which is an effective method to significantly speed-up spectral clustering while maintaining good overall eigenvector accuracy (e.g., [16, 32, 53, 54]). The Nyström method has been analyzed, replicated and improved throughout multiple studies: [6], [13], [45], and [56] focus on the improvements of particular downsides, such as the partial loss of information by sampling landmark points. Furthermore, they provide theoretical evaluations and frameworks on how the quality of the resulting spectral embedding is affected by applying the Nyström approximation. In [43], the impact of the number of landmarks selected as subsample as well as the influence on the overall clustering accuracy is investigated. Thorough studies in [17], [38], and [28] show the impact of sampling techniques picked for identifying the base subset for the Nyström extension. A theoretical analysis of the algorithm's performance and derivations of error bounds are formulated in [11]. We explain the Nyström method in detail in Sec. 3.2.

2.2 Improving Noise Robustness

As spectral clustering has no inherent noise-handling, its quality can suffer from diverse types of noise that often occur in real-world data. In the following, we distinguish between four different notions of noise that are often mixed up in the literature or not clarified: (1) additional noise points, (2) jitter, (3) noisy features, and (4) noisy edges. Even though they are closely interrelated, they can imply different challenges for (spectral) clustering.

Additional noise points. The probably most common notion of noise is that there are additional noise points in the dataset. They are typically uniformly distributed (and iid) and do not belong to any cluster. E.g., NRSC [35] tackles such noise for spectral clustering by assigning all noise points to an extra cluster. However, they work on the fully connected graph and assume that the majority of edges connected to a noise point has a low weight. AHK [23] also tackles this kind of noise and simultaneously robustify spectral clustering regarding the parameter choice by using an aggregated heat kernel. CAHSM [34] use a hypergraph to compensate for outliers and noise.

Jitter. Adding noise to a dataset can also imply adding a small deviation to each point. E.g., noise adjustment for the moons datasets regulates the deviation from the "perfectly-shaped" moons. A similar effect can be achieved by data quantization. In [24], error bounds for spectral clustering on data with jitter, resp., *perturbed data* are evaluated. Robustness against this type of noise for spectral methods is evaluated, e.g., in SpectACL [21], and RSC [7].

Noisy features. Especially in high-dimensional data, we may encounter noisy features. These refer, for example, to uniformly distributed dimensions of the data that are irrelevant for clustering for at least some points. FWKE-SC [26], SSCG [18], [57], and [3] combine feature weighting with spectral clustering to tackle this problem (similarly to subspace clustering). As they mainly focus on the construction of the similarity matrix, they can be combined with our approach in future work.

Noisy edges. Noise in graphs can also occur as additional edges in the affinity graph of the data. RSC [7] (cf. Sections 2.3 and 3.3) focuses on removing edges that connect different clusters, which are also called *corrupted edges.* RSEC [49] regards noisy edges in the context of spectral ensemble clustering. In [4], noise is regarded as "an additive perturbation to the similarity matrix", including noisy edges as well as corrupted weights of existing edges.

In this paper, we focus on robustness w.r.t. noisy edges and jitter. For the other types of noise, we suggest to filter additional noise points in a preprocessing step. For noisy features, our approach can easily be combined with feature weighting approaches that adapt the initial affinity matrix, as SCAR builds on top of the affinity matrix. For weighting the importance of features, one can follow approaches like FWKE-SC [26], using the concept of knowledge entropy, or apply importance scores for attributes that adapt to every point individually, like KISS [5].

2.3 Comparative Methods

In our experiments in Sec. 5 we compare our newly developed method SCAR with standard Spectral Clustering (SC) [40] as well as high-quality state-of-the-art spectral methods that aim at robustness and efficiency: Robust Spectral Clustering (RSC) [7] and SpectACL [21]. Furthermore, we include the very recently introduced method DCF [50] into our analyses. DCF is not a spectral approach, but also aims at robustness and efficiency.

RSC jointly performs the standard Spectral Clustering and the decomposition of the adjacency matrix A. The latter is assumed to be an additive decomposition of two latent factors, a graph containing corrupted edges and a graph representing the noise-free data. As RSC outperforms basic clustering principles like k-Means and density-based clustering methods on noisy datasets [7], it serves as a baseline in our evaluation in Sec. 5.

SpectACL combines approaches from spectral clustering and DBSCAN to solve their major drawbacks regarding noise sensitivity for minimum cut clustering and varying densities for density-based clustering [21]. The core idea is to determine clusters with large average densities while optimizing the density parameters using the spectrum of the weighted adjacency matrix.

DCF aims at improving peak-finding techniques for density-based clustering, which determine groups in a dataset based on their high density as well as distances to clusters of higher density [50]. The approach applies the peak-finding criterion to determine cluster cores instead of point modes, which enables the detection of clusters with varying densities.

3 PRELIMINARIES

In the following we give some preliminary basics for our method SCAR. In Sec. 3.1 we clarify the notation used throughout our work. In Sec. 3.2 we explain the Nyström method that we use to accelerate the eigendecomposition in detail. In Sec. 3.3 we elaborate on the robustificaton method we incorporate in our method SCAR.

3.1 Notation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ be an undirected, weighted graph where $\mathcal{V}$ denotes a set of nodes, $\mathcal{E}$ denotes a set of edges connecting nodes, and w denotes a weight function on the edges $w : \mathcal{E} \to \mathbb{R}^{>0}$. Its adjacency matrix $A \in \mathbb{R}^{n \times n}$ is defined by its entries a_{ij} with $a_{ij} = w(v_i, v_j)$ if $(v_i, v_j) \in \mathcal{E}$, else $a_{ij} = 0$. Let $D := \text{diag}(\deg(v_1), \ldots, \deg(v_n)) \in \mathbb{R}^{n \times n}$ be the degree matrix of $\mathcal{G}$ where $\deg(v_i) := |\{v_j \in V \mid (v_i, v_j) \in \mathcal{E}\}|$ is the degree of node v_i. We define the Laplacian L of $\mathcal{G}$ as $L := D - A$. The Laplacian L is symmetric and positive-semidefinite in $\mathbb{R}^{n \times n}$. Hence, the n eigenvalues $\Lambda = [\lambda_1, \ldots, \lambda_n]$ of L are real and positive. The associated eigenvectors are denoted by $H = [h_1, \ldots, h_n]$, resp., the approximated eigenvectors by $\hat{H}$. Furthermore, we denote by $\mathcal{X} = \{x_i\}_{i=1}^{n}$ the set of n input data samples, where $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector.

3.2 Nyström Method for Eigenvector Approximation

The Nyström method has shown great promise in existing literature to speed-up the eigenvector calculation (e.g., [16, 32, 53, 54]). To accelerate the eigenvector computation, we use only a subsample of the whole dataset. A matrix $M \in \mathbb{R}^{n \times n}$ can be partitioned into:

$$M = \begin{bmatrix} M_1 & {M_2}^T \\ M_2 & M_3 \end{bmatrix}, \tag{1}$$

where $M_1 \in \mathbb{R}^{m \times m}$ represents the affinities between m sampled points in the subset, $M_2 \in \mathbb{R}^{(n-m) \times m}$ contains all weights from the $n-m$ remaining points to the m subsampled points and $M_3 \in \mathbb{R}^{(n-m)\times(n-m)}$ captures the remaining affinities between all points not chosen for the subset. After choosing landmark points for the approximation, the eigenvectors H_1 of M_1 can be calculated. We introduce diverse eigendecomposition approaches that can be used in Sec. 2.1 and compare them empirically in Sec. 5.6.2.

Using the Nyström extension [16], we can extrapolate the eigenvectors for all remaining points. Let H and Λ be the eigenpairs of M, it follows:

$$M = H\Lambda H^T = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \Lambda \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}^T = \begin{bmatrix} H_1\Lambda H_1^T & H_1\Lambda H_2^T \\ H_2\Lambda H_1^T & H_2\Lambda H_2^T \end{bmatrix} \quad (2)$$

Thus, if H_1 denotes the eigenvectors for the subsampled points M_1, we can deduce H_2, the eigenvectors for all remaining points, with $H_2 = M_2 H_1 \Lambda^{-1}$. Sorting the extrapolated eigenvectors for the remaining points back into the calculated eigenvectors for points chosen as subsample yields the approximated eigenvectors $\hat{H} \in \mathbb{R}^{n\times m}$ for M:

$$\hat{H} = \begin{bmatrix} H_1 \\ M_2 H_1 \Lambda^{-1} \end{bmatrix} \quad (3)$$

In the last step, we orthogonalize the approximated eigenvectors $\hat{H}$. By using only a subsample of the data, the time complexity can be reduced from $O(n^3)$ to $O(nm^2 + m^3)$, where usually $m \ll n$ [33].

3.3 Robustifying Spectral Clustering

In order to robustify spectral clustering, we follow RSC [7]. The main idea is to separate the input graph with adjacency matrix A into two latent subcomponents described by A^g and A^c:

$$A = A^g + A^c \quad (4)$$

A^c reflects the **c**orrupted edges in the graph and A^g contains only the noise-free, "**g**ood" edges. The partitioning into two segments can be determined and improved by independently optimizing the spectral embedding for each subgraph. In practice, it is sufficient to resolve only one component, since its counterpart can easily be deduced from the original representation (see Equation 4). In [52], it has been shown that spectral clustering can be transformed into a trace minimization problem for A. Following this idea, in [7], the authors proved that the solution to A^c can be attained by solving a maximization problem for $Tr(H^T L(A^c) H)$, where $L(A^c)$ denotes the Laplacian of matrix A^c. The corresponding objective function for the unnormalized Laplacian (cf. [7]) is defined as:

$$f([a_e^c]_{e\in\mathcal{E}}) := \sum_{(v_i,v_j)\in\mathcal{E}} a_{i,j}^c \cdot ||h_i - h_j||_2^2 \quad (5)$$

Further constraints are given by θ and m. The parameter θ denotes the maximum number of global corruptions that are deleted: $|\{(v_i,v_j)|a_{ij}^c \neq 0\}| \leq 2\cdot\theta$. The parameter m implies the minimum number of nodes that each node in A^g is connected to: $|\{v_j|a_{ij}^g \neq 0\}| \geq m \cdot deg(v_i)$ for each node v_i.

To solve the maximization problem in order to find edges which should be assigned to A^c, we use a greedy approach that is described in [30]. The idea is to sort all edges $e \in \mathcal{E}$ in descending order according to their scores p_e being defined as:

$$p_e = p_{ij} = a_{ij} \cdot ||h_i - h_j||_2^2 \quad (6)$$

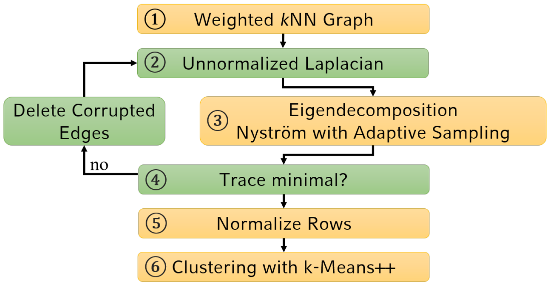


Figure 2: Overview of our method SCAR. Green boxes imply steps in our method that are analogue to RSC [7] and orange implies a significant change or addition.

We iteratively add edges to A^c such that the side constraints defined by parameters θ and m are fulfilled. Further details, proofs, and the reduction to the multidimensional knapsack problem [42] can be found in [7].

4 SCAR - SPECTRAL CLUSTERING ACCELERATED AND ROBUST

We propose our new clustering method SCAR (**S**pectral **C**lustering – **A**ccelerated and **R**obustified). SCAR separates the affinity graph of the data in an iterative approach into two latent components: a clean graph, which is used for the subsequent clustering, and a graph containing noisy edges. Likewise to Robust Spectral Clustering (RSC) [7], it detects noisy edges in each step that are disadvantageous for clustering. Therefore, it reaches overall robustness against noise compared to the original spectral clustering [40]. SCAR is significantly faster than RSC as we accelerate the most time-intensive step, the eigendecomposition, using the Nyström method [16] explained in Sec. 3.2. Furthermore, we improve several aspects of RSC significantly, such that SCAR is not only faster but also achieves significantly better results in real-world experiments as shown in Sec. 5. Fig. 2 gives an overview of our method SCAR and highlights the most important steps that deviate from RSC. In the following, we describe each step of our method in more detail. Algorithm 1 shows the corresponding pseudocode.

Step 1. We calculate the symmetric, weighted kNN graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ of the input data $\mathcal{X}$ (cf. line 1 in the pseudo-code). Each data point $x_i \in \mathcal{X}$ implies a node $v_i \in \mathcal{V}$, where $|\mathcal{X}| = |\mathcal{V}| = n$, i.e., there exists a bijective mapping $\phi : \mathcal{X} \rightarrow \mathcal{V}$. Further, $\mathcal{E} = \{(v_i, v_j) \mid \forall v_i, v_j \in \mathcal{V}, i \neq j : v_i \in kNN(v_j) \vee v_j \in kNN(v_i)\}$, and the weight of edges is defined by the Gaussian similarity function:

$$w(v_i, v_j) = exp\left(\frac{-||\phi^{-1}(v_i) - \phi^{-1}(v_j)||^2}{2\sigma^2}\right) \quad (7)$$

In our experiments, we use $\sigma = \sqrt{nd/2}$ per default. kNN graphs are suitable to find clusters of arbitrary shapes and varying density. Thus, when using spectral clustering, they are on most real world datasets superior to fully connected graphs (FCG), ϵ-neighborhood graphs, or Gabriel Graphs [25]. Our evaluation in Sec. 5 supports these findings. In contrast to RSC [7], we use a weighted kNN graph and apply the Gaussian kernel to weight the edges, which gives more weight to closer points than to points farther away. This further improves clustering results in general [40], as it offers more information that can be relevant for clustering.

Step 2. As elaborated in [22], the unnormalized Laplacian is more suitable than normalized versions to discern between clusters and outliers resp. noise in the eigenspace, amplifying the difference between corrupted and clean edges later. Thus, we calculate the unnormalized graph Laplacian (line 4) as $L = D - A$ based on $\mathcal{G}$.

Step 3. We approximate the eigenvalues of L with the Nyström method as described in Sec. 3.2. Following [38], we use an adaptive sampling approach, where we choose $\alpha \cdot n$ points with the highest degrees as landmarks (cf. line 6). As noise points are unlikely to be nearest neighbors of nodes outside of their own neighborhood (compare to, e.g., ideas of outlier detection algorithms ODIN [19] or *k*NN-LOF [55]), their corresponding nodes in the *k*NN graph tend to have lower degrees. As we sample the nodes with the highest degrees, potential losses regarding the set of edges concern prevalently the noisy edges, we want to remove anyway. We then approximate the first k eigenvectors on $L_1 \in \mathbb{R}^{(\alpha n)\times(\alpha n)}$, which is a small submatrix of $L \in \mathbb{R}^{n\times n}$ (see Equation 1). The resulting matrix $\tilde{H_1} \in \mathbb{R}^{\alpha n\times k}$ contains the first k approximated eigenvectors of L_1. The sampling of the submatrix is outlined in lines 7-10, whereas line 11 shows the eigendecomposition. In line 12 the Nyström extension is carried out. In Sec. 5 we thoroughly investigate several decomposition methods for computing the eigenpairs of L_1. In the following we work on $\tilde{H} \in \mathbb{R}^{n\times k}$, as obtained by Eq. 3.

Step 4. We check in line 15 of the pseudo-code whether the trace of $\tilde{H}^T L(A^g)\tilde{H}$ has decreased compared to the previous iteration. The identification and extraction of corrupted edges in the graph is described in Sec. 3.3 and follows the approach of RSC [7]. We apply Equation 6 in line 19, i.e., p_{ij} is calculated for all edges $(v_i, v_j) \in \mathcal{E}$. High values for p_{ij} indicate a high dissimilarity between the embeddings of nodes v_i and v_j, even though the nodes are connected by an edge. Thus, assigning an edge (v_i, v_j) with a high value p_{ij} to the noise component A^c improves the clustering quality as the edge is disregarded in the subsequent clustering step.

However, to ensure sparsity thresholds, bounds set with θ and m are respected [7]. The parameter θ prevents eliminating too many edges required for reasonable clustering results by limiting the maximum number of overall removable edges. The parameter m ensures a maximum local sparsity, i.e., each node keeps at least a portion m of its originally connected edges. In our experiments, we use $m = 0.5$ per default. If updates on the graph separation still lead to quality improvements, we recalculate L and approximate its eigendecomposition again with the Nyström method. We alternatingly update A^g in line 22 and $\tilde{H}$ until the trace cannot be significantly lowered.

Step 5. As suggested by [7, 11, 16], we orthogonalize and norm the first k resulting approximated eigenvectors row-wise (cf. line 24) which increases clustering quality and stability:

$$\bar{H}_{[i,:]} = \frac{\tilde{H}_{[i,:]}}{\|\tilde{H}_{[i,:]}\|_2} \tag{8}$$

Algorithm 1: SCAR Algorithm

```
   Input: Dataset $X$, user input $k$, $nn$, $\alpha$, $\theta$, $m$
   Output: Clustering containing assigned labels
 1 $A \leftarrow$ kNN_graph($X$, $nn$);
 2 $A^g \leftarrow A$;
 3 for $iter < max\_iterations$ do
 4     $L \leftarrow$ Laplacian($A^g$) ;                    // see Sec. 3.1
 5     /* Nyström method                                   */
 6     $X_l \leftarrow \alpha \cdot |X|$ landmarks;
 7     $i \leftarrow$ indices_of($X_l$);
 8     $j \leftarrow$ indices_of($X \backslash X_l$);
 9     $L_1 \leftarrow L[i, i]$;
10     $L_2 \leftarrow L[j, i]$;
11     $\tilde{H_1}, \Lambda \leftarrow$ eigendecomposition($L_1$);
12     $\tilde{H_2} \leftarrow L_2\tilde{H_1}\Lambda^{-1}$ ;              // see Equation 3
13     $\tilde{H} \leftarrow$ reassemble($\tilde{H_1}, \tilde{H_2}$);
14     $trace \leftarrow$ sum($\Lambda$);
15     if $trace$ is minimal then
16         break;
17     end if
18     /* Removing corrupted edges                        */
19     $p_{i,j} \leftarrow a_{i,j} \cdot \|h_i - h_j\|_2^2$ ;          // see Equation 6
20     $removed\_edges \leftarrow$ edges(argmax($p$), $\theta$, $m$);
21     $A^c \leftarrow$ matrix($removed\_edges$);
22     $A^g \leftarrow A - A^c$;
23 end for
24 $\bar{H} \leftarrow$ row-wise_norm($\tilde{H}$);
25 $Clustering \leftarrow k$_means++(rows($\bar{H}$));
```

Step 6. In the last step (shown in line 25), we cluster the first k rows of $\bar{H}$ (that has the eigenvectors of A^g as columns) with k-Means++ [2]. k-Means++ improves the selection of initial cluster centers for k-Means, leading to an earlier convergence and thus further speed-up compared to traditional spectral clustering approaches using k-Means.

5 EVALUATION

In the following, we examine our method SCAR thoroughly. In Sec. 5.1 we present our experimental setup. In Sections 5.2 and 5.3 we analyze SCAR's clustering quality, noise robustness, efficiency and scalability. In Sec. 5.4 we summarize SCAR's clustering and speed performance and regard their mutual dependencies. In Sec. 5.5 we evaluate the improvements of SCAR over RSC and SC. In Sec. 5.6 we evaluate the influence of various hyperparameters and design choices. SCAR retained an excellent balance between speed and quality over all experiments, while we refrained from hiding experiments that did not deliver desirable results in order to prevent overoptimism [8].

5.1 Experimental Setup

Datasets. In our evaluation, we use two synthetic datasets and ten real-world benchmark dataset. Both synthetic datasets, moons and

Table 1: Dataset properties used in the analysis.

	dataset	n	d	k	noise [%][9]	LB-UB [%][10]
syn.	moons	1,000	2	2	15	
	circles	1,000	2	2	15	
real	iris	150	4	3	7	5-9
	dermatology	366	33	6	9	4-14
	banknote	1,372	4	2	2	0-4
	pendigits$_{16}$	1,499	16	2	1	0-2
	pendigits$_{146}$	2,279	16	3	1	0-2
	pendigits	7,494	16	10	9	2-13
	USPS	11,000	256	10	24	12-33
	MNIST-10K	10,000	784	10	24	13-29
	MNIST-20K	20,000	784	10	21	11-27
	letters	20,000	17	26	46	20-61

circles, are constructed using data generator functions from the scikit-learn library.

Real world benchmark datasets *iris*, *dermatology*, *banknote*, *pendigits*, and *Letter Recognition* (*letters* for short) were obtained from the *UCI- MLR*[1]. *MNIST* and *USPS* were obtained from the repository of the *CS NYU*[2]. Similar to the work of [7], random subsamples were selected for the MNIST dataset. For the pendigits dataset, specific subsets pendigits$_{16}$ and pendigits$_{146}$ were defined as benchmark datasets in the literature [7, 23, 35]. For *dermatology* we omit the feature about the *age* of patients as the dataset is incomplete w.r.t this feature. The data statistics are summarized in Tab. 1.

Competitors. We compare SCAR with standard Spectral Clustering (SC) [40][3], Robust Spectral Clustering (RSC) [7][4], normalized SpectACl [21][5], and Density Core Finding (DCF) [50][6].

Implementation Details. SCAR is implemented in Python, building off of the implementation of RSC [7][6]. We additionally use the libraries scikit-learn, NumPy, Scipy and slepc4py/petsc4py [7]. Experiments were run on an Intel(R) Xeon(R) Silver 4208 CPU @ 2.10GHz using 32GB RAM.

Code: available on GitHub [8]

Hyperparameter Setting. For the synthetic datasets we use per default 0.15 for the parameter *noise* that regulates the jitter. Note that this value is significantly higher than the values applied in, e.g., RSC [7]. We tune $\alpha \in [0.1, 0.2, \dots, 0.9]$. For every dataset, we fix the parameter θ in all our experiments dataset-specific, where $\theta \in \{20, 30, 200, 500, 1k, 10k, 30k, 60k\}$. Following the rule-of-thumb popularized by [14], we used $2\sqrt{n}$ as an upper bound for nn and tested values in 10 percent steps for all methods, accordingly. For a fair comparison to the competitor *DCF*, we also evaluated the parameter β used in their method in the range of $[0.1, 0.2, \dots, 0.9]$ to obtain best scores for the cluster metric.

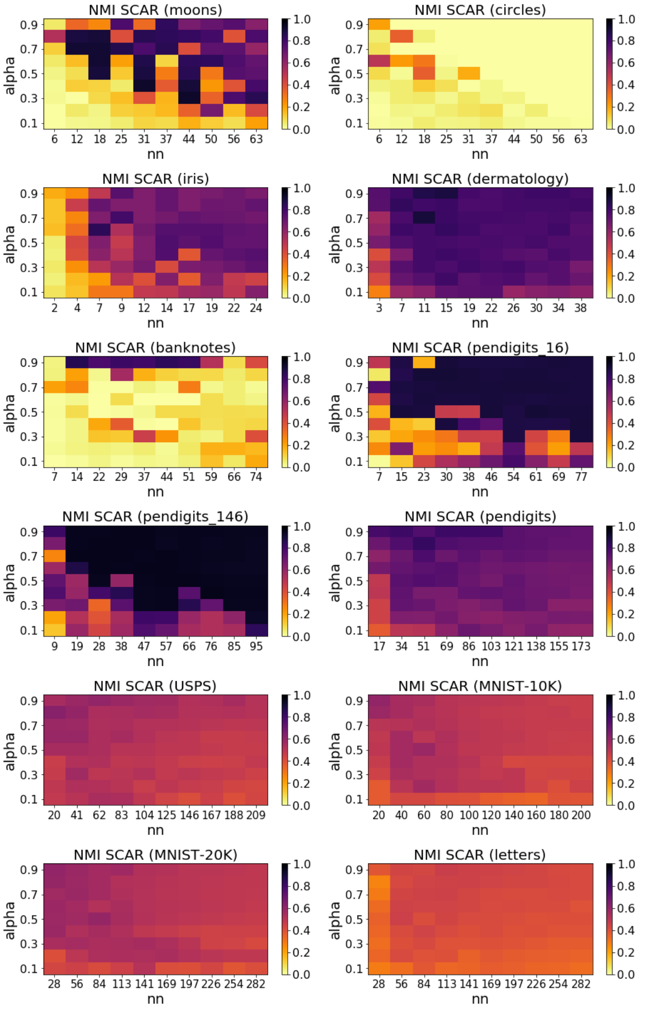


Figure 3: Summary of NMI obtained with SCAR depending on *nn* and α for all datasets.

5.2 Clustering Quality

Clustering quality is measured using the *Normalized Mutual Information* (NMI) [48] and *Adjusted Rand Index (ARI)* scores, which range from 0 to 1. Higher values imply a better accordance to the ground truth. Following the suggestion of [44], ARI should be used when the reference clustering has large equal sized clusters; scores based on mutual information should be used when the reference clustering is unbalanced and there exist small clusters. In the following, we run all experiments for 10 trials and report the average clustering scores per parameter setting if not stated otherwise.

5.2.1 *Effectiveness.*

In Tab. 2 on the left, we summarize the best NMI and ARI scores evaluated on each dataset. In order to obtain the best outcomes for

[1] https://archive.ics.uci.edu/ml/index.php (retrieved: Feb 18, 2022)
[2] https://cs.nyu.edu/~roweis/data.html (retrieved: Feb 25, 2021)
[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html (last accessed: Jul 14, 2022)
[4] https://github.com/abojchevski/rsc(last accessed: Jul 14, 2022)
[5] https://bitbucket.org/Sibylse/spectacl/src/master/ (last accessed: Jul 14, 2022)
[6] https://github.com/tobinjo96/DCFcluster (last accessed: Jul 14, 2022)
[7] https://slepc.upv.es/documentation/ (last accessed: Jul 14, 2022)
[8] https://github.com/SpectralClusteringAcceleratedRobust/SCAR.git

[9] for synth. datasets noise is added in the sklearn function as standard deviation of Gaussian noise, for real datasets, noise is measured as ratio of inter-cluster edges vs. total edges in the k-Graph for all tested nn
[10] LB = noise lower bound, UB = noise upper bound

Table 2: Maximum clustering quality reached, measured by normalized mutual information (NMI) scores and adjusted rand index (ARI), as well as minimum runtimes (in seconds) reached for best NMI scores and overall. Best/Second-best results are **bold**/underlined. Values regarded closer in the text are marked in red for faster readability.

		max NMI ARI										min runtime of best NMI (min runtime overall)				
	dataset	SC		RSC		DCF		SpectACl		SCAR		SC	RSC	DCF	SpectACl	SCAR
syn.	moons	0.43	0.72	0.43	0.72	0.91	0.96	0.88	**0.98**	**0.92**	0.96	0.15 (0.11)	0.19 (0.14)	0.14 (0.13)	0.11 (0.08)	**0.06** (**0.03**)
	circles	0.00	0.00	0.19	0.08	0.33	0.16	**0.79**	**0.86**	0.50	0.57	0.13 (0.11)	0.32 (0.20)	0.09 (0.07)	0.07 (0.06)	**0.05** (**0.03**)
real	iris	0.82	0.83	0.81	0.75	0.77	0.75	0.76	0.73	**0.84**	**0.85**	**0.03** (**0.02**)	0.04 (0.04)	0.08 (0.06)	0.06 (0.04)	**0.03** (**0.02**)
	dermatology	**0.93**	0.91	**0.93**	**0.92**	0.91	0.88	0.88	0.88	0.89	0.89	**0.03** (**0.03**)	0.09 (0.05)	0.09 (0.08)	0.08 (0.08)	0.05 (0.04)
	banknote	0.61	0.62	0.61	0.62	0.62	0.62	0.02	0.03	**0.86**	**0.90**	0.16 (0.15)	0.35 (0.19)	0.11 (0.09)	**0.10** (0.10)	0.12 (**0.03**)
	pendigits$_{16}$	**0.92**	**0.95**	0.90	0.94	0.78	0.76	0.22	0.10	0.90	0.94	0.26 (0.18)	0.37 (0.21)	**0.13** (0.12)	0.17 (0.14)	**0.13** (**0.08**)
	pendigits$_{146}$	0.95	0.96	**0.96**	**0.97**	0.87	0.86	0.70	0.58	0.95	**0.97**	0.41 (0.41)	0.87 (0.69)	0.29 (0.26)	0.29 (0.29)	**0.27** (**0.17**)
	pendigits	0.81	0.67	0.82	0.67	**0.84**	**0.76**	0.74	0.59	0.82	**0.76**	3.88 (2.94)	8.25 (4.05)	**0.96** (**0.80**)	2.09 (1.73)	2.68 (1.38)
	USPS	0.65	0.46	**0.68**	0.45	0.60	0.31	0.58	0.42	0.63	**0.48**	22.22 (22.22)	10.33 (9.70)	55.42 (54.89)	**4.00** (3.86)	4.59 (**3.18**)
	MNIST-10K	0.67	0.50	**0.74**	**0.55**	0.59	0.45	0.62	0.50	0.61	0.44	36.29 (36.29)	10.49 (10.49)	114.03 (111.82)	**5.00** (4.91)	7.34 (**4.41**)
	MNIST-20K	0.68	0.51	**0.76**	**0.55**	0.62	0.49	0.63	0.49	0.60	0.52	244.87 (244.87)	46.45 (31.39)	444.92 (385.94)	**21.18** (**21.18**)	38.83 (**21.18**)
	letters	0.42	0.16	0.42	0.13	**0.56**	0.17	0.38	0.12	0.46	**0.22**	418.02 (62.48)	38.29 (38.29)	**8.94** (**8.91**)	13.88 (12.99)	19.06 (10.84)
	Avg.	0.65	0.60	0.68	0.61	0.70	0.59	0.6	0.52	**0.74**	**0.70**	60.53 (30.81)	9.67 (7.95)	52.1 (46.93)	**3.91** (3.78)	6.10 (**3.44**)

each dataset and each method, we applied a grid-search over the respective parameter spaces as outlined in Sec. 5.1. The dependence of NMI's on the number of neighbors *nn* can be seen in Fig. 4. We discuss the runtimes shown in the right part of Tab. 2 – also in combination with the quality of the clusterings – in Sec. 5.3 and analyze influence of hyperparameter settings in Sec. 5.6.

SCAR reaches on average the best NMI/ARI scores while those results were reached on average with the second best runtime of all tested algorithms. SCARs average runtime is approximately an order of magnitude faster compared to the original standard spectral clustering algorithm (SC) and DCF. SCAR always returns clusterings of solid quality, in contrast to, e.g., SpectACl, which is not able to find an acceptable clustering for the datasets banknote or subsets of the pendigits dataset (marked in red in Tab. 2, see also Fig. 4). Second best values were often reached by SC, which is, however, not designed to reach fast runtimes. SC's good results on our benchmark datasets confirm the high potential of spectral methods for high-quality clustering results. Further, we observe that SCAR can handle highly noisy datasets like moons, where SC as well as RSC could not correctly detect the clusters, reaching NMIs (ARIs) of only 0.43 (0.72). We regard the sensitivity of all methods w.r.t. the parameter *nn* in Fig. 4. Where most methods are rather robust w.r.t. the parameter *nn*, their default settings may not be optimal: in Fig. 1, we applied all algorithms' default parameter settings on the moons dataset. Here, none of our competitors could find the clusters correctly. In contrast, we optimized parameter settings w.r.t. the NMI/ARI via a grid search for Tab. 2. Furthermore, we perceived the banknote dataset as an interesting case, as SCAR significantly surpassed its competitors. SpectACl, e.g., was not able to produce any meaningful clustering results over a variety of tested parameter settings, reaching a maximum NMI (ARI) of 0.02 (0.03). All other competitors reached NMI/ARI scores around 0.62. The banknote dataset contains 4-dimensional representations of forged and authentic banknotes. Its clusters overlap in all dimensions, making its similarity graph highly noisy. Thus, the advantages of SCAR's noise robustness can be seen here, yielding an outstanding NMI (ARI) of **0.86** (**0.90**) for our method.

Even though SCAR yields very good results for most datasets, we still see room to further improve clustering results on high-dimensional datasets in future work. Especially, performance on datasets emerging from pixel-data (USPS and MNIST versions) could benefit from applying feature weighting approaches as outlined in Sec. 2. Tab. 2 shows that despite their different strengths, the clustering metrics do not differ much in how the investigated methods compare. Thus, only NMI is reported in the following as the default metric.

5.2.2 *Robustness against Noise.*

To evaluate SCAR's robustness against noise, we fix the parameter settings for *nn* and α, and only modify the amount of *jitter* in the range of $[0.0, 0.05, 0.1, \ldots, 0.03]$ on the moons dataset. The left graph in Fig. 5 shows that SCAR consistently outperforms other models on the moons dataset for high noise levels ($noise > 0.2$). The NMIs of most comparative methods drop heavily for noise values over 0.1, resp., 0.2. Qualitative results can also be seen in Fig. 1, where SCAR is the only method able to correctly discern the two moons for a comparably high noise level of $noise = 0.15$. The right graph in Fig. 5 gives all runtimes in seconds. SCAR shows an almost constant runtime over different levels for *noise*. For RSC we observe higher runtimes as the eigendecomposition on the whole Laplacian is computed in each iteration. Notably, DCF also shows increased runtimes for low noise values due to higher densities within the clusters. The efficiency of our model evaluated on different benchmark datasets is further discussed in the next section.

Similar to [7], we also examine the robustness against *noisy edges*: We generated Gaussian distributed clusters (blobs) and versions of the moons datasets where we added "corrupted" edges to the associated *k*NN graph. I.e., we added edges between nodes of different clusters using the planted-partition model. Intra-cluster edges were created with a probability of 30% and we added noise edges s.t. 10%, resp., 20% of all edges in the *k*NN graph were corrupted. We evaluate the precision $p = |\mathcal{E}_c \cap \mathcal{E}_r|/|\mathcal{E}_c|$ and recall $r = |\mathcal{E}_c \cap \mathcal{E}_r|/|\mathcal{E}_r|$, where $\mathcal{E}_c$ is the set of corrupted edges and $\mathcal{E}_r$ is the set of edges removed by SCAR. In contrast to [7], we also regard the effect of removing corrupted edges on the clustering

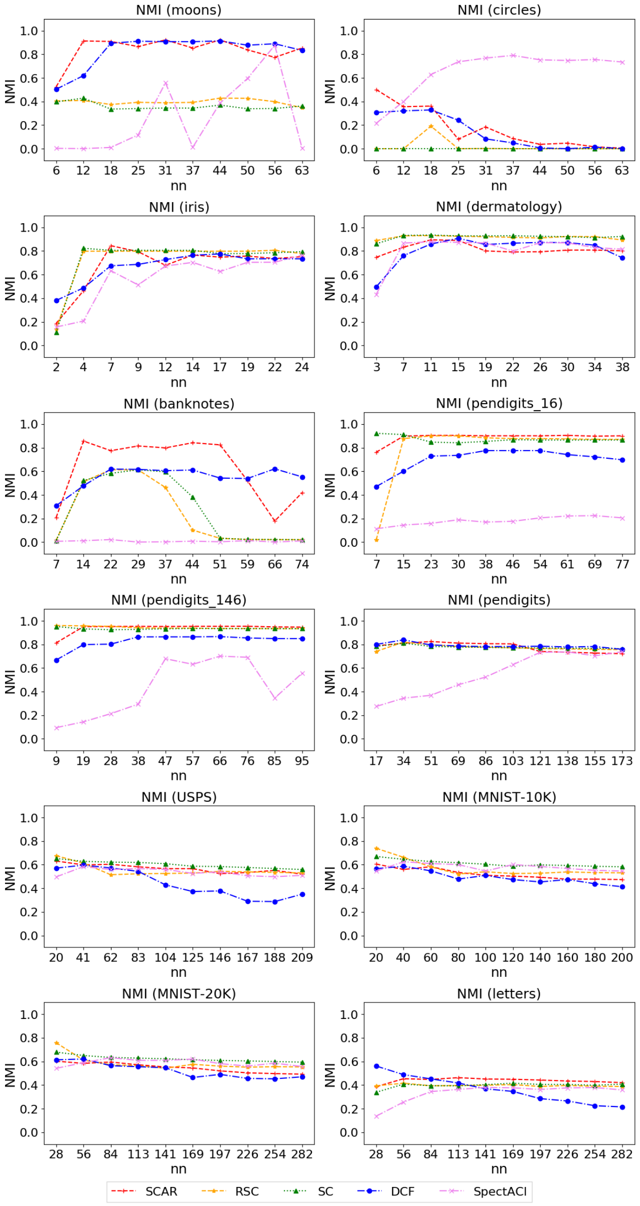


Figure 4: Comparison of NMI of all methods with their best parameter settings on all datasets depending on *nn*.

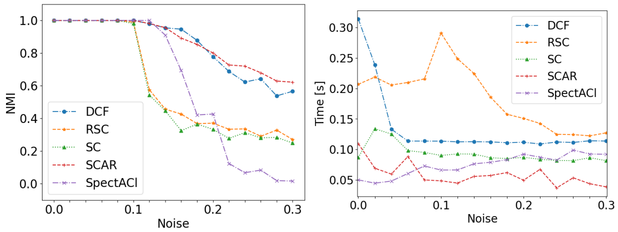
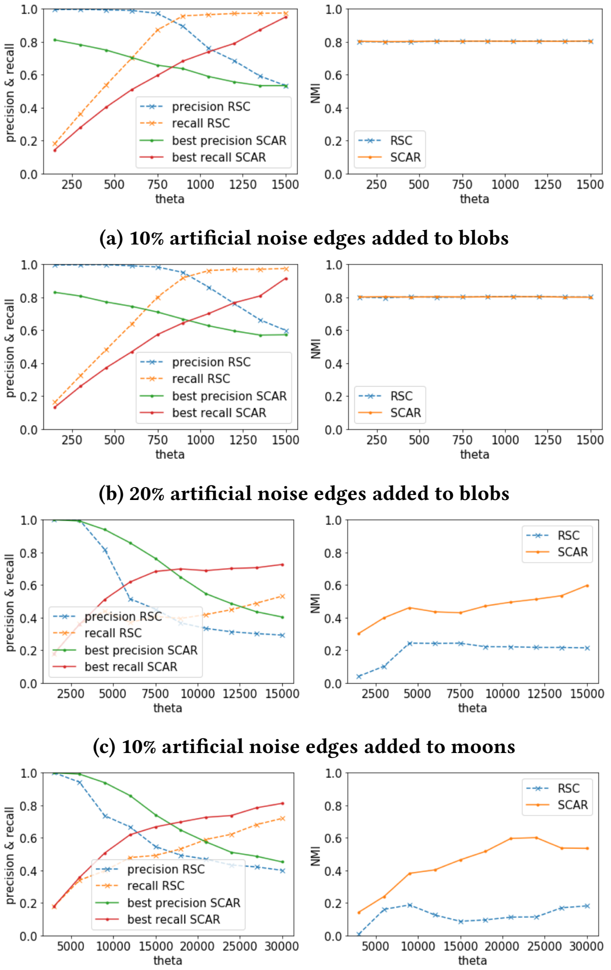


Figure 5: NMI scores (left) and runtime in [s] (right) for $noise \in [0, \ldots, 0.3]$ in 0.02 steps on moons dataset.



(a) 10% artificial noise edges added to blobs

(b) 20% artificial noise edges added to blobs

(c) 10% artificial noise edges added to moons

(d) 20% artificial noise edges added to moons

Figure 6: Precision and recall (left) and NMI (right) for 10% or 20% artificial noise edges added to blobs (n=1000, k=20) resp. moons averaged over random_state=$[0 - 9]$.

quality.Figures 6a and 6b show precision and recall of the detected corrupted edges, as well as the achieved NMIs of RSC and SCAR for increasing θ. Even though precision and recall – implying the quality of detecting corrupted edges – of SCAR's results are lower than for RSC, this does not affect the overall clustering quality. Instead, the constant NMI scores, while increasing θ for both cases (10% and 20% added noise edges), indicate that corrupted edges do not affect the obtained clustering quality for Gaussian distributed clusters. Figures 6c and 6d imply that this is different for moons datasets. Here, SCAR surpasses RSC w.r.t. precision and recall on both noise settings throughout almost all tested values for θ. In contrast to the Gaussian distributed clusters, for the moons dataset, removing corrupted edges enables a higher clustering quality: The NMIs of SCAR are significantly higher than the NMIs for RSC throughout all tested values for θ. Fig. 6 shows that SCAR surpasses RSC in the detection of corrupted edges exactly where these corrupted edges impede a high quality clustering by connecting hard-to-distinguish clusters.

5.3 Runtime Analysis

In this section, we evaluate our model's runtime. In Sec. 5.3.1, we provide a general overview of SCAR's efficiency compared to state-of-the-art methods. We perform scalability experiments in Sec. 5.3.2 and regard the complexity in Sec. 5.3.3

5.3.1 *Efficiency.*

Tab. 2 shows on the right the minimum runtime in seconds for the trials with the highest NMI scores as well as the minimum runtime of all tested parameter settings in brackets. We observe that SCAR yields its best results w.r.t. the NMI almost always as the fastest or second fastest algorithm. SCAR can generally provide faster results than standard SC. The speed-up, in particular, increases with larger datasets. Only on the quite small dermatology dataset, SCAR runs 0.02 seconds longer than SC. The highest speed-up is reached on the letters dataset, where SCAR is more than 20 times faster than SC, while simultaneously increasing the NMI by 10%. Using a similar design and notion of noise as RSC, it is noteworthy that we surpass RSC w.r.t the runtime on every tested dataset. Note also, that RSC already accelerates the eigendecomposition by leveraging IRLM. We reach a maximum acceleration factor of 6.4 in relation to RSC for the heavily noisy circles dataset, where we simultaneously improve the clustering quality from an NMI (ARI) of 0.19 (0.08) to 0.50 (0.57). For further runtime comparisons with RSC and SC, see Sec. 5.5.2. Even though DCF shows fast runtimes as well as good clustering results for most datasets, it cannot reach acceptable runtimes on high-dimensional datasets like USPS or MNIST variations (marked in red). Whereas experiments on lower-dimensional datasets show comparable runtimes in the same order of magnitude for all algorithms, DCF needs on these three high-dimensional datasets more than ten times longer than SCAR. Further investigations on the dependence of all algorithm's runtimes can be found in Sec. 5.3.2. SpectACl shows, similar to SCAR, good runtimes for its best results, but mostly returns significantly worse clustering results. Especially SpectACl's performance on the datasets banknote and the first two pendigits versions (marked in red) is of surprisingly low quality.

For some parameter settings, algorithms may have significantly lower runtimes than for others. E.g., for a small number of nearest neighbors *nn*, the respective nearest neighbors graph has less edges, and thus, most operations performed on it are faster. Analyzing the values in brackets in Tab. 2, we see the best runtimes over all tested parameter settings that can be reached for each experiment and each algorithm. I.e., in contrast to the runtimes regarded in the last paragraph, where we optimized parameter settings for a high NMI, we now optimize parameter settings for a low runtime. Also here, SCAR reaches most often the fastest runtimes. We note, that even for the most suitable parameter settings, DCF cannot achieve an acceptable runtime on high-dimensional datasets like USPS and MNIST variations (marked in red).

We observed that for all datasets, the minimum runtime for the best NMI results were usually close to the respective minimum runtime over all tested settings. More precisely, most of them were at maximum twice as high as the fastest runtime for the respective algorithm. This supports the idea that the model has a stronger influence on its runtime than the selected parameter setting.

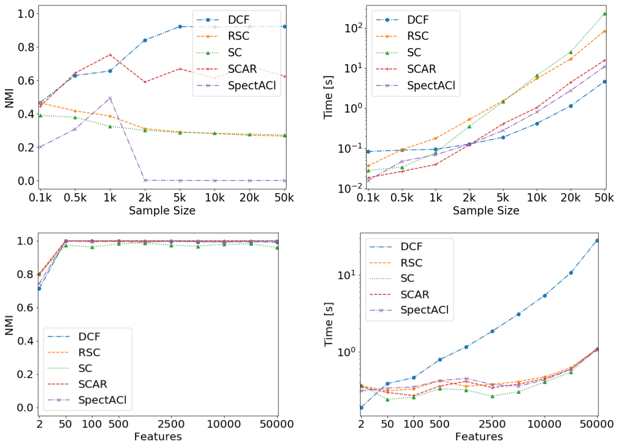


Figure 7: NMI scores (left) and runtime in seconds (right). Top: Moons dataset (noise=0.15) with varying n. Bottom: Blobs dataset (n=1000, k=20, random_state=None) with varying d.

In conclusion, we found that SCAR is stable in its anticipated clustering quality and yields good results at high speed.

5.3.2 *Scalability on Synthetic Datasets.*

Fig. 7 shows the scalability of our approach on the moons and blobs datasets, with a fixed noise of 0.15 for moons and a default cluster standard deviation of 1.0 for blobs. On the former, we scale the number of data points in the range $[100, \ldots, 50k]$. On the latter, we scale the number of features $[2, \ldots, 50k]$. The number of neighbors that are taken into account for the construction of the kNN graph is set to $nn = \sqrt{n}$ in both experiments, where n denotes the number of data samples. All other parameters are frozen. The left diagrams show the obtained NMI scores, and the right diagrams show the elapsed time for all evaluated methods. On the moons dataset, SCAR's scalability oupterforms RSC and SC w.r.t. both, computational time as well as obtained NMI scores for increasing sample size. For smaller datasets, our approach also shows superior performance compared to DCF, which cannot be maintained for increasing the sample-size. However, DCF's runtimes deteriorate for higher dimensionalities, as can be seen in the lower part of Fig. 7 (note the log-scale). SCAR's runtime stays almost linear in the number of features. In Fig. 8, DCF's unfortunate runtime behavior w.r.t. the dimensionality can also be observed on larger real-world datasets with higher dimensionality. While DCF yields low runtimes for large datasets if they are low-dimensional, e.g., for letters, its runtime tremendously increases on USPS, MNIST-10K and MNIST-20k.

5.3.3 *Complexity Analysis.*

Having the same fundamental structure as RSC, we refer for our complexity analysis on the explanations of [7], showing a runtime approximately linear in the number of edges. In the following, we elaborate on the differences between SCAR and RSC that potentially influence the complexity (see also Fig. 2). In Step (1), we calculate the weighted kNN graph in contrast to the unweighted kNN graph for RSC and apply a Gaussian kernel on the edge weights. These changes do not increase the runtime complexity, as all edges of the kNN graph are accessed in both approaches. In Step (3), [7] use the power iteration for the eigendecomposition. We reduce the runtime by using the Nyström method, see Sec. 3.2. In Step (5), we normalize the rows of the approximated, cleansed matrix $\tilde{H} \in \mathbb{R}^{n \times k}$, where $\tilde{H}$ contains the first k vectors that are needed for the subsequent clustering step. Usually, we have $k \ll \sqrt{n}$, s.t. the complexity is not increased when working on a kNN graph (containing approximately $O(n\sqrt{n})$ edges [39]). Overall, we reach a similar complexity as RSC, which is approximately linear in the number of edges, while improving the runtime. Our experiments in Sec. 5.3 confirm the improved runtime w.r.t. RSC.

5.4 Effectiveness and Efficiency

Fig. 8 summarizes the models' performances on the various datasets, where the x-axis shows the runtime and the y-axis shows the clustering scores. Optimal results are located in the upper left reflecting a high NMI score reached within a short amount of time. We show only the best runs of all methods to reduce visual clutter, i.e., only runs that yielded at least 75% of the best NMI score reached by the respective method are shown as single dots. On the highly noisy moons dataset, SCAR's robustness and efficiency dominates the other methods in terms of both, clustering performance and runtime. On small real-world datasets (iris, dermatology, banknotes and the pendigits variations), SCAR is highly competitive with other state-of-the-art models w.r.t. NMI and runtime. As all tested methods have runtimes under one second for all smaller datasets, larger datasets are more expressive for runtime analyses. Thus, we regard in the following (as well as in Fig. 10) the datasets with more than 5000 points (pendigits, USPS, MNIST versions and letters) when investigating runtimes. We note that SCAR is comparably fast on these datasets *and* reaches low runtimes with a comparably low variance. For the low-dimensional datasets pendigits and letters, DCF is even faster than SCAR, but for higher-dimensional datasets (USPS and MNIST versions) advantages of using any of the newer spectral clustering approaches become clear, as DCF's runtime does not scale with the dimensionality. In summary, Fig. 8 demonstrates that SCAR nearly always outperforms its competitors in either runtime, clustering quality, or both and particularly highlights SCAR's reliability.

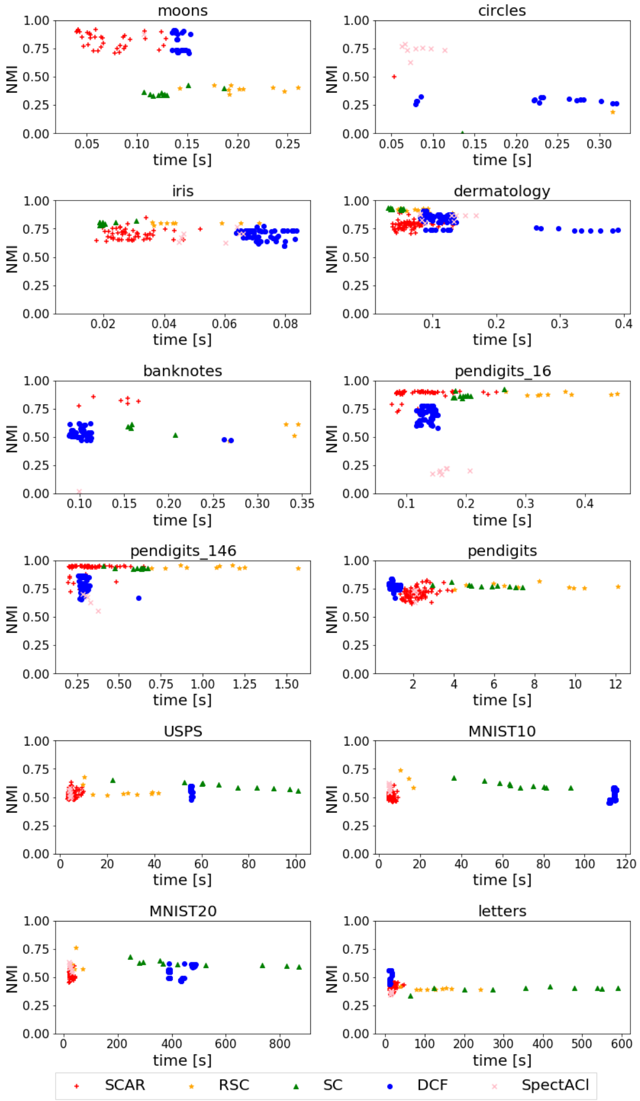


Figure 8: Runtimes and NMIs of all experiments that reached at least 75% of each method's respective best NMI.

5.5 Improvements over RSC and SC

In the following, we examine the improvements of SCAR over RSC and the original Spectral Clustering algorithm (SC) in more detail. Sec. 5.5.1 regards the single components that differentiate SCAR from RSC as well as their functional interaction. Sec. 5.5.2 regards runtime improvements over RSC.

5.5.1 *Effectiveness Improvements of SCAR over RSC and SC.*

Fig. 9 shows NMIs on the highly noisy moons dataset for various settings for nn and different methods: on the left, we compare RSC with a straight-forward Nyström-accelerated version of RSC and our method SCAR. On the right, we perform an ablation study w.r.t. the changes between RSC and SCAR. (We condense the results by setting α to our recommended default value $\alpha = 0.7$). The left part of the figure shows that a simple speed-up of RSC would lead to significantly worse results, whereas SCAR drastically improves the results of RSC. On the right, we can see that each of SCAR's components is chosen meaningfully, leading to an improvement of quality that is reached by the elaborated combination of concepts rather than any single adaption. We regarded the reasons for the individual components in Sec. 4 and explain their impacts and synergies in the following. Using an unweighted graph can deliver good results on the moons dataset, if exactly the right number for nn is chosen (i.e., such that only very few corrupted edges exist). However, as seen in the first line of Fig. 9 on the right, this leads to a strong and unpredictable dependence on guessing a good value for nn. Weighting the edges also allows for a more meaningful sampling of the edges for the Nyström method with the adjusted sampling method we apply: as corrupted edges connect nodes of different

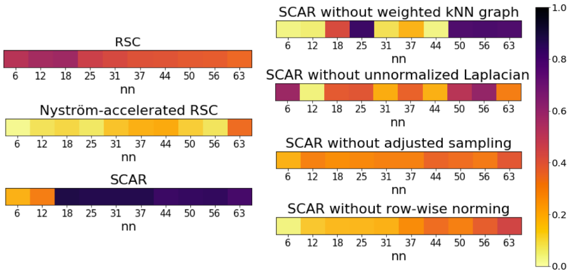


Figure 9: Ablation study of SCAR's NMI performance on moons (avg. over 10 random instantiations) depending on *nn* and for fixed $\alpha = 0.7$. Dark colors imply better NMI scores.

clusters (and distances between clusters are larger than distances inside clusters) they tend to be longer than non-corrupted edges. Applying the Gaussian kernel for calculating the edges' weights, this leads to smaller degrees of nodes connected by corrupted edges. Thus, sampling the nodes with higher degrees allows to sort out corrupted edges (compare with line 3 on the right of Fig. 9). Using an unnormalized Laplacian further enhances the distinguishability between corrupted and non-corrupted edges [22] in the eigenspace, reinforcing the positive effects of our adjusted sampling method heavily (see line 2 on the right of Fig. 9). Small perturbances in the data can be compensated by normalizing the rows [16]. That accounts for jitter and pushes points of a cluster even closer in eigenspace, which robustifies the adapted sampling step and further improves the clustering (line 4 on the right of Fig. 9).

5.5.2 *Efficiency Improvements over RSC and SC.*

Fig. 10 shows SCAR's substantial runtime accelerations over RSC depending on their parameter settings on four larger benchmark datasets (where $n \geq 5000$, resp., $runtime > 1s$). In Fig. 8, runtimes and their respective NMIs are shown for all methods. In particular for larger datasets, SCAR nearly always outperforms SC and RSC regarding runtime significantly. A more thorough discussion on the proper choice of hyperparameters α and *nn* is given in Sec. 5.6.1.

5.6 Hyperparameter Tuning

In this section we examine the influence of various parameter settings on our model's performance. In Sec. 5.6.1, we examine the portion α denoting the number of landmarks chosen for the Nyström subsample and the number of nearest neighbors *nn* for the construction of the kNN graph. In Sec. 5.6.2, we evaluate the performance of various decomposition methods on the sampled submatrix and how the clustering quality and computational time depends on different configurations. In Sec. 5.6.3, we investigate the influence of the parameter θ on the models' performances of SCAR and RSC.

5.6.1 *Number of Landmarks and Number of Neighbors.*

In Fig. 3, we show the NMI scores for all tested datasets depending on *nn* and α, where darker, resp. lighter, colors reflect higher, resp. lower, NMI scores. For a more thorough analysis of the impact of the number of neighbors, we use $2\sqrt{n}$ as an upper bound for *nn* [14]. We see that the choice for *nn* and α has a strong effect on the clustering quality: The quality of smaller datasets depends more heavily on a proper choice of α compared to larger datasets. Our experiments show that a higher amount of landmarks improves clustering results. Furthermore, the illustration reveals that on larger datasets, the performance is improved whenever the kNN graph retains its sparse nature, i.e., by lowering the amount of *nn*. This effect also heavily improves the efficiency of our proposed method as discussed in Sec. 5.3.1. On USPS, as well as on the MNIST datasets, we observe higher peaks for lower values of *nn*. On smaller datasets, it is more likely that the kNN graph connects samples from distinctive cluster, i.e., the graph contains misleading information. Comparing iris and dermatology, we found that for the latter, it is more favorable to choose a smaller *nn* to identify the six clusters properly, whereas on iris, with three clusters, we can choose higher values without mingling the information of separate clusters. Per default, we suggest to use values $\alpha = 0.7$ and $nn = \sqrt{n}$.

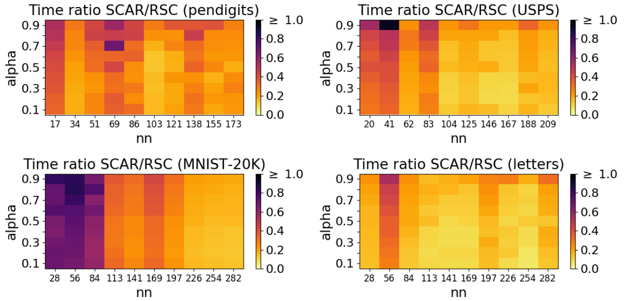


Figure 10: Summary of time ratios SCAR / RSC depending on *nn* and α for large real world datasets. Lighter colors imply better (i.e., faster) results for our method SCAR.

While good clustering results are a prerequisite for useful clustering algorithms, SCAR's major benefit is its runtime acceleration. Fig. 10 summarizes obtained runtime quotients of SCAR compared to RSC for four large real-world datasets and their dependence on *nn* and α. We only display the larger datasets here, as they require runtimes for RSC $\gg 1s$ (see Table 2), and therefore an acceleration analysis is more meaningful. The values in each heatmap depict the ratio of runtimes between SCAR and RSC, i.e., $runtime(SCAR)/runtime(RSC)$. Consequently, smaller values indicate faster runtimes of SCAR compared to RSC. The effect and strength of the Nyström method can be observed for all larger datasets. By sampling only a submatrix in order to approximate the spectrum as a whole, we observe a performance boost compared to RSC. The impact of the choice of α is shown on the y-axis, whereas the effect of *nn* is shown on the x-axis. The experiments show that SCAR has significantly lower runtimes than RSC even for high values of α, further supporting our quite high recommended choice for $\alpha = 0.7$. For larger values of *nn* SCAR's speed-up becomes even clearer: Larger *nn* lead to more edges in the kNN graph and therefore more acceleration potential for SCAR over RSC as the graph is more dense. Fig. 4 indicates that SCAR's clustering results are relatively robust against the choice of *nn*. Thus, SCAR's runtime improvements over RSC do not have a negative effect on its clustering performance.

5.6.2 *Decomposition of Submatrix.*

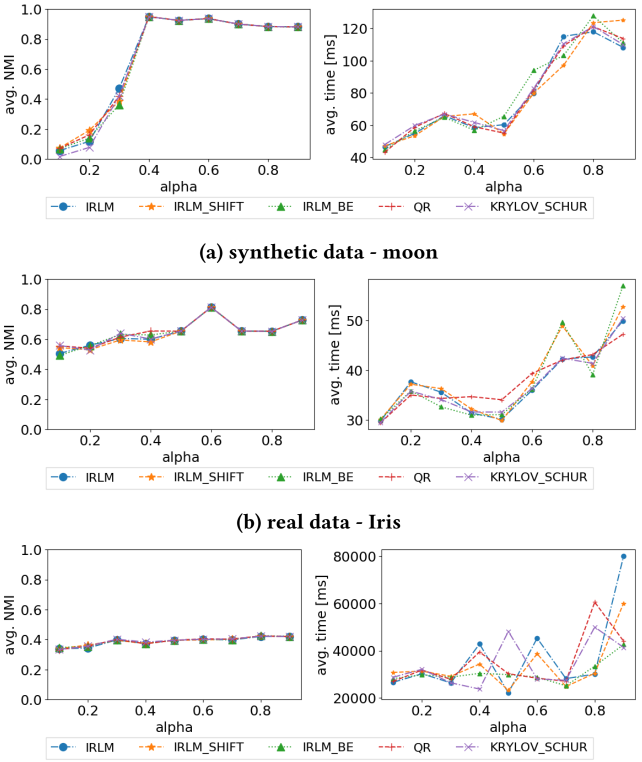
In the following, we evaluate commonly used decomposition methods on the sampled submatrix of the Nyström Approximation explained in Sec. 3.2. Fig. 11 shows the highest observed NMI scores (left) within 10 trials as well as the respective runtimes (right) with a fixed value of *nn* for each dataset. As the submatrix in the Nyström method is symmetric, we apply the *Implicitly Restarted Lanczos Method* (IRLM) which is based on power iterations and has also been used in [7] as decomposition heuristic. Additionally, we evaluate variants of IRLM with *-Shift* applying a shift-inversion on the spectrum to transform the smallest eigenvalues to be the highest, and *-BE* for which eigenvalues are approximated from both ends of the spectrum. For the latter, [41] showed, that approximating eigenpairs from both ends of the spectrum can speed-up the convergence. We also evaluate a standard QR decomposition, as well as the *Krylov-Schur* decomposition as proposed by [47].[11] Empirically, all decomposition methods yielded similar qualitative results w.r.t the NMI score. Examining the runtimes on smaller datasets, we observe a slight overhead in the computation of the shifting operation for *IRLM-Shift*, as well as in applying a sampling from both ends of the spectrum. On larger datasets, this effect flattens out and the *Krylov-Schur* decomposition that is optimized towards large, sparse matrices shows a marginal benefit for larger α values. In our experiments we used the standard IRLM as default heuristic for the computation of the eigenpairs as it showed competitive results over the full range of the tested datasets.

5.6.3 *Influence of Parameter θ.*

In Fig. 12, we evaluate the influence of parameter θ on the clustering's quality and runtimes for SCAR and RSC [7]. As argued in Sec. 5.6.1, we fix nn to $nn = \sqrt{n}$. We scale the number of expected corruptions in the dataset logarithmically: $\theta \in [10, 100, 1k, 10k]$. On the moons dataset, our approach outperforms RSC almost over the full range of chosen θ whilst drastically reducing the computational time as shown on the right. Generally, increasing the sparsity threshold might lead to a clearer separation, however, the clustering quality suffers for very large values as clean edges might be attached to the corrupted graph A^c.

6 CONCLUSION

We introduced SCAR, a novel robust and efficient clustering method. It elucidates the benefits from Robust Spectral Clustering [7] enhanced by the Nyström method for an accelerated computation of the eigendecomposition. We reduced the sensitivity to noisy input data as well as the runtime complexity compared to standard Spectral Clustering significantly. In a thorough experimental study, we compare SCAR's clustering quality with state-of-the-art models showing highly competitive results on real-world benchmark datasets, as well as its robustness against noise on artificial data. We evaluated robustness w.r.t. noisy edges in the similarity graph of the data as well as robustness w.r.t. jitter in the original data, tackling the two most difficult types of noise for clustering. SCAR consistently yielded low runtimes, in particular it is significantly faster than RSC and SC, while returning highly competitive clustering qualities on real-world and synthetic data. SCAR is recommendable when looking for a reliable, fast and robust clustering method on large and high-dimensional datasets that tend to be noisy.



(a) synthetic data - moon

(b) real data - Iris

(c) real data - Letter Recognition

Figure 11: Avg. NMI scores (left) and runtimes (right) for decomposition methods IRLM, IRLM-Shift, IRLM-BE, QR and krylov-schur on different datasets.

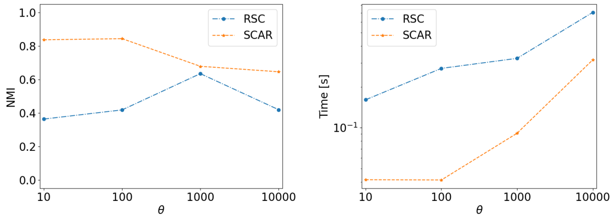


Figure 12: NMI scores (left) and runtime in [s] (right) for $\theta \in [10, 100, 1k, 10k]$ on moons dataset (noise=0.15).

ACKNOWLEDGMENTS

This work has been partially funded by VILLUM FONDEN, the Technical University of Munich's Institute for Ethics in Artificial Intelligence (IEAI) and the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

[11] We use state-of-the-art libraries, where IRLM and its variants are implemented in ARPACK, QR in LAPACK, and krylov-schur as part of SLEPc/PETSc

REFERENCES

[1] Walter Edwin Arnoldi. 1951. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics* 9, 1 (1951), 17–29.

[2] David Arthur and Sergei Vassilvitskii. 2006. *k-means++: The advantages of careful seeding*. Technical Report. Stanford.

[3] Francis Bach and Michael Jordan. 2004. Learning spectral clustering. *Advances in neural information processing systems* 16, 2 (2004), 305–312.

[4] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. 2011. Noise thresholds for spectral clustering. *Advances in Neural Information Processing Systems* 24 (2011).

[5] Anna Beer, Ekaterina Allerborn, Valentin Hartmann, and Thomas Seidl. 2021. KISS-A fast kNN-based Importance Score for Subspaces. In *EDBT*. 391–396.

[6] Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik. 2002. Spectral partitioning with indefinite kernels using the Nyström extension. In *European conference on computer vision*. Springer, 531–542.

[7] Aleksandar Bojchevski, Yves Matkovic, and Stephan Günnemann. 2017. Robust Spectral Clustering for Noisy Data: Modeling Sparse Corruptions Improves Latent Embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 737–746.

[8] Anne-Laure Boulesteix. 2015. Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Computational Biology* 11, 4 (2015), e1004191.

[9] Eamonn Cahill, Alan Irving, Christopher Johnston, James Sexton, Ukqcd Collaboration, et al. 2000. Numerical stability of Lanczos methods. *Nuclear Physics B-Proceedings Supplements* 83 (2000), 825–827.

[10] Xiaojun Chen, Weijun Hong, Feiping Nie, Dan He, Min Yang, and Joshua Zhexue Huang. 2018. Spectral clustering of large-scale data by directly solving normalized cut. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1206–1215.

[11] Anna Choromanska, Tony Jebara, Hyungtae Kim, Mahesh Mohan, and Claire Monteleoni. 2013. Fast Spectral Clustering via the Nyström Method. In *International Conference on Algorithmic Learning Theory*. Springer, 367–381.

[12] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. 2004. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 551–556.

[13] Petros Drineas and Michael W Mahoney. 2005. Approximating a gram matrix for improved kernel-based learning. In *International Conference on Computational Learning Theory*. Springer, 323–337.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork. 2001. *Pattern Classification* (2 ed.). Wiley, New York.

[15] Thomas Ericsson and Axel Ruhe. 1980. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Math. Comp.* 35, 152 (1980), 1251–1268.

[16] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. 2004. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 2 (2004).

[17] Alex Gittens and Michael Mahoney. 2013. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*. PMLR, 567–575.

[18] Stephan Günnemann, Ines Färber, Sebastian Raubach, and Thomas Seidl. 2013. Spectral subspace clustering for graphs with feature vectors. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 231–240.

[19] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. 2004. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. IEEE, 430–433.

[20] Bruce Hendrickson and Robert W. Leland. 1995. A Multi-Level Algorithm For Partitioning Graphs. *Supercomputing '95: Proceedings of the 1995 ACM/IEEE Conference on Supercomputing (CDROM)* (1995), 1–14.

[21] Sibylle Hess, Wouter Duivesteijn, Philipp Honysz, and Katharina Morik. 2019. The SpectACl of nonconvex clustering: a spectral approach to density-based clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3788–3795.

[22] Desmond J Higham and Milla Kibble. 2004. A unified view of spectral clustering. *University of Strathclyde mathematics research report* 2 (2004).

[23] Hao Huang, Shinjae Yoo, Hong Qin, and Dantong Yu. 2011. A robust clustering algorithm based on aggregated heat kernel mapping. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 270–279.

[24] Ling Huang, Donghui Yan, Nina Taft, and Michael Jordan. 2008. Spectral clustering with perturbed data. *Advances in Neural Information Processing Systems* 21 (2008).

[25] Tülin Inkaya. 2016. A Parameter-Free Similarity Graph for Spectral Clustering. *Expert Syst. Appl.* 42, 24 (dec 2016), 9489–9498. https://doi.org/10.1016/j.eswa.2015.07.074

[26] Hongjie Jia, Shifei Ding, Hong Zhu, Fulin Wu, and Lina Bao. 2013. A Feature Weighted Spectral Clustering Algorithm Based on Knowledge Entropy. *J. Softw.* 8, 5 (2013), 1101–1108.

[27] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20, 1 (1998), 359–392.

[28] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. 2009. Sampling techniques for the nystrom method. In *Artificial Intelligence and Statistics*. 304–311.

[29] Cornelius Lanczos. 1950. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA.

[30] Daniel Lehmann, Liadan Ita Oćallaghan, and Yoav Shoham. 2002. Truth revelation in approximately efficient combinatorial auctions. *Journal of the ACM (JACM)* 49, 5 (2002), 577–602.

[31] Richard B Lehoucq, Danny C Sorensen, and Chao Yang. 1998. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM.

[32] Mu Li, James Tin-Yau Kwok, and Baoliang Lu. 2010. Making large-scale Nyström approximation possible. In *ICML 2010-Proceedings, 27th International Conference on Machine Learning*. 631.

[33] Mu Li, Xiao-Chen Lian, James T Kwok, and Bao-Liang Lu. 2011. Time and space efficient spectral clustering via column sampling. In *CVPR 2011*. IEEE, 2297–2304.

[34] Xi Li, Weiming Hu, Chunhua Shen, Anthony Dick, and Zhongfei Zhang. 2014. Context-Aware Hypergraph Construction for Robust Spectral Clustering. *IEEE Transactions on Knowledge and Data Engineering* 26, 10 (2014), 2588–2597. https://doi.org/10.1109/TKDE.2013.126

[35] Zhenguo Li, Jianzhuang Liu, Shifeng Chen, and Xiaoou Tang. 2007. Noise robust spectral clustering. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

[36] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[37] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[38] Mahesh Mohan and Claire Monteleoni. 2017. Exploiting sparsity to improve the accuracy of Nyström-based large-scale spectral clustering. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 9–16.

[39] Prakash Nadkarni. 2016. Chapter 10 - Core Technologies: Data Mining and "Big Data". In *Clinical Research Computing*, Prakash Nadkarni (Ed.). Academic Press, 187–204. https://doi.org/10.1016/B978-0-12-803130-8.00010-5

[40] Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 14 (2001), 849–856.

[41] Beresford N. Parlett. 1998. *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, Philadelphia.

[42] Jella Pfeiffer and Franz Rothlauf. 2007. Analysis of greedy heuristics and weight-coded eas for multidimensional knapsack problems and multi-unit combinatorial auctions. In *Proceedings of the 9th annual Conference on Genetic and Evolutionary Computation*. 1529–1529.

[43] Farhad Pourkamali-Anaraki. 2020. Scalable Spectral Clustering With Nyström Approximation: Practical and Theoretical Aspects. *IEEE Open Journal of Signal Processing* 1 (2020), 242–256.

[44] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2016. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research* 17, 1 (2016), 4635–4666.

[45] Tomoya Sakai and Atsushi Imiya. 2009. Fast Spectral Clustering with Random Projection and Sampling. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 372–384.

[46] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.

[47] G. W. Stewart. 2002. A Krylov-Schur Algorithm for Large Eigenproblems. *SIAM J. Matrix Anal. Appl.* 23, 3 (2002), 601–614. https://doi.org/10.1137/S0895479800371529

[48] Alexander Strehl and Joydeep Ghosh. 2002. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal on Machine Learning Research (JMLR)* 3 (December 2002), 583–617.

[49] Zhiqiang Tao, Hongfu Liu, Sheng Li, and Yun Fu. 2016. Robust spectral ensemble clustering. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 367–376.

[50] Joshua Tobin and Mimi Zhang. 2021. DCF: An Efficient and Robust Density-Based Clustering Method. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 629–638.

[51] Nicolas Tremblay and Andreas Loukas. 2020. Approximating spectral clustering via sampling: a review. *Sampling Techniques for Supervised or Unsupervised Tasks* (2020), 129–183.

[52] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 4 (2007), 395–416.

[53] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek. 2009. Approximate Spectral Clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (Advances in Knowledge Discovery and Data Mining)*. Springer Berlin Heidelberg, 134–146.

[54] Christopher Williams and Matthias Seeger. 2001. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*. 682–688.

[55] He Xu, Lin Zhang, Peng Li, and Feng Zhu. 2022. Outlier detection algorithm based on k-nearest neighbors-local outlier factor. *Journal of Algorithms & Computational Technology* 16 (2022), 17483026221078111. https://doi.org/10.1177/17483026221078111

[56] Kai Zhang, Liang Lan, Jun Liu, Andreas Rauber, and Fabian Moerchen. 2012. Inductive kernel low-rank decomposition with priors: A generalized nystrom method. *arXiv preprint arXiv:1206.4619* (2012).

[57] Xiatian Zhu, Chen Change Loy, and Shaogang Gong. 2014. Constructing robust affinity graphs for spectral clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1450–1457.