Technische Universität München

TUM School of Social Sciences and Technology

Understanding moral wiggle room: Situational conditionalities and interindividual differences in prosociality

Fiona Maria tho Pesch

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Doris Holzberger

Prüfer*innen der Dissertation:

1. Prof. Dr. Anna Baumert

2. Prof. Dr. Susann Fiedler

Die Dissertation wurde am 17.04.2023 an der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 25.05.2023 angenommen.

Table of contents

General introduction **03**

Chapter 1 What's moral wiggle room? A theory specification and advancement **19**

Chapter 2 Attributional ambiguity reduces charitable giving by relaxing social norms **43**

Chapter 3 Conflict in willful ignorance: A mouse-tracking investigation **61**

Chapter 4 Who ignores - and why? **84**

General discussion **122**

References **137**

General introduction

In 2022, private German households gave €5.7 billion to charity, with 18.7 million people having given to charity on average 7 times a year (Deutscher Spendenrat e.V., 2023). Such charitable giving is therefore an important economic factor, as well as widespread in society. But why do people give to charity? The drivers of prosocial behaviors, such as charitable giving, but also helping, sharing, cooperation or punishment of unfair behavior, are still highly debated in the literature. However, these behaviors are vital for the functioning, not only of interindividual relationships, but also of societies at large. Cooperation, for example, enables human societies to achieve common goals by working in groups, pooling their resources and talents, and dividing labor accordingly (Henrich & Muthukrishna, 2021). The overarching aim of this dissertation is to isolate the role of individual and situational factors driving prosocial behavior in different contexts. Traditionally, psychological research has focused on identifying interindividual differences that predict prosociality on the one hand, such as differences in cognitive development (Kohlberg & Kramer, 1969), personality (Doris, 2002; Penner & Finkelstein, 1998; Thielmann et al., 2020) or identity (Aquino & Reed, 2002; Blasi, 1980; Narvaez & Lapsley, 2009), and on contextual factors that influence prosocial behaviors on the other hand, such as environmental cues (Bateson et al., 2006; Cramwinckel et al., 2013; Haley & Fessler, 2005), cognitive resources (Gino et al., 2011; Teoh & Hutcherson, 2022), or emotional reactions (Tangney et al., 2007). Both perspectives have proved to be very relevant for understanding drivers of prosociality. In this dissertation, I will combine both approaches in complementary ways to tackle the question of what drives prosociality.

Prosociality

People engage in a wide range of prosocial behaviors on a daily basis, such as helping strangers, volunteering for non-profit organizations, and giving to charity. Researchers from various fields, including psychology, philosophy, economics, and other social sciences have been interested in understanding why people choose to behave prosocially and what factors influence prosociality. Though scientific interest has been increasing in the last two decades, the question of what constitutes prosocial behavior is still up for debate, and its conceptualization is heterogeneous in the literature (Pfattheicher et al., 2022), ranging from definitions centering on intentions (Batson & Powell, 2003) to definitions emphasizing the importance of moral norms and societal expectations (Dovidio, 1984). Throughout this dissertation, I will refer to prosociality as behavior that is “costly to the actor and beneficial to the recipient” (S. A. West et al., 2011, p. 232), focusing on the consequences of behavior. As will be evident throughout the dissertation, it is important to distinguish between action and intention, as the intentions behind prosocial behavior are diverse, and the main subject of this dissertation. Note here that altruism as a concept is closely related to, but distinct from prosociality. Historically speaking, the term prosociality was coined by social scientists to describe the opposite of antisocial behavior (Pfattheicher et al., 2022). As such, it relates to a behavioral phenomenon. Altruism, by contrast, is the antonym of egoism, and thus describes a certain motivation for behavior. Though altruism most often leads to prosocial behavior, not all prosocial behavior is motivated by altruistic concerns (Batson & Powell, 2003).

When people enter a situation in which they have the option to behave prosocially, they face a tradeoff decision: They can either engage in prosocial behavior, which implies costs to themselves by profiting others, or they can choose to engage in selfish behavior, maximizing their own resources at the cost of others. Thus, prosocial behavior often requires the inhibition of self-interested behavior (i.e., resisting the temptation to choose the selfish option) to the benefit of others or society as a whole, as self-interest is at odds with other-regarding concerns. For this reason, researchers have conceptualized prosocial decision contexts as representing a conflict of competing motives (Baumeister & Alghamdi, 2015; Fishbach & Woolley, 2015; Locey et al., 2013; Sheldon & Fishbach, 2015). The resolution of this conflict is at the heart of economic game paradigms. Economic games are used to study and model economic behavior in controlled experimental settings. They are used to test hypotheses about how people make decisions and interact with one another. Research on prosocial behavior has long relied on different economic games such as the ultimatum game (Güth et al., 1982) and the dictator game (Forsythe et al., 1994) to investigate different drivers of prosociality. The dictator game is perhaps the most popular economic game used to study social preferences, fairness, and prosociality, which is likely due to its simplicity. In the game, one player, the dictator, is given a certain amount of money and must decide how much, if any, to give to the other player, the recipient. The dictator can choose to give any amount of money, from zero to the entire amount, to the recipient. The recipient receives whatever amount the dictator decides to give. The game is designed to test how much the dictator values prosociality and fairness, as well as how much weight they place on their own self-interest. A meta-analysis on the last 25 years of research on dictator games shows that dictators on average give 28.35% of their resources (Engel, 2011).

Why do people behave prosocially?

People behave prosocially, or in ways that benefit others at their own cost, for a variety of reasons. Evolutionary accounts have mainly focused on theories of kin selection (Hamilton,

1964) or reciprocity and group fitness (Nowak, 2006; Trivers, 1971). Meanwhile, researchers in psychology and economics rely on concepts such as preferences, norms, and emotions to explain prosocial behavior. In the following sections, I will present different theoretical accounts that explain prosociality. This list is by no means exhaustive, and is meant to introduce the accounts most relevant to this dissertation.

Social preferences

Standard economic theory assumes that the main principle guiding human behavior is rationality, and that rationality is motivated by pure self-interest. A rational person is assumed to know what is best for them and thus act in ways that maximize their utility (i.e., the subjective satisfaction or well-being that individuals derive from consuming goods or from engaging in certain actions or behaviors). However, as early as Adam Smith, the father of modern economics, it has been argued that humans are also interested in the well-being of others, though they “derive nothing from it, except the pleasure of seeing it” (A. Smith, 1759, p. 4). By now, it is undisputed that humans behave more prosocially than expected of a *homo economicus*. Theories trying to pacify the experimental evidence on prosociality with economic theoretical accounts led to the inception of social preferences, where the agent cares not only about their pure material outcomes, but also the consequences of their actions on others (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Fischbacher, 2002; Fehr & Schmidt, 1999). In other words, prosocial behavior is still conceptualized as maximizing the agent’s utility; this utility, however, is not only dependent on the agent’s payoffs, but also the well-being of others. There are different ways in which the well-being of others could factor into the utility of an agent. People may, for example, gain a *warm glow*, an expression describing the positive feeling or emotional reward that individuals experience when they engage in acts of kindness or prosociality. This would imply the existence of impure altruism (Andreoni, 1990), whereby people give to charity because they *feel good*. Alternatively, people could exhibit a form of inequality aversions (Fehr & Schmidt, 1999), speaking to the idea that individuals have a negative preference for inequality and would rather have a more equitable distribution of resources, even if it means giving up some of their own.

Social norms

Social norms can be understood as reflecting a group’s implicit rules and standards, describing what is collectively perceived as socially appropriate (Bicchieri, 2005; Cialdini & Trost, 1998). Generally speaking, most people have an intrinsic preference to conform to social norms, and are willing to sacrifice material gain in order to comply with them (Cappelen et al., 2007; Gächter et al., 2017; Krupka & Weber, 2009; López-Pérez, 2008). Norm compliance thus often drives behavior and decision-making. Social norms have been identified to be important for all kinds of social behavior, including prosocial behavior (Andreoni & Bernheim, 2009; Bénabou & Tirole, 2006; Kimbrough & Vostroknutov, 2016; Krupka & Weber, 2013), lying (Bicchieri et al., 2019; Gächter & Schulz, 2016), and costly punishment (Fehr & Fischbacher, 2004; Fehr & Gächter, 2000). As such, social norms are crucial to cooperation and the functioning of human societies (Henrich & McElreath, 2003). Social norms can be upheld by sanctions, such as being punished or ostracized for deviating from a norm, or rewards in the form of reputational gains for following a certain norm for example. In public good experiments, an economic game that models cooperation, it is typical to observe a steady decline of participants' cooperation over different trials. However, once the option to punish other players for free-riding is introduced, it has been shown that cooperation rates remain stable, at a high level (Fehr & Fischbacher, 2004). Punishment opportunities therefore have a decisive impact on upholding social norms. People also have been demonstrated to internalize social norms, thereby judging their own behavior according to its conformity to a given norm. When social norms are internalized, sanctions or rewards are administered by the individual themselves, in the form of experiencing guilt or pride (Burger et al., 2009; Gavrilets & Richerson, 2017; Kimbrough & Vostroknutov, 2016; Schwartz, 1977). Social norms can be measured using the social norm elicitation method (Krupka & Weber, 2013). It typically involves asking individuals questions about whether they believe a behavior is socially appropriate within a specific context. Crucially, participants are incentivized to choose the option that they believe is chosen by most other participants in the study, meaning that they earn money for correctly guessing the most common response. The responses are then used to estimate the social norm for that behavior in the specific context.

Cognitive dissonance

The classical psychological theory of cognitive dissonance by Leon Festinger (1957) suggests that individuals experience psychological discomfort or dissonance when their attitudes, beliefs, or behaviors are inconsistent with one another. To reduce this discomfort, people engage in various strategies, such as changing their beliefs, behaviors, or attitudes to restore consistency. Advancements of the theory concentrate on the idea that people experience cognitive dissonance when their behavior is in conflict with their self-concept (Aronson, 1992; Beauvois & Joule, 1996). When applied to prosociality, the theory suggests that individuals experience cognitive dissonance when their beliefs or attitudes towards prosocial behaviors are inconsistent with their actual behavior. For instance, if someone believes that it is important to help others but fails to do so in a specific situation, they may experience discomfort due to the inconsistency between their beliefs and their behavior. Konow (2000) used the concept of cognitive dissonance to explain sharing behavior in simple economic games, such as the dictator game. Agents are assumed to balance between, on one hand, their desire to maximize their own payoffs, and on the other, the experience of cognitive dissonance engendered by sharing less than they believe to be fair. Within this framework, prosocial behavior is thought to reduce cognitive dissonance, as people generally hold general prosocial attitudes and sharing norms. Importantly, people can also engage in self-deception to avoid cognitive dissonance while also maximizing their own payoffs (Konow, 2000). Indeed, individuals may also rationalize their failure to behave prosocially by justifying why they could not do so, such as by blaming external circumstances. Overall then, Festinger's (1957) Cognitive Dissonance Theory provides insight into the psychological processes underlying prosocial behavior, and how individuals may reconcile inconsistencies between their beliefs and actions.

Person perception/image

Any given choice one makes - including the decision to act prosocially or selfishly - not only creates material realities in their surroundings, but is also an expressive act, sending a signal about the agent’s preferences, motivations, and intentions. Behaving prosocially therefore enables people to establish and maintain a positive moral image. Research on image and identity suggests that people are generally motivated to maintain a positive moral image of themselves (Blasi, 1980; Dunning, 2007; Monin & Jordan, 2009), and several studies have shown the importance of image concerns for prosocial behavior (Andreoni & Bernheim, 2009; Ariely et al., 2009; Bagwell & Bernheim, 1996; Glazer & Konrad, 1996). Image concerns have subsequently been incorporated into models of decision-making, assuming that people want to signal their moral image to others and themselves (Andreoni & Bernheim, 2009; Bénabou & Tirole, 2006; Bodner & Prelec, 2003). For example, Bénabou and Tirole (2006) developed a theory of prosocial behavior that combines people’s prosocial motivations with their image concerns. The model suggests that prosocial behavior depends on the context. Making choices observable by others typically increases the likelihood of prosocial behavior. Reducing the transparency between actions and outcomes, on the other hand, decreases prosocial behavior, as it makes it harder to infer the agent’s intentions. In these cases, selfish behavior can also be attributed to the context, instead of having to connect selfish behavior to the agent’s intentions.

One can broadly distinguish between social and self-image concerns. Social image concerns describe a desire to be seen by others in a certain way. As most people want to be seen as moral and prosocial, social image concerns have a large impact on behavior in the prosocial domain. Distinctly, as much as people care about the opinion of others, they also care about their own self-image. Self-image concerns describe our desire to perceive ourselves as moral and prosocial. According to Bodner and Prelec (2003), agents may alter their behavior to maintain a positive self-image, a process known as self-signaling. They argue that because one cannot perfectly introspect the motivation underlying one's own behavior, a person may also adjust their behavior in order to manage their impression of themselves; As they cannot accurately introspect their motivations, they may end up distorting their behavior. Bénabou and Tirole (2006) see self-signaling as a way of influencing the beliefs of a future self, who may not remember the original reasoning behind the behavior. As such, self-image concerns have been hypothesized to drive prosocial behavior. However, distinguishing self- from social image concerns is challenging. Take the example of charitable giving. Research indicates that people are less likely to give to charity when given an excuse, such as not having noticed the option to give (Adena & Huck, 2020). While it is obvious that an excuse for not giving relieves social pressure, by weakening the signal that selfish behavior sends to others about the agent's intentions and motivations, it simultaneously allows the agent to protect their self-image, by engaging in self-deception strategies. In one of the few studies trying to tease apart the effects of self- and social image concerns on giving behavior, Grossman (2015) varied the information that would be revealed to others, and the likelihood of the individual's choice being implemented in a lab experiment. Results showed evidence of social image concerns but not self-image concerns. Other researchers, however, have found evidence of image effects in completely anonymous situations, and argue that only self-image can play out in such settings (Grossman & van der Weele, 2017).

Who behaves prosocially?

Unsurprisingly, the motivations I described above differ between individuals in relatively stable ways, and this is reflected in substantial interindividual variability in prosocial behavioral tendencies. While some people are willing to share their resources with others, a large share of the population consistently decides not to (e.g., Camerer, 2003; Engel, 2011; Sally, 1995). There is some consistency in terms of who behaves selfishly or prosocially across different decision contexts and across time (Baumert et al., 2014; Blanco et al., 2011; Galizzi & Navarro-Martinez, 2019; Haesevoets et al., 2015; McAuliffe et al., 2019; Peysakhovich et al., 2014; Yamagishi et al., 2013). In the last couple of decades, researchers have identified relevant personality traits that predict prosocial behavior in economic games. A meta-analysis of this research line recently highlighted the interindividual differences in Social Value Orientation, Honesty-Humility, and Guilt Proneness as the most relevant predictors (Thielmann et al., 2020).

Social value orientation (SVO) refers to an individual's dispositional tendency to prioritize the welfare of either oneself or others in social decision-making (e.g., McClintock, 1972; Murphy & Ackermann, 2014; van Lange, 1999). People with a pro-self orientation tend to prioritize their own interests and goals, while those with a pro-social orientation tend to prioritize the interests and goals of others. The most prominent measure of SVO is the slider task (Murphy et al., 2011), a behavioral and incentive-compatible measure. The task involves a series of scenarios, in which participants have to make dictator game decisions, distributing money between themselves and another participant. Meta-analytic evidence suggests a strong average correlation between SVO and prosocial behavior of $p = .26$ (Thielmann et al., 2020). Note, however, that SVO might well be such a good predictor for prosociality in economic games, due to its similarities in assessment methods, as the measurement of SVO relies on a series of game-like distribution decisions.

Honesty-Humility is one of the six broad factors of personality in the HEXACO Personality Inventory (Ashton & Lee, 2007), referring to the degree to which an individual is honest, fair, and genuine in dealing with others, versus the degree to which they are boastful, arrogant, and insincere (Ashton & Lee, 2007). There are different scales for measuring each of the six dimensions of the HEXACO model, varying in length (e.g., Ashton & Lee, 2009; K. Lee & Ashton, 2018). In game settings, Honesty-Humility shows an average correlation of $p = .20$ with prosocial behavior (Thielmann et al., 2020).

Guilt proneness refers to an individual's tendency to experience feelings of guilt in response to their own wrongdoing or moral transgression (Tangney et al., 2007). It is a personality trait that can be measured by self-report questionnaires, such as the Guilt and Shame Proneness scale (Cohen et al., 2011). Individuals who are high in guilt proneness tend to feel guilty even when they have not done anything wrong, and are more likely to take responsibility for their actions and make amends. In contrast, individuals who are low in guilt proneness tend to be less likely to feel guilty, and may not take responsibility for their actions as readily. Guilt proneness is considered as a dispositional characteristic that can influence the way people think, feel, and behave. In Thielmann et al's (2020) meta-analysis,, guilt proneness was positively related to prosocial behavior with an average correlation of $p = .22$.

When do people behave prosocially?

In addition to differences in personality influencing whether an agent decides to act prosocially or not, situational characteristics also factor in. There are several strands of research exploring different situational characteristics influencing prosociality. For example, research on empathy and prosociality suggests that the more we empathize with the entity we are asked to help or share with in a given situation, the more likely we are to act in prosocial ways (Batson et al., 2002). This can result in ingroup favoritism (Fiedler et al., 2018), and the identifiable victim effect (S. Lee & Feeley, 2016). Classic psychological research on the presence vs. absence of others has also demonstrated its influence on prosociality. On one hand, some research has suggested that having others observe one's own behavior leads to more prosociality (Bradley et al., 2018), while other investigations suggest it can also result in a diffusion of responsibility, leading to less prosociality (Darley & Latané, 1968).

Another domain of situational characteristics influencing prosociality includes factors that blur the relationship between intentions and behavior. Intentions have long been recognized as crucial to moral judgment (Greene et al., 2009; Waldmann et al., 2012), and person perception (Jones & Davis, 1965; Kelley & Michela, 1980): We judge others according to the assumed intentions behind their observable behavior. As most people strive towards being seen as moral and prosocial by others (Aquino & Reed, 2002; Dunning, 2007; Monin & Jordan, 2009; Rachlin, 2002), situational factors that blur the relationship between intentions and behaviors can reduce the (social) pressure to be prosocial, making prosocial behavior less likely. A research stream which centers on the concepts of moral wiggle room (Dana et al., 2007) shows that selfish behavior increases when there is reduced transparency between intentions and outcomes (Dana et al., 2007; Exley, 2016; Grossman, 2014; Grossman & van der Weele, 2017). The literature identifies several situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one's intention to be selfish. For example, introducing uncertainty as to whether the agent or a computer has made a certain distribution decision leads to more selfish choices (Dana et al., 2007; Regner, 2021); Similarly, outcome risk and ambiguity can be exploited to behave selfishly while appearing risk averse or risk seeking (Exley, 2016). The most studied form of moral wiggle room is willful ignorance, the phenomenon whereby allowing people to ignore the consequences of their own behavior for other entities leads to more selfish behavior (Dana et al., 2007; for a review see Vu et al., 2023). In their original conception of moral wiggle room, Dana and colleagues (2007) suggested a reduction in social pressure, or the availability of an excuse not to give as potential explanatory mechanisms. Other researchers advance feeling less guilty (Feiler, 2014; Garcia et al., 2020; Thunström et al., 2014) or less conflicted (Grossman, 2014; Lin & Reich, 2018; Matthey & Regner, 2011; Woolley & Risen, 2018) when choosing the selfish option under moral wiggle room, potentially because moral wiggle room protects one's self- or social image (Adena & Huck, 2020; Andreoni & Bernheim, 2009; Grossman, 2015; Grossman & van der Weele, 2017). Research on the perception of behavior under moral wiggle room supports the notion that it could serve as an effective excuse, by showing that selfish behavior was perceived as less socially inappropriate (Krupka & Weber, 2013) and punished less by others (Bartling et al., 2014; Conrads & Irlenbusch, 2013) when selfishness was enacted under chosen ignorance. In that sense, moral wiggle room seems to reduce the reputational costs for behaving selfishly.

Theoretical approach and research agenda

This dissertation aims at understanding the motivational foundations of prosocial behavior, more specifically of charitable giving. To this end, I examined the role of social norms, cognitive dissonance and image concerns for prosociality. My work has focused on the concept of moral wiggle room, as it allowed me to systematically study potential mechanisms of prosociality, by observing covariation of certain factors in situations with and without moral wiggle room. As such, moral wiggle room is not only an interesting concept to study in terms of the situational conditionality of prosocial behavior, but also to understand why people behave prosocially more generally.

Research in psychology often wants to know why people act in the way they do. To answer this question, social psychologists have traditionally focused on situational elements, while personality psychologists concentrated on interindividual differences in so-called traits (i.e., stable and enduring characteristics or patterns of behavior, thought or emotion that distinguish one individual from another). For a long time, researchers from these two research traditions have argued that either situations or persons have a stronger effect on behavior (Donnellan et al., 2009; Funder, 2008; Kenrick & Funder, 1988). Moving beyond this debate, interactionists acknowledge the interactive influences of situations and persons on behavior, moving toward a more complete understanding of why people do what they do. Researchers can also utilize person-situation interactions to learn about underlying motivations driving behavior. In the domain of moral wiggle room, for example, researchers have found that dispositional guilt proneness is related to prosocial choices in transparent, but not in intransparent situations (Regner, 2021), suggesting that once transparency is reduced, guilt proneness does not lead to interindividually different prosocial behavior. This might be seen as suggestive evidence that selfish behavior in an intransparent situation elicited less guilt. Generally, utilizing person-situation interactions can be a fruitful way of investigating the motivational foundations of prosocial decision-making.

Throughout my dissertation, I combined different research methods, encompassing theoretical approaches and empirical ones, including experiments, correlational studies, and process tracing techniques. I furthermore made use of insights from personality psychology, to explore the ways in which people systematically vary in behavioral tendencies, and combine these with a more traditional social psychological focus on situational influences on behavior, to gain a more in-depth understanding of why, when and who behaves prosocially. Next, I briefly introduce the research methods included in the following chapters, before providing an overview of each chapter in itself.

Theory specification

Theories of social psychology differ widely concerning their degree of specification. While some theories include a large degree of formal specifications and broadly accepted operationalizations (e.g., the theory of planned behavior, Ajzen, 1991) others do so to a much lesser degree (Smaldino, 2019). Various authors have suggested different approaches to improve theories through specification by formalizing verbal theories as sets of equations (Borsboom et al., 2021) or propositions (R. West et al., 2019). Some of the attempts concern identifying and specifying core predictions of a theory and separating them from auxiliary assumptions (Glöckner & Betsch, 2011; Lakatos, 1968; Oberauer & Lewandowsky, 2019). The specification of core propositions allows conducting decisive tests of competing theories, which could foster sciences to advance faster by an order of magnitude (Platt, 1964).

Glöckner and Betsch (2011) aimed to improve theories and specifically their *empirical content* (Popper, 1934) by developing standards for theory specification. A well-specified theory consists of one or more specific propositions regarding relationships or causal effects among concepts (Glöckner & Betsch, 2011). When specifying unidirectional causal effects, a proposition consists of two elements: (1) the precondition or antecedence and (2) the result or consequence. Often a theory consists of more than one proposition and individual propositions may be connected to each other (such as when a mediating mechanism is proposed). All concepts that appear in the antecedence as well as in the consequence have to be defined, in a way that makes them unambiguous and testable. This step is also crucial for avoiding any tautologies in the theory. A good theory specification also offers insights on how to measure and manipulate these concepts (i.e., operationalizations of concepts).

Lastly, a theory may be based on additional, often implicit, boundary conditions that have to be met for the theory to be applicable. These mostly represent a tradeoff between specificity and generalizability, as they hold information regarding the subgroup of individuals or situations the theory applies to. Such boundary conditions should be specified (and thus made explicit) as auxiliary assumptions of the theory to avoid inefficiencies in theory testing. Verbal theory specification can be seen as a useful tool for researchers across disciplines, as it can reach a broad readership due to its simplicity, and thus can foster interdisciplinary collaborations. It is therefore a useful approach for specifying the theory of moral wiggle room, as researchers from economics, psychology and philosophy work closely together on this topic. Verbal theory specification can help to heighten efficiency and enables researchers to strive towards more unified theoretical frameworks, which can function as catalysts of knowledge about human behavior and its underlying psychological mechanisms.

Economic games

Economic games are laboratory experiments that are used to study human decision-making and behavior, and designed to mimic real-world decision contexts. They are a valuable tool for researchers in economics, psychology, and other social sciences because they allow for controlled manipulation of variables and offer a way to test hypotheses about human behavior. The underlying logic of economic games is based on game theory, which is a branch of mathematics that is used to model and analyze human behavior. Game theory provides a set of mathematical tools for analyzing the interactions between decision-makers, testing hypotheses about human behavior, and investigating the factors that influence decision-making. Economic games differ from traditional psychological methods because they rely on incentivizing choices by paying participants real money according to their choices, instead of simply asking participants to self-report how they would behave in certain situations. Incentivizing participants' choices in the lab thus forces them to put their money where their mouth is, creating more externally valid decision settings in the lab (for a discussion on the usefulness and impact of incentivisation, see Hertwig & Ortmann, 2001).

Mouse-tracking

Research on human decision-making has long concentrated on observing human behavior in experimental settings. However, this approach is limited in terms of investigating the process leading up to a decision. Process-tracing techniques go beyond the mere observation of choice as the behavioral outcome by tracking the temporal development of cognitive processes, such as information acquisition or preference formation. While eye-tracking (Russo, 2019), Mouselab (Willemsen & Johnson, 2019), or Flashlight (Schulte-Mecklenbeck & Huber, 2003) track the processes of information acquisition, mouse-tracking also sheds light on preference formation (Schulte-Mecklenbeck et al., 2019).

To gain access to this information, mouse-tracking records hand movements indirectly by continuously sampling the cursor position of a computer mouse while participants decide between options that are spatially separated on a screen (Freeman & Ambady, 2010). Most studies use a forced choice, two-alternative paradigm. The underlying assumption of mouse-tracking is that motor movements can be used as an indicator for cognitive processing (Spivey & Dale, 2006). More specifically, it is assumed that mouse movements mirror the tentative commitments to the different choice options during the decision process (Freeman et al., 2011). When an agent strongly considers an option, this option exerts a *pull* of the mouse movement in its direction. Thus, mouse trajectories can be seen as an indicator for how much each option was considered during the decision process (Koop & Johnson, 2011). The most relevant indicator for the work presented in this dissertation is the curvature of the response trajectory. A greater deviation towards non-chosen options in the trajectory indicates a higher level of consideration for that option (Freeman & Ambady, 2010; Spivey & Dale, 2006). Thus, the degree of curvature can be seen as an indicator of the overall cognitive conflict experienced during the decision-making process.

Empirical evidence supports the interpretation of response trajectory curvature as an indicator of cognitive conflict during decision-making (Spivey et al., 2005): Participants were asked to select one of two images while the distractor image displayed either a phonologically similar or dissimilar word. The results showed that the response trajectories were more curved towards the phonologically similar distractor, indicating a higher level of attraction and cognitive conflict (Spivey et al., 2005). Later studies further validated the use of response trajectory curvature in understanding cognitive processes in various domains such as decision-making, language, social cognition, and learning (Dshemuchadse et al., 2013; Freeman et al., 2011; Koop, 2013; Koop & Johnson, 2011, 2013). The response dynamics approach has also been supported by neurophysiological evidence (Spivey, 2007).

Cognitive conflict as measured by mouse-tracking has been related to prosocial behavior in the domain of social dilemmas (Kieslich & Hilbig, 2014). Participants playing simple two-person social dilemma games with two options (cooperation and defection) showed less cognitive conflict in cooperation, meaning that response trajectories were more curved towards the non-chosen option when individuals defected than when they cooperated. Interestingly, this effect was driven by participants high in Honesty-Humility, indicated by an interaction of Honesty-Humility and choice when predicting mouse curvatures (Kieslich &

Hilbig, 2014). This can be seen as further evidence for the usefulness of an interactionist view on research in the domain of prosociality.

Overview of chapters

In **Chapter 1**, I provide a formal verbal theory specification for the theory of moral wiggle room. As the remainder of this thesis centers around the notion of moral wiggle room, this specification lays the foundations for the following chapters. More generally speaking, the theory of moral wiggle room, though very popular amongst researchers both from psychology and economics, has not been formally specified as of yet. The literature on moral wiggle room has been characterized by a very heterogeneous reading of the concept, which arguably is detrimental for scientific efficiency.

Chapter 2 identified another form of moral wiggle room, and investigated potential psychological mechanisms driving the effect of moral wiggle room on prosocial behavior. Borrowing from a classic paradigm in social psychology (Snyder et al., 1979), this line of studies investigated whether people take advantage of attributional ambiguity as an excuse to be selfish when facing charitable requests. In four experiments (*N* = 2,147), participants faced a binary choice between a prosocial and a selfish option. I manipulated whether the charities associated with each option were the same or different, and counterbalanced which charity was associated with the selfish option. I explored the role of self-image concerns and social norms for the effect of attributional ambiguity on charitable giving.

In **Chapter 3**, I utilized the process tracing technique of mouse-tracking to investigate the role of cognitive conflict for willful ignorance, a widely-studied form of moral wiggle room (Dana et al., 2007; Vu et al., 2023). Research on willful ignorance in prosocial decision-making suggests that people ignore information that would otherwise be instrumental to their decisions. This behavioral effect has been hypothesized to be driven by a desire to avoid cognitive conflict when behaving selfishly (Grossman, 2014; Lin & Reich, 2018; Matthey & Regner, 2011; Woolley & Risen, 2018). In a fully incentivized experiment (*N* = 210), cognitive conflict was implicitly measured by tracking people's mouse trajectories while they made binary decisions on a computer screen, distributing money between themselves and a charity. In the first decision, they had the option to ignore the consequences of their choice for the charity. Subsequently, participants made 18 choices in which their mouse movements were tracked, 12 of which were unaligned trials (i.e., the option with the higher payoff for the participant was giving the charity a lower donation). Through the analysis of participants' mouse trajectories, the level of cognitive conflict experienced while making binary decisions was inferred. Combining these behavioral measures with interindividual differences in Guilt Proneness, Social Value Orientation and Honesty-Humility provided further insights into who and why people engage in willful ignorance.

Chapter 4 takes a critical perspective on selfishness as the main mechanism behind willful ignorance, addressing the question of whether willful ignorance is indeed driven by wiggling-related motivations. Traditionally, willful ignorance has been conceptualized to be driven by motivations such as image concerns or other wiggling-related factors (Dana et al., 2007; Grossman & van der Weele, 2017; Vu et al., 2023). I critically investigated other potential motivations for ignorance, such as tradeoff aversion and inattention, by intraindividually varying different aspects of the decision context. Within an online study with four waves of data collection, subjects made multiple decisions following the general setup of Dana et al. (2007), whereby the receiving parties and the kind of information that can be ignored were manipulated. Observing multiple ignorance decisions from the same person and combining those with personality measures allowed me to shed light on the underlying motivations for ignorance.

Within the **General Discussion**, I highlighted the interconnections between the different chapters, and what they separately and collectively can tell us about the theory of moral wiggle room. In a second step, I also drew conclusions from my empirical investigations for the motivations underlying prosociality more generally. Finally, I attended to the limitations and potential future directions that arise from the studies presented in this dissertation.

Chapter 1:

What’s moral wiggle room? A theory specification and advancement

This chapter is based on Fahrenwaldt, A.*, tho Pesch, F.*, Fiedler, S., & Baumert, A. (under review at the Journal of Judgment and Decision-Making). What’s moral wiggle room? A theory specification and advancement.

*The first authorship is shared by Alina Fahrenwaldt and Fiona tho Pesch, as they contributed equally to the paper.

Abstract

The term *moral wiggle room* (MWR) describes the phenomenon that the likelihood for selfish behavior is increased if situational characteristics reduce the transparency between behaviors and their consequences. It is based on the idea that, without MWR, prosocial behavior is not only the result of prosocial preferences, but also of a desire to appear like a prosocial person. Studies testing the effect of MWR reveal substantial heterogeneity in the understanding of core concepts and boundary conditions, the implemented operationalizations and size of the effect. We argue that systematic theory specification is needed to avoid ambiguities and strengthen this research field. Using a novel method of formal verbal theory specification, we outline the original postulation of MWR and identify its loopholes. On this basis, we refine the original postulation by fully specifying all concepts and their appropriate operationalizations as well as necessary auxiliary assumptions. Most importantly, we advance it as a (fully testable) theory by redefining the concept of MWR, by specifying three underlying psychological mechanisms of the behavioral MWR effect (i.e., anticipated image damage, perceived social norms, and anticipatory emotions) and the role of interindividual differences in the susceptibility to MWR (i.e., the joint effects of dispositional other-regarding preferences and social image concerns). Lastly, we relate it to existing theories and draw a roadmap for future work on the theory and its empirical tests. With our contribution, we hope to stimulate more rigorous and efficient research on the effect of MWR and provide an example for the utility of formal verbal theory specification in general.

Keywords: moral wiggle room, theory specification, ignorance, prosocial behavior, social image, self-image, social preferences, norms, guilt

Authors' note: For the appendix, please see https://osf.io/b7ezp/?view_only=b5a2d1178acc47cc9d5658fb29aa4391.

Introduction

The nature of human social behavior has fascinated philosophers and scientists for centuries. In the past, behavioral theorists have postulated that prosocial behavior arises because some people have a preference for prosocial outcomes, meaning they receive utility from behaving prosocially (e.g., sharing, donating, and helping; Andreoni & Miller, 2002; Charness & Rabin, 2002; Fehr & Schmidt, 1999). This assumption has been challenged by studies demonstrating the situational conditionality of prosocial behavior. The *moral wiggle room* (MWR) framework proposes that prosociality is partly driven by a desire to appear prosocial instead of genuine prosocial preferences over outcomes alone (Dana et al., 2007, DWK hereafter). In their original postulation, DWK defined MWR as a reduction in the transparency of the link between behavior and outcomes, and they argued that this intransparency allows people to behave selfishly without appearing selfish. In line with their predictions, they discovered reduced prosociality in situations providing MWR. Albeit intriguing, we argue that the original postulation of MWR is underspecified and its formulation holds some critical loopholes and inconsistencies. For example, the psychological mechanism proposed to be underlying the behavioral MWR effect necessitates the implicit assumption that reduced transparency obfuscates the signal which the outcome of a selfish behavior sends to others about one's intentions. Such lack of explicitness in a theory's formulation is problematic for rigorous testing and may render the scientific process inefficient. As we demonstrate in the present paper, formal theory specification is a powerful tool to solve these issues and allow for theoretical advances.

In the following, we provide a brief outline of the existing research on MWR and explain why we need a formal verbal specification. Subsequently, we specify the original postulation of MWR by DWK and discuss its loopholes and problems. We advance the theory by incorporating solutions to these loopholes, providing precise definitions, boundary conditions and operationalizations of MWR. Moreover, we specify three underlying psychological mechanisms of MWR as well as the role of interindividual differences in the susceptibility to MWR. Finally, our paper resumes with a general discussion, summarizing our key insights from the theory specification process and highlighting empirical questions to be addressed by future research.

Brief outline of research on MWR

Since its inception, the concept of MWR has sparked great research interest. More than 1500 papers have mentioned the term *moral wiggle room*, and the original paper (DWK) has been cited over 1690 times according to google scholar (retrieved December 7th 2022). Within this literature, one can identify different operationalizations of MWR. The operationalization receiving most attention is *strategic ignorance*, encompassing experimental setups designed to demonstrate that people avoid certain information and subsequently are more likely to behave selfishly (DWK; Bartling et al., 2014; Bell et al., 2017; D’Adda et al., 2018; Ehrich & Irwin, 2005; Grossman, 2014; Matthey & Regner, 2011; Momsen & Ohndorf, 2020). Another operationalization of MWR is the introduction of *uncertainty between behavior and outcome*, meaning experimental manipulations of the recipient’s ability to know whether a decision has been made by the agent or by another entity (DWK). Furthermore, it has been proposed that outcome risk and ambiguity can be exploited to behave selfishly while appearing risk averse or risk seeking (Exley, 2016). Additional study designs seem connected, but have not been linked to the term MWR so far. One example are designs introducing information asymmetry where the recipient does not know the initial endowment of an agent, hindering judgment whether the agent was fair or not (Ockenfels & Werner, 2012). Another example are studies testing default and omission effects, in which agents can plausibly claim to have missed the chance to choose prosocially (Gärtner & Sandberg, 2017). All of these operationalizations of MWR have in common that they reduce prosociality in comparison to a more transparent baseline condition. For a list of operationalizations used in research on MWR, see Appendix A, Table A1.

Why and how to achieve a formal verbal theory specification of MWR

This diversity in operationalization could support comprehensive theory tests by focusing on different conceptual angles. However, a review of the literature indicates that researchers not only diverge in their operationalization of the MWR concept, but also that there is substantial heterogeneity in their understanding of the concept itself. For example, while DWK proposed that the label MWR describes certain situational characteristics, some employ it for any kind of justificatory cognition for immoral behavior (e.g., D’Adda et al., 2018). Heterogeneity is also evident in the researcher’s understanding of the postulation’s specificity (i.e., boundary conditions for its application). For example, some generalize the notion of MWR to reciprocal (e.g., van der Weele, 2014), strategic (e.g., Bolton et al., 2019) or even purely vicarious decision-making (Cerrone & Engel, 2019; for a list of decision settings in which MWR was tested, see Appendix A, Table A1).

One reason for this heterogeneity may be that the notion of MWR has never been formally specified in terms of a theory. The resulting divergence in the understanding of the central concepts and boundary conditions is problematic as it yields a lack of comparability, interpretability, and replicability of research findings (Camerer et al., 2018; Smaldino, 2019), and thus scientific inefficiency. These issues can be prevented by improving the specification of theoretical models, including their core propositions, definitions of concepts, operationalizations, and boundary conditions (e.g., Asendorpf et al., 2016), thus offering a roadmap efficient testing (e.g., Glöckner & Betsch, 2011; Gollwitzer & Schwabe, 2020). Glöckner and Betsch (2011) introduced standards for such a verbal theory specification. Their approach aims at increasing the *empirical content* of theories (i.e., the clarity of predictions and avoidance of contradictions and tautologies; Popper, 1934). They suggested that a well-specified theory should consist of a finite set of clear-cut propositions regarding relationships or causal effects among concepts, which together fully describe the theory.

When specifying unidirectional causal effects, a *proposition* consists of two elements: *antecedence* and *consequence* (written as if-then-statements). Often, a theory consists of multiple (interconnected) propositions (i.e., mediating mechanisms). *Concepts* appearing in the set of propositions have to be defined in an unambiguous and testable manner. The propositions link concepts through logical (AND, OR, etc.) operations. A good theory specification also offers insights into how to measure and manipulate these concepts, thus exemplifying the concepts' *operationalizations*. Lastly, any boundary conditions for the theory to be testable should be made explicit as *auxiliary assumptions*. They are necessary to isolate the effects of interest and rule out potentially confounding factors. Such auxiliary assumptions also concern the tradeoff between *specificity and generalizability*, as they hold information regarding the subgroup of people or situations that the theory applies to. Additionally, specifications should identify critical properties allowing for theory falsification.

We argue that a specification of MWR, following the standards proposed by Glöckner and Betsch (2011), will reduce existing and prevent future misunderstandings in the literature. Moreover, it serves to identify the most relevant open questions to be addressed by future research.

Specification of the original postulation of MWR

The original paper on MWR by DWK included only a rudimental verbalization of their theoretical assumptions, but a specification of propositions and core concepts can be deduced from their employed study designs. This specification shows that the original paper left critical aspects underspecified. Therefore, we present it directly together with important loopholes.

Propositions

At the heart of the postulation lies the proposition that in situations containing MWR (versus no MWR) there will be a higher likelihood of selfish behavior (see Table 1, Prop. no. 1). This behavioral effect of MWR is theorized to be mediated by agents having an "excuse" or "justification" not to give (DWK, p. 69) or "feeling [less] compelled to give" (DWK, p. 77-78). The authors use these terms interchangeably with a change in "norms and constraints" (DWK, p. 78). Specifically, they propose that the fairness norm could be perceived as less relevant in terms of less binding or less important in comparison with competing norms. The authors additionally state that other mechanisms could be possible, without further specifying them.

Table 1

Propositions derived from the original paper on MWR (DWK)

Prop. no.	Antecedence	Consequence
1	IF MWR (versus no MWR)	THEN higher likelihood of selfish behavior
2a	IF MWR (versus no MWR)	THEN reduced relevance of fairness norms and constraints (i.e., not feeling compelled to give or having an excuse or justification not to give)
2b	IF reduced relevance of fairness norms and constraints (i.e., not feeling compelled to give or having an excuse or justification not to give)	THEN higher likelihood of selfish behavior

Concept definition and operationalization

Selfish behavior. Selfish behavior is defined as decisions maximizing one's own profit while disregarding other people's payoff. This is operationalized as a binary decision with one option profiting the agent more and the recipient less compared to a second option which is more egalitarian.

MWR and no MWR. MWR is defined as situational characteristics that remove the transparency between (selfish) behavior and outcomes. Specifically, the authors speak of transparency as the "commonly known one-to-one mapping between the [agent's] actions and the outcomes to both parties" (DWK, p. 69). MWR was operationalized in three different ways: by utilizing treatments termed (i) hidden information, (ii) plausible deniability and (iii) multiple dictator (see Appendix A for a detailed description of the three original treatments). No MWR consequently describes situations with full transparency between behavior and outcomes (i.e., the baseline setting).

➔ **Loophole 1: Inconsistency between definition of MWR and mechanism proposition**

According to its original definition, MWR makes it more difficult to infer whether an observed *outcome* resulted from the agent's *behavior*, i.e., whether the agent can be held accountable for the outcome. At the same time, a change in norms and constraints is postulated to be the psychological mechanism driving the behavioral MWR effect (Props nos 2a&b). However, reduced accountability does not imply changed norms and constraints, and accountability alone can liberate the agent to behave selfishly, even if norms and constraints remain unchanged (Krysowski & Tremewan, 2021). The proposed norms mechanism makes more sense with a redefinition of MWR. Specifically, MWR could describe situational characteristics which make it difficult to infer an agent's *intentions* behind a behavior by allowing for other plausible reasons for the behavior (e.g., being short on time, not wanting to be nosy, being overwhelmed by the decision). In such situations, third-party observers of the agent's behavior as well as the agents themselves may perceive a change in the relevant social norms such that selfish behavior is less socially inappropriate.

Loophole 2 : Unsuitable operationalizations of the concept MWR

The *multiple dictators* treatment presented in the original paper is at odds with the definition of MWR. Specifically, in this treatment, the prosocial outcome for a passive recipient is implemented if one member of a group of agents chooses this option (over an agent-profiting but recipient-disadvantaging outcome). Thus, intransparency only pertains to the case of a fair outcome. In contrast, if all agents decide selfishly, every agent's *behavior* is clearly inferable from the implemented unfair outcome. Moreover, in this case, the agents' selfish *intentions* are also directly inferable. Thus, increases in selfish behavior in this treatment may be explained solely by effects of diffusion of responsibility (Darley & Latané, 1968), but not by MWR as defined in DWK.

Relevance of fairness norms and constraints. The original theory contained no clear definitions and operationalizations of the concept. This also applies to the synonymously used concepts "excuse", "justification", and "feeling [less] compelled to give".

➔ **Loophole 3: Lack of clear definition and operationalization of mechanism concepts**

In order to construct appropriate operationalizations it needs to be clarified whether MWR affects behavior through an objective change in the prevailing social norms (i.e., changes in what most people find appropriate) or through a change in an agent's perception of these prevailing norms.

➔ **Loophole 4: Differentiation of psychological mechanisms**

To allow for falsification of the theory, there needs to be clarity on the (differences between) underlying psychological mechanisms. DWK propose that MWR is effective, because it provides the agent with "excuse[s]" or "justifications" for not giving and that agents do not feel compelled to give in situations with MWR, but this is used interchangeably with the idea that there is a change in (the bindingness of, or availability of competing) norms and constraints. However, psychologically, perceptions of norms are not the same as feelings. Moreover, DWK state that the proposed psychological mechanism is only "one way" to account for their results. This is problematic, as allowing for other, not-specified mechanisms renders the theory non-falsifiable.

Auxiliary assumptions

In the original paper, no auxiliary assumptions were specified. However, four assumptions regarding the decision structure can be derived from the design of the experimental setups and their additional explanations.

1. The decision must have consequences for oneself and others, and parties' interests must be conflicting.
2. MWR must not restrict an agent's choice (i.e., their ability to implement the fair outcome).
3. The effects of MWR are only testable in non-strategic, unilateral interaction.[1]
4. There are two independent and sometimes conflicting motives influencing social behavior in the population: an agent's preferences over payoff distributions AND an agent's self- or social image concerns.[2]

[1] Generally, social decisions may additionally be driven by concerns for reciprocity and recipients' intentions. In such settings the proposed main driver of the effects of MWR (i.e., "not feeling compelled to give", DWK, p. 68) is no longer clearly separable from other, more rational concerns (e.g., economic disutility from potential punishment by the recipient). DWK highlighted that they chose to test the effect of MWR in settings which preclude such concerns.

[2] Note that DWK did not use any specific term for this concept. However, we believe that the term "image concerns", which is common in the literature on moral and social behavior, captures best what DWK mean when

➔ **Loophole 5: No explicit specification of agents' available action space**
So far, it is not clear whether the behavioral effect of MWR only applies to binary decision settings (pitting a selfish against a fair outcome) or also to settings offering multiple options, including giving that exceeds fairness.

➔ **Loophole 6: Lack of clear definitions and operationalizations of the motivations**
The concepts of *image concerns* and *preferences over payoff-distributions* are not sufficiently specified. This also gives room for theoretical inconsistency concerning the relevance of self-image versus social image concerns in MWR. On the one hand, MWR is conceptualized as a reduction in the "*commonly* known one-to-one-mapping" (DWK, p. 69) in the original theory, suggesting a focus on the social image. On the other hand, it is proposed that MWR is effective, when people are motivated by selfish distributional preferences but "do not want to appear selfish - either to themselves or others" (DWK, p. 68). Thus, it is unclear whether the original theory proposes that manipulations protecting only one's self-image OR only one's social image in case of selfish behavior each effectively provide MWR, or that MWR manipulations should primarily protect one's *social* image, but can, as a side product, also protect one's self-image.

➔ **Loophole 8: Interindividual differences**
From the original paper we derived the auxiliary assumption (no. 4) that the combination of two motives (preferences over payoff distributions and image concerns) determines the effect of MWR. This suggests that there could be differential effects of MWR on social behavior, depending on how important each of these two motives is for a specific agent in a certain distribution context. The original studies revealed such heterogeneity in MWR effects[3], which should be incorporated theoretically.

Theory advancement

Incorporating the identified loopholes we propose the following fully specified theory of MWR. Lists of all propositions, precise definitions and operationalizations (resolving

they state that many people "do not want to appear selfish" (p. 68). Other readers may be more familiar with the term "signaling", which we would like to treat as a synonym for "image concerns".

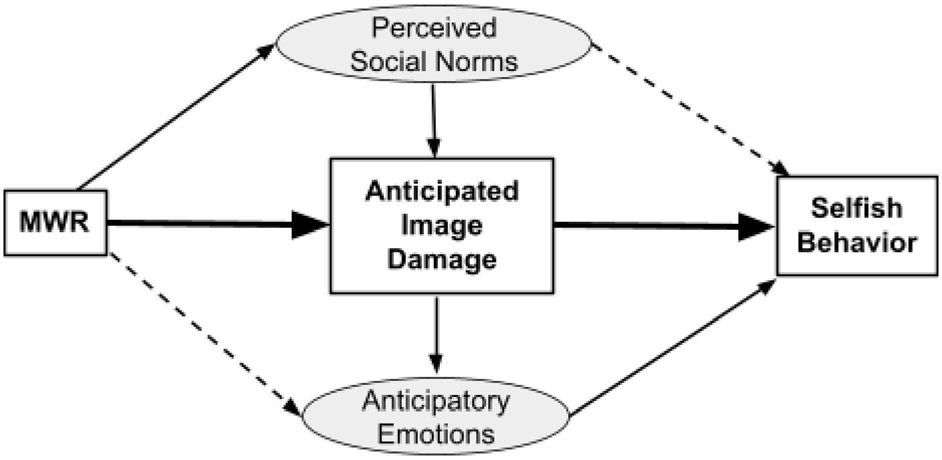
[3] In DWK *even without* any MWR, roughly one quarter of participants still behaved selfishly, and *even with* MWR, roughly one third of their participants decided against the selfish option. Therefore, only a fraction of people appeared to have been affected by MWR.

Loopholes 3 & 6) as well as auxiliary assumptions of the theory advancement can be found in Appendix C.

The behavioral proposition remains unchanged as compared to the original postulations (see Table 2). Remedying the issue of underspecification of the involved psychological mechanism (resolving Loophole 4), we extend the original postulations (see Table 1) by specifying three interrelated psychological mechanisms underlying MWR. These include (a) anticipated damage to one's social image[4], (b) perceived social norms, and (c) anticipatory emotional reaction to the situation (see Table 2; see Figure 1 for their interrelations).

Figure 1

Proposed psychological mechanisms underlying the effect of MWR on social behavior. Style of arrow lines indicates strength of effects when accounting for interrelations (bold line = main mechanism; continuous line = indirect effects; dotted line = weakest effects when considering respective other pathways).



Overview of the psychological mechanisms underlying the effect of MWR

Image mechanism: Change in anticipated social image damage. We propose the main psychological mechanism underlying the behavioral MWR-effect to be agents' reduced anticipation of damage to their social image in case of selfish behavior under MWR (see Table 2, Props 2a & 2b, resolving Loophole 6). The idea that agents might factor in anticipated social image damage when deciding whether to behave selfishly has been

[4] Meaning MWR may take effect, because it reduces the damage the agent anticipates for their social image in case of selfish behavior (DWK, 2007; but also see van der Weele et al., 2010). This was derived from the auxiliary assumption no.4 of the original specification, and DWK's claim that agents "do not want to appear selfish" (DWK, 2007, p. 68).

considered before (Exley, 2016). Whenever situations reduce the inferability of one's intentions from observed outcomes (see our definition of MWR), selfish behavior can be expected to have less negative effects on one's social image (e.g., Grossman & van der Weele, 2017) and individuals may behave more in line with their actual (selfish) preferences (DWK).

Normative Mechanism: Change in perceived social norms. Similarly to the original theory, we propose that MWR may take effect by changing perceived social norms, such that selfish behavior is perceived as less socially inappropriate (see Table 2, Props 3a & 3b). Social norms reflect a group's (implicit) rules and standards and are consequently often used as behavioral decision-making heuristics (Bicchieri, 2005). MWR changes (perceptions of) the prevailing norm. For example, choosing the self-profiting option after deciding to ignore relevant information about others' payoffs was perceived as less socially inappropriate by observers (Krupka & Weber, 2013) and recipients (Bolton et al., 2019; Grossman & van der Weele, 2017) compared to knowingly choosing the selfish option in these experiments. It also resulted in less ultimatum game rejections (Conrads & Irlenbusch, 2013), and lower third-party punishment (Bartling et al., 2014).

Emotional mechanism: Change in anticipatory guilt. MWR may also be effective by changing the agent's emotional reaction to the decision situation (see Table 6). Specifically, we propose that MWR could affect behavior via the anticipatory experience of guilt, which precedes and thus can inhibit behavior (Table 2, Props 4a & 4b). Anticipatory guilt has been closely connected to moral standards and has been shown to inhibit selfish or immoral behavior (Tangney et al., 2007). Indeed, some authors have suggested an anticipatory feeling of guilt for selfish behavior in transparent giving situations (i.e., situations without MWR; Feiler, 2014; Garcia et al., 2020). Vice versa, reduced selfishness-related anticipatory guilt in non-transparent decision settings might play a role in mediating the effect of MWR (e.g., Thunström et al., 2014).

Interrelations of the three psychological mechanisms

We propose the change in anticipated image damage to be the main psychological mechanism through which MWR affects behavior. It is likely interlinked with the other two psychological mechanisms (see Figure 1 and Table 2, props. nos. 5a & 5b).

Normative Mechanism and Image Mechanism. Adhering to social norms may be used as a signal of one's own morality, in an attempt to create a positive social image (Andreoni & Bernheim, 2009), while not adhering to social norms may result in anticipated image damage. Consequently, we propose that a MWR-induced change in perceived social norms also reduces anticipated image damage.

Emotional Mechanism and Image Mechanism. Guilt has been associated with one's anticipated social image (Larson & Capra, 2009). For the theory of MWR, we propose that reductions in anticipatory guilt may result from reduced anticipated social image damage. This is plausible because anticipatory guilt is thought to result from appraisal of immoral actions attributed to oneself.[5]

Propositions

Table 2

Main propositions of the theory advancement

Prop. no.	Antecedence	Consequence
1	IF MWR (versus no MWR)	THEN increase in selfish behavior
2a	IF MWR (versus no MWR)	THEN decrease in anticipated damage to one's social image by selfish behavior between these situations
2b	IF decrease in anticipated damage to one's social image by selfish behavior between situations	THEN increase in selfish behavior between these situations
3a	IF MWR (versus no MWR)	THEN change in perceived social norms (decrease in perception of selfish behavior as socially inappropriate between these situations)
3b	IF change in perceived social norms (decrease in perception of selfish behavior as socially inappropriate between situations)	THEN increase in selfish behavior between these situations
4a	IF MWR (versus no MWR)	THEN decrease in anticipatory guilt between these situations
4b	IF decrease in anticipatory guilt between situations	THEN increase in selfish behavior between these situations

[5] Note that this form of guilt is more in line with the "guilt-from-disapproval" rather than the "guilt-from-disappointment" type that has been outlined by (Hauge, 2016). This also sets our ideas apart from those on empathy-avoidance explaining reductions in prosociality (Shaw et al., 1994).

5a	IF change in perceived social norms (decrease in perception of selfish behavior as socially inappropriate between situations)	THEN decrease in anticipated damage to one's social image by selfish behavior between these situations
5b	IF decrease in anticipated damage to one's social image by selfish behavior between situations	THEN decrease in anticipatory guilt between these situations

Note. Since all concepts appearing in the propositions are operationalized to be estimated at the population level, all proposed links should be viewed as probabilistic rather than deterministic.

Falsification of the theory. Note that the behavioral and the mechanism propositions can be tested independently and the behavioral proposition does not rely on the mechanism proposition to be confirmed. However, if proof of the behavioral proposition (prop. no. 1) is not accompanied by proof of at least one of the three psychological mechanisms (props. nos. 2a-4b), this would constitute a falsification of the theory of MWR. Since the interrelation propositions (props. nos. 5a and 5b) only add a theoretical layer, lack of proof for these propositions would not constitute a falsification of the whole theory.

Concept definition and operationalization

For a list of possible operationalizations of all concepts, see Appendix C, Table C2. We exclude any unsuitable operationalizations (e.g., the *multiple dictators* treatment for inducing MWR, resolving Loophole 2).

Selfish behavior. We adapt the definition from the original theory of *selfish behavior* as choosing a (more) selfish distribution option over one or several (more) prosocial distribution option(s). This definition allows for operationalizing selfish behavior as a binary, ordinal, or continuous variable at the individual level. Here, a (more) selfish option yields a higher payoff for the agent and a lower payoff for the recipient, while (more) prosocial options are always less profitable for the agent but reduce inequality between agent and recipient. Note that an increase in selfish behavior is estimated at the population level, either as an increase in the likelihood of selfish behavior (estimated from choice frequencies) or the average degree of selfish behavior (for continuous operationalizations).

MWR and no MWR. We redefine MWR as situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one's intention to be selfish. Thus, the focus lies on whether the agent's selfish *intention* can be inferred from the observable outcomes, and not whether the agent's *behavior* can be observed or inferred from the outcomes. No MWR consequently describes situations with full transparency between intention and outcomes (i.e., the baseline setting).

Logically, this definition includes intransparency between behavior and outcomes: if an agent’s behavior is unknown, the intention behind it cannot be inferred either. However, with this redefinition, MWR can exist (and take effect) even when an agent’s behavior is observable. This definition is more consistent with the proposition of psychological mechanisms, such as a change in norms and constraints (resolving Loophole 1). Empirical findings support the idea that MWR is created by ambiguity regarding the agent’s intention behind their behavior. For example, third parties punished an agent’s selfish decisions less, albeit being fully able to observe these decisions, as long as the agent had not revealed the recipient’s outcome prior to their decision (Bartling et al., 2014). In other words, even if agents could be held accountable for their *behavior*, others still punished them less because they could not be completely sure about the *intentions* behind this behavior.

Anticipated image damage. This added concept can be defined as cognitive appraisal when faced with a distribution decision, in which agents appraise how much their social image would be damaged by their selfish behavior. Anticipated image damage is estimated at the population level.

Perceived social norms. Social Norms describe behavioral rules based on social consensus regarding appropriate or prevalent behavior. Notably, we believe that any objective change in social norms can only be behaviorally relevant for agents if they perceive such change and therefore propose to measure the perceived (change in) social norms (for a discussion of the difference between objective and perceived social norms, see Tankard & Paluck, 2016). Perceived social norms are estimated at the population level.

Anticipatory guilt is a negative emotion experienced when contemplating the more unethical of available behavioral options. In situations, where the relevant decision pits own-payoff-maximization against fairness considerations, anticipatory guilt would be the negative emotion caused by contemplated selfish behavior. Anticipatory guilt is estimated at the population level.

Auxiliary assumptions

To address Loopholes 5 & 7, we extend and revise the list of auxiliary assumptions (for the full list of auxiliary assumptions, see Appendix C, Table C3).

First, resolving Loophole 5, we do not restrict the response format to binary decisions, but we limit the behavioral space to decisions ranging from sharing nothing to sharing 50% of the resources. This decision is motivated by previous research indicating a monotonous increase in appropriateness ratings for this behavioral range, but a flattening or even reversing relationship for giving more than half (Krupka & Weber, 2013). Thus, we refrain from making any predictions for this behavioral range.

4. The behavioral space should range from purely selfish distribution options (i.e., keeping the whole initial endowment) to equal distribution options (i.e., 50:50 split).

In order to specify the role of social image concerns (resolving Loophole 7), auxiliary assumption no. 4 (now no. 5, see below) had to be revised. Specifically, we assume that *social* image concerns are decisive for observing the effect of MWR. Empirical results lend preliminary support to this hypothesis (Andreoni & Bernheim, 2009). In contrast, self-image concerns seem to have little or no effect on behavior in MWR settings (Grossman, 2015) and may be more relevant for positive deviations from existing prosocial behavior than for the shift from selfish to prosocial behavior (Bénabou & Tirole, 2006; Bodner & Prelec, 2003; Grossman, 2015; Lazear et al., 2012), which are not the focus of the postulations regarding MWR. We believe all three psychological mechanisms logically require agents to have (a minimum level of) social image concerns.

5. There are two independent and sometimes conflicting motives influencing social behavior in the population: an agent's preferences over payoff distributions (henceforth *other-regarding preferences*) AND an agent's *social* image concerns.

Additional propositions accounting for heterogeneity

To account for interindividual heterogeneity of preferences and motivations (resolving Loophole 8), we posit that the effect of MWR depends on relatively stable interindividual differences (see Table 3) in *other-regarding preferences* and *image concerns*.[6]

Image concerns proposition. We already proposed that the effect of MWR should only be observable in a population where image concerns are present (aux. assumption no. 5). Here, we further specify that the MWR effect increases with agents' dispositional image

[6] Note that proposition no. 6 and 7 rest on the implicit assumption that the strength of the three psychological mechanisms should increase with image concerns. This is supported by research associating image concerns with norm-adherence (Gross & Vostroknutov, 2022), anticipated image damage (Bursztyn & Jensen, 2017) and guilt-proneness (Regner, 2021).

concerns (see proposition no. 6).[7] This is motivated by research showing a positive relationship between image concerns and prosocial behavior (Gotowiec & van Mastrigt, 2019), especially if it is visible or not incentivised (Müller & Moshagen, 2019; Winterich et al., 2013). We propose this effect to be further moderated by agents' other-regarding preferences.

Other-regarding preferences proposition. Agents also differ in the degree to which they care about fairness and prosociality (Murphy et al., 2011). We propose a non-linear effect of dispositional other-regarding preferences for all individuals who have social image concerns (see proposition no. 7): For individuals, who already have a strong inclination to act prosocially or selfishly, MWR (and image concerns) should be less important for their decision compared to those with more moderate other-regarding preferences (Grossman & van der Weele, 2017).

To summarize, we propose that these two dispositions interact, resulting in the most pronounced MWR effect among people with strong social image concerns and moderate other-regarding preferences.

Table 3

Differential effects of MWR propositions

Prop. no.	Antecedence	Consequence
6	THE HIGHER the image concerns	THE GREATER the effect of MWR (versus no MWR)
7	THE MORE extreme (i.e., selfish or prosocial) the other-regarding preferences	THE SMALLER the relevance of image concerns for the effect of MWR (versus no MWR) AND THE SMALLER the effect of MWR (versus no MWR)

Falsification of the theory. We do not consider the propositions of differential effects (i.e., moderation of the MWR effect by image concerns and other-regarding preferences) to be core to the theory of MWR. Thus, lack of proof for props. 6 and 7 would not constitute a falsification of the whole theory.

[7] This proposition is in line with findings from recent research investigating the interactive effect of other-regarding preferences and image concerns (Friedrichsen & Engelmann, 2013), indicating that misrepresentation of one's own other-regarding preferences in transparent but not intransparent situations is specifically driven by image concerns.

Definition and operationalization of the relevant concepts in these propositions

For a list of possible operationalizations of the two concepts, see Appendix C, Table C2. Note that both concepts are continuous rather than binary constructs.

Social image concerns describe how much an agent generally tends to value to be evaluated positively by relevant others. Social image concerns are determined by general dispositional social image concerns as well as domain specific deviations from the general concerns (e.g., when the relevant others are their own family or anonymous others).

Other-regarding preferences describe the true preferences of an agent over the distribution of resources between themselves and a recipient. These true preferences are independent of image concerns, and thus not influenced by MWR. We assume that agents' true other-regarding preferences are determined by their general other-regarding preferences as well as domain specific preferences (e.g., when the recipients are children, animals or the environment; or when the shared good is money, time, etc.).

Discussion

Since its introduction to the literature by Dana, Weber and Kuang (DWK) in 2007, researchers from all over the world have tested MWR with different operationalizations and showed its effect on prosocial behavior in a multitude of settings. The present level of diversity, and ambiguity in studies investigating MWR is one indication that the notion of MWR has not been sufficiently specified. This is problematic for interpretability and comparability of study results and may hinder scientific progress. In the present paper, we set out to remedy this shortcoming in the literature by providing a formal verbal theory specification, following standards developed by Glöckner and Betsch (2011). This approach allowed us to reveal and resolve loopholes of the postulations made by DWK, to advance them towards a strictly testable theory of MWR.

Key aspects of the theory advancement

The three most important loopholes of the original postulation of MWR concern (1) the definition of MWR, (2) the specification of underlying psychological mechanisms and their interrelation, and (3) the relevance of interindividual differences. Filling these gaps though specification we provide:

(1) A redefinition of MWR as situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing behavior sends to others about one's intention to be selfish. Thus, MWR is not explained by a mere reduction in accountability and its effects can be shown even in settings where an agent's behavior is observable. This redefinition of MWR also makes the propositions of underlying psychological mechanisms (props 2a-4b) more plausible.

(2) A disentanglement of the three potential psychological mechanisms underlying the behavioral MWR effect as conceptually different, yet interrelated: (a) the anticipation of less social image damage in case of selfish behavior; (b) changes in perceived social norms; and (c) a reduction in anticipatory guilt related to selfish behavior. Moreover, we conceptualize anticipated image damage as the main mechanism that receives input from social norm perceptions and provides input to anticipatory guilt. With this clear differentiation between the three psychological mechanisms and with the specification of potential pathways, we provide the basis for rigorous tests of the mechanisms of MWR in future work.

(3) Additional propositions specifying the role and interplay of interindividual differences in other-regarding preferences and social image concerns. While the effect of MWR should generally increase with the degree of dispositional image concerns, we argue that this interactive behavioral effect of image concerns and MWR is weaker for agents with more extreme (i.e., very prosocial or very selfish) other-regarding preferences. Notably, this three-way interaction sets our theorizing apart from earlier work that considered only additive effects of other-regarding preferences and image concerns (e.g., Andreoni & Bernheim, 2009). If this proposition is supported, it will have vast implications for research (e.g., the development of new auxiliary assumptions) and applications (e.g., intervention-designs tailored to the multi-trait personality of social agents).

Based on the new specifications, we provide a list of suitable manipulations of MWR that may be utilized in future studies (see Appendix A, Table A1). We invite other researchers to test and potentially falsify the different elements of our theory advancement (i.e., propositions, concept definitions and operationalizations as well as auxiliary assumptions), and thereby contribute to further theory revision and development.

Empirical roadmap

The theory advancement as specified in this paper needs to be tested rigorously. This means isolating the psychological mechanisms underlying MWR, quantifying their unique contributions and interdependencies, and testing the relevance and interaction of other-regarding preferences and image concerns for the effect of MWR. For the latter, it could be tested whether the effect of MWR linearly increases with image concerns or whether this two-way interaction is better captured by models assuming exponential increases or discontinuity (e.g., a single cutoff-point or repeated step-functions). Similarly, it needs to be tested whether other-regarding preferences indeed modulate the relevance of image concerns with the proposed inverse-u-shape quadratic function or rather in a different manner.

Open questions possibly requiring theory revision

There are some persistent open questions, which, once answered, may demand theory revision:

(1) Is self- or social image damage more decisive for the effect of MWR? This debate becomes more understandable when taking a closer look at the employed MWR operationalizations. For instance, in the ‘willful ignorance’ treatment, recipients are denied any information about whether or not the agent revealed the full outcome matrix. This protects agents’ social image, but it does not explain why a substantial fraction of agents still avoid revealing the outcome information (DWK; Grossman & van der Weele, 2017). The latter can be explained with self-image protection, which is why some researchers highlight its importance for the MWR effect (Grossman & van der Weele, 2017; Matthey & Regner, 2014). It is, however, not clear yet how exactly people can fool themselves into thinking that they would be less blameworthy for selfish behavior under MWR (see Grossman, 2010). Future research is needed to investigate the interrelations between self- and social image, and how exactly such a form of self-deception works in different MWR-settings.

(2) Can MRW be conceptualized and measured as a continuous construct? If so, how would that look like and what would it imply? For simplicity purposes, we conceptualized MWR as a binary construct: situations either contain MWR, or they do not. However, it is much more likely that MWR is a continuous construct, with situations offering more or less MWR. As we mentioned above, the different operationalizations of MWR (see Appendix A, Table A1) most likely contain different degrees of MWR, and different factors could play a role in determining the degree of MWR of a specific operationalization. For example, MWR-operationalizations may differ in the observability of the agent’s behavior, in the effort needed to exploit MWR, or in the possibility to fool oneself in addition to others (see previous discussion on social versus self-image). For a more in-depth discussion of this idea, see Appendix D.

Possibilities for extending and differentiating the theory of MWR

Once these open questions have been answered, further extension and specification of the theory might be warranted.

Possible Extension. Starting out with the proposed boundary conditions, it may be interesting to test what happens if one relaxes the auxiliary assumptions that the action space shall not include sharing above the equal split, taking options, or the possibility for moral balancing (i.e., repeated social decisions). Guiding questions here would be: are more-than-equal shares not socially demanded (Andreoni & Bernheim, 2009), or not socially desired (Duncan, 2009; Tasimi et al., 2015)? Does the inclusion of unethical options provoke feelings of entitlement (Cullis et al., 2012), because it changes the reference point for what counts as selfish (e.g., Bardsley, 2008)? Does this always reduce fair shares (Cappelen et al., 2013; List, 2007; Zhang & Ortmann, 2014)? Will repeated MWR exposure result in moral balancing (Birkelund & Cherry, 2020)? Such extensions may increase the theory's ecological validity and range of applicability.

Next, the main *mechanism of anticipated image damage* from selfishness could be extended to anticipated image benefit from prosociality, because MWR could render one's social image generally less malleable. In other words, agents may not only give less under MWR, because they fear less damage, but also because they expect less image benefit from being prosocial. This differentiation may be relevant for behavioral predictions (e.g., loss aversion, Tversky & Kahneman, 1991) and the role of interindividual differences (Sassenberg & Hansen, 2007).

Possible differentiation. Differentiation could concern the *mechanism of perceived social norms,* targeting the distinction between injunctive and descriptive norms (Jacobson et al., 2011; Reno et al., 1993; J. R. Smith et al., 2012). So far, we focused on the agent's perception of what others find socially appropriate (the *injunctive norm*). However, people frequently base their norm-following behavior on observations of others' behavior (the *descriptive norm*; (Bicchieri et al., 2022). Future research should tease apart how the two types of norm perceptions relate to the effect of MWR.

We also see potential for a more fine-grained specification of the *emotional mechanism,* such as an elucidation of the role of *shame* or *conflict*. Similar to guilt, shame is a negative self-directed moral emotion (Tangney et al., 2007), and some researchers include shame in their explanation of the effect of MWR (e.g., Bonner et al., 2017; Regner, 2021). Lastly, there are other interindividual differences which might come into play in modulating the effect of MWR. These could be differences in Guilt Proneness (Regner, 2021), HEXACO factors (e.g., Ashton et al., 2014), in social norm espousal (Bizer et al., 2014), in need for cognition (Petty et al., 2009) or in (social) loss and reward processing (e.g., Boyce et al., 2016; BIS/BAS, Fricke & Vogel, 2020; Sassenberg & Hansen, 2007). Future theorizing should take these distinctions into account and test empirically for their unique contributions.

How does MWR relate to other theories

Lastly, we would like to discuss the contribution of MWR in the context of related theories in moral psychology. Albeit relating to several other theories, the theory of MWR captures unique aspects of the moral decision-making process. It applies ideas about a combined impact of situational factors and interindividual differences known from moral judgment theory (e.g., Kohlberg, 1981) to the behavioral domain. It also adds to behavioral theory (e.g., Ajzen, 1991) as it goes beyond the interactive effect of attitudes and norm perceptions by defining situational circumstances which change norm perceptions, social image anticipations and anticipatory emotional reactions.

A theory closely related to the theory of MWR, and thus requiring especially careful assessment, is social cognitive theory of moral thought and action (Bandura, 1999). It proposes a self-regulatory system bearing resemblance to the proposed mechanisms of MWR: own moral (mis)conduct is judged against internal and external factors (MWR theory: preferences and social norms), and reacted towards, for example by means of self-sanctioning (MWR theory: feeling guilt). As in MWR theory, this process can be anticipatory to inhibit misconduct prior to its initiation. Social cognitive theory also covers the idea of moral flexibility by listing four different strategies of *moral disengagement*, including ignoring harmful consequences and obscuring one's causal role in bringing about such consequences. However, the theory of MWR differs in 5 important aspects from social cognitive theory: (1) While social cognitive theory focuses on trait-like strategies for moral disengagement, the theory of MWR focuses on the situational characteristics allowing for flexibility in moral behavior. (2) While the disengagement strategies in social cognitive theory aim at mis-construal of the action, the outcome or the action-outcome-contingency, the theory of MWR is concerned with obscuring the intentions behind actions. (3) While diffusion of responsibility is listed as a disengagement strategy in social cognitive theory, it is excluded as MWR-operationalization. (4) While social cognitive theory highlights the importance of *self*-image concerns, the theory of MWR is concerned with situations protecting agents' *social* image. (5) While the theory of MWR explicitly specifies the importance of interindividual differences in (social) image concerns and their interplay with other-regarding preferences in effects of MWR, this is not specified by social cognitive theory. Apart from these content-wise differences, the theory of MWR disentangles input and output of the *judgment* stage (i.e., norm perceptions and anticipated image damage), offers precise concept definitions and operationalizations and sets boundary conditions for testing proposed effects. Taken together, these discrepancies clarify that the theory of MWR does not just describe one variant of moral disengagement and has a right to exist on its own.

The importance of situational justifications for immoral behavior has also been highlighted by attribution theory (Kelley, 1967; Kelley & Michela, 1980). Interestingly, this theory approaches the topic from a different perspective: it states that third-party observers are less likely to infer agents' intentions from their behavior when additional plausible causes explaining the behavior can be factored in. MWR could be viewed as offering such additional plausible causes. Indeed, people are judged less harshly by others when behaving selfishly under moral wiggle room, compared to when no moral wiggle room is present (Bartling et al., 2014). Thus, this theory can be used to explain how people know when a situation offers MWR, assuming some kind of meta-cognition.

To sum it up, other theoretical accounts lend interesting insights into various aspects underlying moral judgment and decision-making, but we still need a theory of MWR to actually capture this specific behavioral effect and its underlying psychological mechanisms.

The theory of MWR can also be incorporated into broader theoretical frameworks, such as the utility framework (Fishburn, 1970). The underlying idea of this framework is that people gain utility from decision outcomes and are assumed to strive for utility maximization with their choices. Crucially, expected utilities resulting from a specific decision depend on agents' preferences and situational circumstances. Originally, utility only referred to self-serving gain. Experimental evidence from behavioral economics and psychology broadened this concept. Observing a multitude of prosocial behaviors led researchers to argue against a purely selfishly motivated *homo economicus* and develop several social preference theories (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Fischbacher, 2002; Fehr & Schmidt, 1999b). These theories tried to explain prosociality by adding social preferences to the list of factors determining overall utility. Though there are different approaches to model how and why people have prosocial preferences, they all share the idea that agents who behave prosocially gain utility from behaving prosocially. However, research on MWR indicates that this does not explain all prosocial behavior. Specifically, situations without MWR may come with certain perceived social norms, anticipated image damage and anticipatory guilt, all of which carry their own (dis-) utility. The degree to which these (dis-) utilities (and their changes in case of MWR) enter into the final overall utility of a specific action depends on the weight attached to them by an agent's image concerns. Our theory specification spells out how exactly image concerns can be conceived to impact people's utility, both by pointing to its interrelations with situational characteristics and its moderation by other-regarding preferences. It can thus be seen as a further specification of the model put forward by Levitt and List (2007).

More generally speaking, theories could be formalized as sets of equations (e.g., Borsboom et al., 2021) or verbal propositions (e.g., West et al., 2019). Though utility theory is often expressed in econometric equations, we decided to employ verbal theory specification for the theory of MWR. A verbal specification has the advantage of reaching a broader readership, thus fostering interdisciplinary collaborations. We hope that our verbal theory specification will serve as a blueprint for other theories to be verbally specified. The more theories follow formal (verbal) specification rules, the easier it will be to connect them to related theories and integrate them in the network of theories. Furthermore, by relating our theory to already existing theories, we were able to identify overlaps and unique contributions. We recommend this to become a standard procedure in each new theory specification as it will help to track connections between theories and spot potential for theory-synthesis. As a discipline, and to heighten efficiency, we should strive towards more unified theoretical frameworks functioning as catalysts of knowledge about human behavior and its underlying psychological mechanisms.

Conclusion

Like many theories, the original theory of MWR suffered from underspecification. We identified the lack of clear and concise definitions of key concepts, their operationalizations and auxiliary assumptions. The three most important loopholes of this original theory concerned (1) the definition of MWR, (2) the underlying psychological mechanisms and their interrelation, and (3) the relevance of interindividual differences. Such underspecification has adverse effects on theory testing, making it inefficient. We tackle this issue by providing the first formal verbal specification of the original postulation of MWR, identifying and resolving loopholes, and by suggesting an empirical roadmap for future research. We also set the advanced theory of MWR apart from existing related theories. With these contributions, we hope to stimulate fruitful and efficient future research on MWR. Moreover, we hope that we could demonstrate the utility of verbal theory specification, which may also motivate other researchers to employ this method.

Chapter 2:

Attributional ambiguity reduces charitable giving by relaxing social norms

This chapter is based on tho Pesch, F., & Dana, J. (under review at the Journal of Experimental Social Psychology). Attributional ambiguity reduces charitable giving by relaxing social norms.

Abstract

A growing literature demonstrates reluctant giving: Many people who give to charity voluntarily would have preferred to find an excuse not to give, but the mechanisms remain unclear. Consistent with this literature, we found that attributional ambiguity significantly reduces donations to charity. Participants in our studies (*N* = 2,147) faced a binary choice between a prosocial option (i.e., giving more to charity) and a selfish option (i.e., keeping more for themselves). We manipulated whether the donation went to the same charity in both options or to two different charities. When different charities were used, participants were less likely to choose the prosocial option than when the charities were the same, regardless of which charity was associated with the more prosocial option, revealing a hidden preference for selfishness. Using incentive compatible elicitations, we found no evidence that people developed a preference for the charity associated with the selfish option, which would be predicted by self-image concerns. Instead, we found that self-serving choices were seen as less selfish under attributional ambiguity so that the presence of attributional ambiguity reduced shared expectations that not giving was socially inappropriate. Attributional ambiguity thus lowered donations by relieving social pressure to give.

Keywords: Moral wiggle room, attributional ambiguity, prosocial behavior, charitable giving, social norms

Authors' note: For all data, analysis scripts and materials see https://osf.io/6jp9q/?view_only=3543196fdd004ffb828d8a10c8c2e01f. For the pre-registration see https://osf.io/hgycv/?view_only=d25322e34d794c108c4fbe053d0de188.

Introduction

Charitable giving in the US is a growing multi-billion dollar business: In 2020, Americans donated an all-time record of $471 billion (Giving USA Foundation, 2021)**.** Why people donate is less clear. The question of motivational drivers of prosociality has occupied philosophers and scientists for centuries. Some people give to charity for truly altruistic reasons (Batson & Shaw, 1991). Carefully constructed economic experiments show that people give to even anonymous others who cannot reciprocate, consistent with a preference for more equal outcomes (e.g., Fehr & Fischbacher, 2002). Others may give because the act gives them a positive *warm glow* (Andreoni, 1990) and conforms with their moral identity as a good person (Aquino & Reed, 2002).

A burgeoning area of research in psychology and behavioral economics suggests, however, that some giving is *reluctant*. That is, some people would prefer not to give, but experience psychological costs associated with refusing to do so (Bénabou & Tirole, 2006; Berman & Small, 2012; Dana et al., 2007; Lindsey et al., 2007). People generally want to establish and maintain a positive moral image of themselves (Aquino & Reed, 2002; Dunning, 2007; Monin & Jordan, 2009; Rachlin, 2002). But this motivation stands in conflict with self-interest when faced with prosocial requests. In these settings, selfish behavior comes with psychological costs such as self-reproach (Bandura et al., 1996, 2001; Higgins, 1997) and negative self-evaluation (Jordan et al., 2015; Rothmund & Baumert, 2014). By behaving prosocially, people can avoid these psychological costs by paying the material costs of prosociality. As such, people may seek excuses why selfish behavior does not violate their own moral standards in such settings (Andreoni et al., 2017; Bandura et al., 1996; DellaVigna et al., 2012; Exley, 2016; Lin et al., 2016). This way, people can reap the benefits of the selfish choice without paying the psychological costs of doing so.

We borrowed from a classic paradigm in social psychology (Snyder et al., 1979) to investigate whether people take advantage of attributional ambiguity as an excuse to be selfish when facing charitable requests. In our experiments, participants faced a binary choice between a prosocial option (i.e., giving more to charity) and a selfish option (i.e., keeping more for themselves). We manipulated whether the charities associated with each option were the same or different, and counterbalanced which charity was associated with the selfish option. If participants systematically give less when the charities differ, they apparently use attributional ambiguity as an excuse to give less. We also used incentivized elicitation mechanisms to see whether participants subsequently preferred the charity associated with the selfish option, and whether the perception of what is normatively appropriate changed under attributional ambiguity. We furthermore specify how attributional ambiguity changes norms.

Reluctant giving

Though we see a great deal of prosociality around us, some of it is not as genuine as we would hope. Sometimes people help others or give to good causes when they would rather not, but give into the social pressure such a request creates (Bursztyn & Jensen, 2017). As such, they also look for excuses for justifying self-interested behavior (Batson et al., 2002; Monin & Norton, 2003), reflecting a fundamental desire to be seen in a positive and moral light (e.g., Kruglanski, 1989; Kunda, 1990). In this intrapersonal conflict between a need to keep and a need to give, their "want-self" would rather keep its resources (i.e., time or money), while the "should-self" feels an obligation to give (Bazerman et al., 1998).

Studies on so-called "moral wiggle room" show that people often solve this intrapersonal conflict by exploiting situational characteristics as excuses not to give. For example, participants playing simple economic games frequently give some part of an experimental endowment to other anonymous participants or to charities, even though the recipients will never learn their identity and cannot retaliate if they give nothing (Eckel & Grossman, 1996). This behavior would seem to reflect a preference for fair outcomes. But levels of generosity significantly decline once participants have a way to be selfish without revealing selfishness. For example, many participants avoid free information about the consequences for others of choosing a selfishly preferred option (Dana et al., 2007; Feiler, 2014; Grossman, 2014; Grossman & van der Weele, 2017), thus allowing them to avoid the risk of feeling self-reproach. People also strategically use the risk that charitable donations will be wasted as an excuse not to give (Exley, 2016). In field studies, people engage in costly avoidance of charitable requests, for example avoiding the exits of a supermarket where Salvation Army bell ringers are asking for donations (Andreoni et al., 2017) or not being home if they know a charity solicitor will be coming (DellaVigna et al., 2012). Similar avoidance of prosocial requests occurs in the lab, where participants will accept a smaller monetary payment if they cannot be asked to voluntarily share that payment with another participant (Dana et al., 2006; Lazear et al., 2012; Lin et al., 2016). These studies have in common that they allow one to behave selfishly without experiencing the self-reproach associated with violating one's moral standards (Higgins, 1987).

Attributional ambiguity as an excuse

Classic studies of attributional ambiguity demonstrate experimental strategies for revealing participants' hidden, undesirable motives. When Snyder et al. (1979) asked participants to choose between two rooms to sit and watch the same movie, they were equally likely to sit next to a disabled person as a person without physical disabilities. However, when the movies in the two rooms were different, only 17% of the people sat in the room with the disabled person, even though the movies were counterbalanced across rooms. The authors thus concluded that people truly desired to avoid the disabled, but only did so if this motive was not clearly revealed.

Correspondent inference theory (Jones & Davis, 1965) holds that the strength of inference one can make from observing someone's choices depends on the number of *noncommon effects* between the chosen and the forgone options. Noncommon effects are outcomes that are brought about by selecting one specific alternative, but not another. Decisions that differ only on one dimension (e.g., whether the person in the room has a disability) allow observers to attribute the agent's intention to that dimension. Having multiple noncommon effects (e.g., the people and the movies in the rooms) creates attributional ambiguity: It is not clear which dimension drove the agent's choice. Attributional ambiguity thus allows one to choose according to their intrinsic preference with reduced concern about revealing undesirable motives to others. The effects of attributional ambiguity on behavior have been observed in studies of discrimination (Batson et al., 1986; Norton et al., 2004; Snyder et al., 1979) and willful ignorance (Woolley & Risen, 2021).

Attributional ambiguity in prosocial decision settings can thus be understood as providing moral wiggle room. When people are asked to choose between donation amounts, the amount of money is the only noncommon effect. When introducing a second noncommon effect, for example different charities associated with the different donation amounts, it is not clear whether giving less reveals a selfish motive or a preference for a particular charity.

Identifying mechanisms

Though the effect of providing moral wiggle room has been demonstrated in a variety of decision settings, the underlying mechanisms remain unclear. The literature commonly assumes that people were hiding their true preference for selfish outcomes to avoid feeling guilt (D'Adda et al., 2018; Feiler, 2014; Garcia et al., 2020; Larson & Capra, 2009; Momsen & Ohndorf, 2020; Thunström et al., 2014) or to maintain a positive self-image while still behaving selfishly (Grossman & van der Weele, 2017; Matthey & Regner, 2014). Yet, direct evidence of self-image concerns is lacking. Using a series of control choices, for example, Exley and Kessler (2021) suggest that much of the information avoidance (i.e., one form of moral wiggle room) found in prior studies on giving may not be due to selfishness-related conflicts, but perhaps other factors such as confusion or inattention.

Self-image concerns have long been recognized as an important source of prosocial behavior (Barclay, 2004; Baumeister, 1998; Bem, 1972; Festinger, 1957; Fiske, 2009; Hardy & Van Vugt, 2006; Kawamura et al., 2021; Konow, 2000; Willer, 2009). Conceptually, self-image concerns can be represented by the emotional experience of guilt, as guilt can be defined as the emotional reaction to a private transgression (Cohen et al., 2011). Empirically, guilt has been shown to be connected to self-image concerns (Erlandsson et al., 2016). However, little direct evidence exists to show that self-image concerns drive reluctant giving. Individual differences in guilt and social value orientation at least correlate with how having moral wiggle room will change behavior (Regner, 2021). Woolley and Risen (2021) find that attributional ambiguity changes behavior both in public and private decision settings, and thus conclude that the effect is driven by self-image related factors. Self-image has also been hypothesized to drive the effect of moral wiggle room more broadly by several authors (Dana et al., 2007; Lazear et al., 2012; Matthey & Regner, 2014; Momsen & Ohndorf, 2020). Returning to the example of using the risk that charity might be ineffective as an excuse not to give, a self-image concerns account would suggest that people actually believe that the reason they did not give was risk; i.e., that they used the excuses to themselves so that they could maintain a positive self-image (Bem, 1972; Goffman, 1959) without paying the monetary cost. In their studies on avoiding people with disabilities, Snyder et al. (1979) do not find quantitative support for this preference-change mechanism by way of participants' ratings of the movies, but some participants in their study claim that their decision was based on movie preference during debriefing.

Another possibility, however, is that the mere presence of an excuse relaxes the perception that generosity is socially normative in a given situation. Social norms can be seen as rules and standards of behavior within a group that proscribe selfish interests in favor of group interests by way of cooperation and prosociality (Ohtsuki & Iwasa, 2006; Thøgersen, 2008). They often function in an internalized way, meaning that people also follow social norms under complete anonymity and when not being observed by others at all (Bicchieri, 2005; Conte et al., 2010). When social norms are internalized, sanctions or rewards are administered by the individual themselves in the form of experiencing guilt or pride (Kimbrough & Vostroknutov, 2016). Social norms have been implicated as a mechanism in experiments demonstrating reluctant giving (Bartling et al., 2014; Conrads & Irlenbusch, 2013; Krupka & Weber, 2013). If excuses for not giving are readily apparent in a situation, then not giving does not clearly reveal a selfish motive and therefore, the normative pressure to give will be weakened. Returning again to the example of risk as an excuse for not giving, if risk could be a reason that people do not give, then they do not necessarily reveal a selfish motive by not giving. Since social norms exist to proscribe selfishness, the norm of giving in this situation could thus be weaker. If people care about doing what is socially appropriate, they may give less when they have the excuse of risk, even if they do not have self-image concerns. That is, even if they do not fool themselves into believing the reason they are not giving is risk, people may give less because they are less likely to believe that giving is socially normative in the situation. Within this paper, we will tease apart these two possible pathways of self-image and social norms.

Overview of Studies

In three studies, we investigated the effects of attributional ambiguity on charitable giving. We predicted that people gave less to charity when given a reason other than selfishness. We further contributed to the understanding of the effect by investigating potential mechanisms.

In the first two studies, participants faced a binary choice between giving more or less to charity. We manipulated whether the donation went to the same charity in both options (Same Charity condition) or whether the charity differed between the options (Different Charities condition). We hypothesized that participants in the Different Charities condition were less likely to choose the prosocial option. In study 2, we replicated the effect in a larger sample, and explored potential mechanisms. Did people fool themselves into thinking that they actually preferred the charity associated with the selfish option (self-image account)? Or did people simply know that selfish behavior was seen as less socially inappropriate under attributional ambiguity because others cannot judge whether they wanted to be selfish or whether they preferred a certain charity (social norms account)? In our design, we tested the self-image account by asking participants to vote for which charity should receive an additional donation of $50, with the money going to the charity that received the most votes. If participants in the Different Charities condition systematically voted for the charity that was linked to the selfish option in their choice setting, it would be strong suggestive evidence that self-image concerns played a decisive role in the effect of attributional ambiguity. For testing the social norms account, we elicited participants' perception of the prevailing social norm using the incentive compatible elicitation method of Krupka and Weber (2013). If selfish behavior was perceived as less socially inappropriate under attributional ambiguity, this would be strong suggestive evidence for the social norms account. Because social norm ratings can be influenced by people's prior behavior in the experimental part of the study, we elicited the social norms for the Same and the Different Charities conditions in an independent sample in study 3a, again using a between-subject design. Study 3b finally sheds light on the role of perceived selfishness in the social norm change.

Study 1

Study 1 examined the effect of introducing attributional ambiguity into a charity decision context where charities were chosen to be equally attractive[8]. Following Snyder et al. (1979), we expected participants to be more likely to choose the selfish option when there was attributional ambiguity (i.e., in the Different Charities condition), compared to when there is no attributional ambiguity (i.e., in the Same Charity condition). We furthermore examined self-image related elements of the mechanism, such as a change in charity preferences due to our manipulation.

Methods

Participants and design. The pooled data of Snyder et al.'s study 1 and study 2 revealed a large effect size of V = .425. Because replications generally lead to effect sizes that are considerably smaller than the original (Open Science Collaboration, 2015) and our context was slightly different, we recruited 240 subjects. Eighteen participants were released from the study without making a choice after failing comprehension questions that ensured they understood the task, leaving us with 222 of subjects who took part in the study (see Materials in https://osf.io/6jp9q/?view_only=3543196fdd004ffb828d8a10c8c2e01f for comprehension questions). This final sample size yielded a power above .8 to detect an effect half the size of the original. In all studies, including study 1, we recruited US-based participants from Amazon's Mechanical Turk with an approval rate of at least 98% and a minimum of 50 approved HITs. We excluded participants who had participated in any of our prior studies. The study took about 5 minutes. Participants received a flat fee of $0.35 in addition to a bonus payment of $0.40 to $0.50, depending on their choice. We furthermore incentivized the question of which charity was more popular amongst the participants of this study with $0.10. The study was a between-subjects design, with 111 in each condition.

Procedure. In all conditions, participants first read the same general instructions, and answered two comprehension questions. Participants who failed the comprehension questions twice were then exited from the study and did not proceed to the decision stage. All other participants were then randomly assigned to one of the two conditions, and chose

[8] For the results of our pre-study, see Appendix A.

between two options that allocated money to themselves and a charity (see Figure 1). In the Same Charity condition, the donation went to the same charity (either No Lean Season or the END Fund) in both options. In the Different Charities condition, the charities differed between the options. We counterbalanced the charities in the Different Charities condition, and randomly assigned participants to one of the charities in the Same Charity condition. Both the participants and the charities were paid according to participants' choices.

Figure 1

Options in the Different Charities condition (left) and the Same Charity condition (right)

A	You: \$0.50 No Lean Season: \$0.20
B	You: \$0.40 END Fund: \$0.40

A	You: \$0.50 No Lean Season: \$0.20
B	You: \$0.40 No Lean Season: \$0.40

After the allocation decisions, all participants were asked to rate on a scale from 1 ("not at all") to 5 ("very much") how conflicted they felt about their decision, how satisfied or happy they were with their decision, and four questions gauging how trustworthy the charities were and how important charitable giving was. All participants then read the description of both charities, and answered two incentive compatible questions about the charity they preferred. First, they indicated their personal preference for one charity by voting for which charity should receive an additional donation of \$50, with the money going to the charity that received the most votes (personal preference question). They then indicated which charity was more popular in general, receiving a bonus of \$0.10 if their answer was the most commonly given (popularity question). For exploratory reasons, we asked participants to make a couple of hypothetical sharing decisions at the end of the experiment (see Appendix B2).

Results

Neither the two counterbalanced Same Charity conditions, nor the two Different Charities conditions significantly differed from each other (*p*s > .472), indicating that participants did not generally prefer one charity to the other. We therefore collapsed the data in each condition. Participants were more likely to behave prosocially in the Same Charity condition

(64%) than in the Different Charities condition (41%), Chi^2(1)= 11.29, *p* = .001, V = .23, 95% CI [.10, .35] (see Table 1).

Table 1

Being exposed to the Different Charities condition makes it more likely for participants to select the selfish option

	Same Charity	Different Charities	
selfish	40 (36.0%)	65 (58.6%)	105
prosocial	71 (64.0%)	46 (41.4%)	117
	111	111	222

In the personal preference question, 62% of all participants voted for the END Fund to receive the additional donation.n. Also, 60% guessed that the END Fund would be more popular amongst the participants. Comparing the two counterbalanced versions of the Different Charities condition, personal preference for the END Fund was not significantly stronger when it was associated with a selfish choice than when it was not (62.5% vs. 58.2% voting for the END Fund to receive the additional donation), Chi^2(1)= 0.22, *p* = .642, V = .04, 95% CI [-.14, .23]. Similarly, there was no difference in the popularity question between the two counterbalanced Different Charities conditions (58.9% vs. 54.5% guessing that others preferred the END Fund), Chi^2(1)= 0.22, *p* = .641, V = .04, 95% CI [-.14, .23]. When looking at how trustworthy participants perceive the two charities in the self-report measures, we did not find any differences between the two counterbalanced Different Charity conditions, all *p*s > .100 (for more analyses on these self-report measures, see Appendix B1).

Discussion

Introducing attributional ambiguity to a prosocial decision context decreased choice of the prosocial option from 64% to 41%. We thus conceptually replicate Snyder et al. (1979) in the domain of charitable giving, revealing a hidden preference for the selfish option. We did not find direct evidence of self-image concerns by way of people indicating that they chose according to the noncommon effect: Participants were not more likely to report a personal preference for the charity that matched the selfish option. In other words, our participants did not claim to base their selfish choices on a preference for a specific charity. Study 2 investigated whether attributional ambiguity weakens the social norm against selfishness and whether social norms mediate the effect of attributional ambiguity on giving.

Study 2

Study 2 sought to replicate the behavioral effect of attributional ambiguity and to investigate its effect on perceived social norms. We used a method that elicits true beliefs about norms through incentivized choices (Krupka & Weber, 2013). Specifically, participants indicate the social appropriateness of certain behavioral responses, while being incentivized for picking the option that is chosen by most other participants. This way, participants are incentivized to state their true belief about the social appropriateness of the behavior in question.

Methods

Because we were adding social norms as a factor to investigate and wish to distinguish it from other image-based explanations, we preregistered a highly-powered study seeking 750 participants per condition including study design, hypotheses and analysis plan. We collected data from 1492 participants from Amazon Mechanical Turk (54% female), because 8 participants failed the comprehension questions, meaning that they did not reach the decision stage of the experiment. A sensitivity analysis revealed an 80% power to detect an effect size of w = .073. We replicated the set-up used in study 1, adding the social norm elicitation method of Krupka and Weber (2013) to the post-experimental questionnaire. On a scale from 1 (“Very socially inappropriate”) to 4 (“Very socially appropriate”), participants were asked to indicate the social appropriateness of each option in their respective experimental conditions. Participants were informed that one of the two options was selected randomly to determine a bonus payment of $0.10 if the participant’s response was the same as the most common response. We also asked participants to indicate how appropriate they personally found each of the two behavioral options, ranging from 1 (“Very inappropriate”) to 4 (“Very appropriate”).

Results

As in study 1, we first tested whether participants significantly favored one charity over another, and found no significant differences across counterbalanced conditions, *p*s > .322, The main effect of our Different Charity manipulation was smaller than in study 1, but statistically significant, Chi^2(1) = 25.13, *p* < .001, V = .13, 95% CI [.08, .18]: In the Same Charity condition, about 67.8% of participants behaved prosocially, dropping to 55.2% in the Different Charities condition (see Table 2).

Table 2

Being exposed to the Different Charities condition makes it more likely for participants to select the selfish option

	Same Charity	Different Charities	
selfish	240 (32.2%)	335 (44.8%)	575
prosocial	505 (67.8%)	412 (55.2%)	917
	745	747	1492

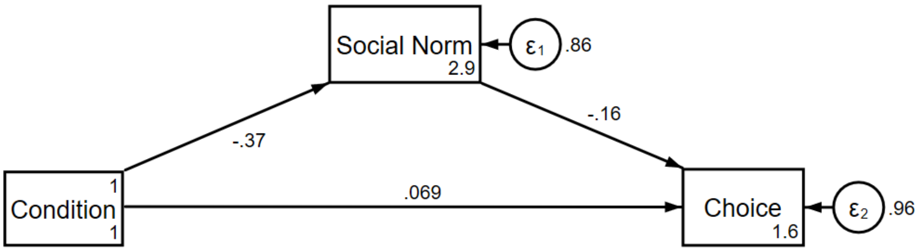
As in study 1, we did not observe a subsequent effect of our manipulation on participants' personal preferences over the two charities in the two counterbalanced versions of the Different Charities condition, Chi^2(1) = 0.07, *p* = .799, V = -.01, 95% CI [-.08, .06]. However, these participants thought that the charity linked to the selfish option would be more popular, Chi^2(1) = 7.77, *p* = .005, V = .10, 95% CI [.03, .17]. A similar pattern can be seen in the two counterbalanced versions of the Same Charity condition: Participants thought that the charity they just donated to would be more popular amongst others, Chi^2(1) = 5.36, *p* = .021, V = .09, 95% CI [.01, .16], but this was not significantly related to their personal preference, Chi^2(1) = 2.48, *p* = .115, V = .06, 95% CI [-.01, .13].

Our manipulation also had a significant effect on the perceptions of social norms: People perceived selfish behavior to be more socially permissible in the Different Charities condition (*M* = 2.97, *SD* = 0.98) than in the Same Charity condition (*M* = 2.21, *SD* = 0.92), *t*(1490) = 15.51 *p* < .001, *d* = .80, 95% CI [.70, .91]. The social permissibility of behaving prosocially decreased slightly, but significantly from *M* = 3.80 (*SD* = 0.54) in the Same Charity condition to *M* = 3.66 (*SD* = 0.65) in the Different Charities condition, *t*(1490) = -4.43, *p* < .001, *d* = -.23, 95% CI [-.33, -.13]. We conducted a mediation analysis to see if perceived social norms regarding selfishness mediated the effect of attributional ambiguity on giving (see Figure 2). The indirect effect of attributional ambiguity through perceived social norms was significant, β = .059, Sobel *Z* = 5.56, *p* < .001, mediating about 47% of the effect.[9]

[9] For a discussion on the role of personal norms in our setup, see Appendix B.

Figure 2

Path diagram (with standardized coefficients) displaying the mediation of attributional ambiguity (Condition: 0 = Different Charities, 1 = Same Charity) on prosocial behavior (0 = selfish choice, 1 = prosocial choice) through social norms concerning selfish behavior



Discussion

In study 2, we replicated the behavioral effect of attributional ambiguity on charitable giving. We also shed light on the potential mechanism behind it: While participants did not seem to change their personal preference for one of the charities, they did think that others would do so, as the popularity of the charities is impacted by our experimental manipulation. Importantly, they also perceived a change in the social norm: Selfish behavior was expected to be seen by others as more socially permissible in the Different Charities condition. Our mediation results suggest that attributional ambiguity increased selfish behavior by making it more socially permissible, however, causality cannot be inferred from our experimental setup. Specifically, the reported social norm ratings could reflect a shared belief that most participants want to rationalize their selfish choices after the fact. Study 3a thus examined whether attributional ambiguity changes the perceived social norm in a sample that does not make the charity decision. Study 3b will investigate the relation of this change in norm to the perception of behavior as selfish.

Study 3a & b

Methods

We recruited 485 participants from Amazon Mechanical Turk for each study 3a and b, so as to achieve a power of .8 to detect a small effect of d = .23. In study 3a, the final sample size was N = 433, while in study B a total of 460 participants passed the comprehension questions. The first part of the instructions was identical to studies 1 and 2, but instead of explaining that they will make the decision in the decision stage, instructions talk about person X making a decision. After passing the comprehension questions, participants of study 3a were asked to give ratings about the social permissibility of behavior as described in study 2 (Krupka & Weber, 2013). In study 3b, we asked participants about how selfish each behavioral option would be on a scale from 1 ("Not selfish at all") to 5 ("Very selfish"). Participants then answered three questions about how much they would feel guilty/regret/have a bad conscience if choosing option A on a scale from 1 ("Not at all") to 5 ("Very much"). We manipulated between-subject whether participants saw the decision scenario of the Same or the Different Charities condition. In study 3a, participants received an extra $0.50 if their response was the most frequent response given by other participants. Participants also indicated how permissible they personally perceived the two different behavioral options as described in study 2. In study 3b, all participants received an extra $0.20 if they passed the comprehension questions.

Results

In study 3a, participants in the Different Charities condition rated selfish behavior (i.e., choosing option A) to be more socially permissible (*M* = 2.76, *SD* = 0.89) than participants in the Same Charity condition (*M* = 2.42, *SD* = 0.88), *t*(431) = 4.04, *p* < .001, *d* = .388, 95% CI [.20, .58]. At the same time, prosocial behavior (i.e., choosing option B) was perceived as slightly but significantly less socially permissible in the Different Charity condition (*M* = 3.63, *SD* = 0.66**)** then the Same Charity condition (*M* = 3.78, *SD* = 0.51), *t*(431) = -2.66; *p* = .004, *d* = .26, 95% CI [-.44, -.07]. Personal norms follow the same pattern (see Appendix B3).

In study 3b, participants in the Different Charities condition rated choosing option A as less selfish (*M* = 3.20, *SD* = 1.20) than participants in the Same Charity condition (*M* = 3.88, *SD* = 1.53), *t*(458) = -5.32, *p* < .001, *d* = -.50, 95% CI [-.68, -.31]. When asking participants how they would anticipate feeling if they had chosen option A, participants in the Different Charities conditiion anticipate feeling less guilty (*M* = 2.83, *SD* = 1.44) as compared to participants in the Same Charity condition (*M* = 3.06, *SD* = 1.40), *t*(458) = -1.68, *p* = .047, *d* = -.16, 95% CI [-.34, .03]. Similarly, participants report higher levels of anticipated bad conscience, *t*(458) = -2.36, *p* = .009, *d* = -.22, 95% CI [-.40, -.04], and a trending effect in this direction on antcipated regret, *t*(458) = -1.54, *p* = .06, *d* = -.14, 95% CI [-.33, .04].

Discussion

Participants who did not make charitable decisions, and thus had no motivation to justify selfishness, still perceived that choosing option A was not as normatively inappropriate when attributional ambiguity was introduced. The effect size, however, was smaller for these participants than in study 2, possibly indicating additional post-decision rationalization in study 2. Furthermore, choosing option A is seen as less selfish, indicating that attributional ambiguity changes the way we evaluate behavior in terms of its morality. In the Different Charities condition, participants also anticipated feeling less guilty and having less of a bad conscience if they would choose option A compared to participants in the Same Charity condition.

General discussion

Using a classic paradigm that builds on correspondent inference theory, our studies suggest that some people who give to charity have a hidden preference for selfish outcomes. When selfish behavior could not be unambiguously attributable to selfish motives, people more often chose the selfish option. Using incentivized decisions and self-reports, we found no support for the idea that these people fooled themselves by adjusting their charity preference to fit their (more selfish) choices. We did, however, find support for a social norms account: Introducing attributional ambiguity reduced the perception that generous behavior was socially expected, thus relieving normative pressure to give. We identified a reduction in the perceived selfishness of own-payoff-maximizing behavior under attributional ambiguity as a potential driver behind this normative change.

Our findings conceptually replicate the findings of Snyder et al. (1979) in the domain of charitable giving. While Snyder et al. show a hidden preference for avoiding physical proximity to disabled people, we show a hidden preference for selfish behavior. Our results are in support of correspondent inference theory (Jones & Davis, 1965): By introducing a second noncommon effect to the decision setting (i.e., different charities associated with the selfish and generous options), a potential observer cannot draw clear dispositional inferences about the decision-maker from observing one single decision. Only a large number of observations across conditions allows us to identify selfishness as a hidden motive. We thus identified attributional ambiguity as one more form of moral wiggle room (Dana et al., 2007). Attributional ambiguity reduces the transparency between people’s intentions and the outcome of their behavior. It is hard to tell from observing the decision of one person what motivated the person to decide this way, i.e., whether the person chose due to the specific charity or due to the specific distribution of the decision.

We found support for the hypothesis that a change in social norms is related to the effect of attributional ambiguity on prosocial behavior. Not only did we find a mediation of social norms, we also observed this social norm change in a separate sample of participants who did not make a prosocial decision themselves. Selfish behavior was seen as less socially inappropriate under attributional ambiguity. Decisions are apparently judged less harshly when observers cannot draw clear dispositional inferences about the decision-maker from the observed behavior. Attributional ambiguity weakens the signal of behavior, which results in less pressure coming from social norms. Note, however, that our mediation analysis is only one possible way of understanding our data. We back up this conception by showing that our experimental manipulation not only leads to a change in prosociality, but also changes the way people conceive the social norm in a given situation. We furthermore show that attributional ambiguity leads to changes in the perceived selfishness of behavior: Choosing the selfish option is perceived as less selfish under attributional ambiguity. As such, our results are also in line with correspondent inference theory (Jones & Davis, 1965), postulating that the strength of inference one can make from observing someone's behavior depends on the number of noncommon effects between the different options. As a second noncommon effect such as different charities creates attributional ambiguity as to which dimension drove the agent's choice, outside observers cannot be sure whether opting for the own-payoff-maximizing option is driven by a desire to maximize one's own gains, or whether a charity preference was the decisive factor. Thus, choosing option A was seen as less selfish in the Different Charities condition, as the agent could have acted according to a charity preference, and not in order to maximize their own payoffs.

Following up on the discussion in the moral wiggle room literature regarding potential mechanisms, our data supports the idea that the observed behavioral effect is actually driven by selfishness-related elements. Because selfish behavior is seen as less selfish under attributional ambiguity, the social norm relaxes, and people experience less pressure to give to charity. Though our participants did not try to fool themselves, they still seemed to consume what other people thought of them, and anticipated feeling less guilty if they chose the selfish option. It seems sufficient that outside observers cannot judge their prosocial type clearly, even though they themselves seem to understand that their motivation for choosing the selfish option is selfishness. Our results can also inform us about the mechanism behind behavioral changes observed in other forms of moral wiggle room. For example, people have been shown to use risk as an excuse for selfish behavior (Exley, 2016). The level of risk can be seen as a second noncommon effect within this decision setting, and thus also diffuses the signal that a selfish choice sends to others about the agent's underlying motives.

Our findings on the importance of social norms for the effect of attributional ambiguity on sharing decisions suggest that concerns about social image may play a role in the moral wiggle room effect (see Andreoni & Bernheim, 2009; Grossman, 2015). However, it should be noted that in our experiments, sharing decisions were not observed by others. Despite this, participants still changed their behavior according to changing norms, which could be considered evidence of internalized norm-following (Bicchieri, 2005; Conte et al., 2010). Future research should explore whether the effect of attributional ambiguity on behavior is even more pronounced when sharing decisions are made in the presence of others.

So, why do people give to charity? Often, they give reluctantly in the presence of a request. Indeed, it has been suggested that as much as 50% of the time someone chooses to give in a lab or field experiment, they do so reluctantly and would have preferred to avoid the request or have an excuse not to give (D. M. Cain et al., 2014). Our results suggest that social expectations, rather than self-image concerns, are a key driver. But norms about what is appropriate in this arena are fragile. The introduction of attributional ambiguity reduces shared notions of what is socially appropriate, and people give less as a consequence. For practitioners on the ground, our results suggest caution about injecting ambiguity into the charity choice, such as might occur when providing more variety in options, at least when targeting a demographic that may be reluctant.

Conclusion

Attributional ambiguity allows people to behave more selfishly by making selfish behavior less socially inappropriate. People do not try to fool themselves or others into thinking that they actually prefer the charity that is attached to the selfish choice. However, people selfishly benefit from the ambiguity of what motivated their choice because of the reduction in social expectations to give. Future research should investigate the causality in the link between a change in social norms and subsequent behavior.

Open practices

We report how we determined our sample size, all data exclusions, and all manipulations in the study. All studies were granted exemption by the University’s Human Subjects Committee (protocol number: 2000020511). All data, analysis code, measures, and research materials for all three studies, and the two studies reported in the Appendix are available at https://osf.io/6jp9q/?view_only=3543196fdd004ffb828d8a10c8c2e01f. Data were analyzed using STATA, version BE 17.0. Design and analysis were not pre-registered for study 1 and 3, but for study 2 (see https://osf.io/hgycv?view_only=d25322e34d794c108c4fbe053d0de188).

Chapter 3:

Conflict in willful ignorance: A mouse-tracking investigation

This chapter is based on tho Pesch, F. (prepared for submission). Conflict in willful ignorance: A mouse-tracking investigation.

Abstract

People ignore information that would otherwise be instrumental to their decisions. In prosocial decision-making, willful ignorance has been suggested to stem from a desire to avoid cognitive conflict when acting selfishly. A possible method to implicitly measure cognitive conflict is by tracking mouse trajectories of individuals as they make binary decisions on a computer screen. In a fully incentivized experiment (*N* = 210), participants made several binary decisions regarding the distribution of money between themselves and a charity. In the first decision, they had the option to ignore the impact of their decision on the charity. Subsequently, participants made 18 choices, with their mouse movements tracked, of which 12 were unaligned trials (wherein the option with a higher payoff for the participant resulted in lower donation for the charity). Analyzing the mouse trajectories, the study shows that participants experienced more cognitive conflict in unaligned trials than in aligned trials (proof of concept). The study also demonstrates that individuals who experienced more cognitive conflict in unaligned trials were more likely to engage in willful ignorance. Also, participants experienced more cognitive conflict when choosing selfishly than when choosing prosocially. Additionally, the interaction between allocation choice and interindividual differences in Guilt Proneness, Social Value Orientation, and Honesty-Humility in predicting cognitive conflict suggested that participants who were dispositionally selfish were equally conflicted, regardless of their choice, while those who were dispositionally prosocial felt less conflicted when choosing the prosocial option. As dispositionally selfish participants were more likely to ignore, willful ignorance could be viewed as a strategy to avoid the conflict that arises from either choice. The results suggest that willful ignorance may serve to simplify the decision context, rather than avoid conflict in case of selfish behavior.

Keywords: Willful ignorance, conflict, cognitive dissonance, prosociality

Authors’ note: For all data, analysis scripts and materials, see https://osf.io/w6s3d/?view_only=2eec5741e24c413b81853931f9473ff4. For the pre-registration, see https://osf.io/f8x6g/?view_only=a3242fdb10a043759cbc1ac91f874692.

Introduction

Sometimes, we find ourselves torn between what we want to do and what we feel obligated to do (Bazerman et al., 1998). One area where we can encounter this intraindividual conflict is in prosocial decision-making. For example, we may feel obliged to assist a friend with moving, but we may also want to have a lazy Saturday instead; We feel we should donate to charity, but we may prefer to go shopping; Or we may want to keep the entire cake to ourselves, even though we should share it with others. To avoid this type of should-want conflict, people may engage in willful ignorance, where they ignore the potential negative consequences of their behavior on others. Research on willful ignorance suggests that people indeed avoid certain information that would otherwise be instrumental for their decisions. In this study, we address the question of whether willful ignorance is indeed related to cognitive conflict in prosocial decision-making by using mouse-tracking as an implicit measure for conflict.

Prosociality and willful ignorance

Prosociality can be defined as behavior that is “costly to the actor and beneficial to the recipient” (S. A. West et al., 2011, p. 232). On a psychological level, most people feel morally obliged to behave prosocially, as they want to maintain a positive moral image (Aquino & Reed, 2002; Dunning, 2007; Monin & Jordan, 2009; Rachlin, 2002). As such, prosociality represents a tradeoff between self-interest and the desire to be seen as moral, whether by others or by oneself. It can also be viewed as an internal conflict between the want-self and the should-self. The want-self aims to maximize personal gain, while the should-self aims to maintain a positive moral image (Sezer et al., 2015). Cognitive dissonance theory (Festinger, 1957) postulates that a person holding two psychologically conflicting cognitions experience the aversive emotional state of cognitive dissonance (i.e., negative drive state, Festinger, 1957). Within this framework, the intrapersonal conflict can be seen as holding beliefs about which behavior one considers appropriate that simultaneously conflicts with a desire to maximize one’s own resources (Aronson, 1992; Beauvois & Joule, 1996). Applied to simple sharing decisions as one form of prosociality, this means that people experience cognitive dissonance when they share less than their regard as fair (Konow, 2000). As this divergence between attitude and behavior would result in the negative emotional state of cognitive dissonance, people behave prosocially to pacify their behavior with their moral standards.

In some situations, it is possible to avoid this should-want conflict altogether by ignoring unwanted information. By not knowing the negative consequences of one's own behavior for others, one can engage in selfish behavior without having to acknowledge it (Vu et al., 2023). Though rational choice theory assumes that people should value information to the extent that it helps them make more informed decisions (Stiegler, 1961), empirical evidence on willful ignorance in prosocial decision-making challenges this notion by providing several counterexamples (Dana et al., 2007; Ehrich & Irwin, 2005; Grossman, 2014; Grossman & van der Weele, 2017; Thunström et al., 2014). The literature largely relies on the hidden information treatment of Dana and colleagues (2007). The authors employed a binary dictator game, in which participants distributed money between themselves and another entity. In the baseline condition, participants decided between one selfish and one prosocial option. In the hidden information treatment, the payoffs to the other entity were hidden initially, and participants could decide to reveal the payoffs at no costs by clicking a button. In case participants revealed, they would either face unaligned payoffs, which were the same as in the baseline, or aligned payoffs, meaning that one option was better both for the participant as well as for the other entity. This way, participants who ignored did not know whether choosing the own-payoff-maximizing option would actually harm the other entity. Results indicated that some participants exploited ignorance as an excuse to choose the selfish option (Dana et al., 2007). There are different proposed mechanisms as to why people engage in willful ignorance. While some authors argue that image concerns drive this effect (Adena & Huck, 2020; Andreoni & Bernheim, 2009; Grossman, 2015; Grossman & van der Weele, 2017), others relate to concepts such as guilt (Feiler, 2014; Garcia et al., 2020; Thunström et al., 2014) or conflict (Grossman, 2014; Lin & Reich, 2018; Matthey & Regner, 2011; Woolley & Risen, 2018).

Our study centers on understanding the role of cognitive conflict for willful ignorance. The theoretical model of Konow (2000) assumes that people reduce cognitive dissonance in prosocial decision-making either by aligning their moral standards with their actual behavior by behaving prosocially, or they engage in self-deception strategies. One such strategy potentially is willful ignorance. Thus, applying cognitive dissonance theory to the domain of willful ignorance, we tested whether participants experiencing higher levels of said should-want conflict were more likely to remain ignorant about the consequences of their choice for the other entity. We built on Matthey and Regner (2011), who showed that ignorance was related to cognitive dissonance in simple sharing decision: Participants who experienced a “negative drive state” (Festinger, 1957), operationalized by longer decision times, as well as higher self-reported choice difficulty in transparent binary dictator decision, were more likely to engage in willful ignorance in settings allowing for ignorance. Though this study can be seen as first evidence for a connection between willful ignorance and cognitive dissonance, the authors only infer the experience of dissonance from reaction times and self-reports. With our study, we added to this first evidence by directly and implicitly measuring cognitive conflict in prosocial decision settings and relating these measures to willful ignorance.

Mouse-tracking as a measure for cognitive conflict

Cognitive conflict can be measured implicitly and unobtrusively by continuously tracking people’s mouse movements in computerized experiments while participants choose between two options, which are spatially separated on a screen (Freeman & Ambady, 2010). The underlying idea of this measure is that cognitive processes are continuously translated into motor responses. In computerized experiments, this would be mouse movements. The attractiveness of an option is thought to be translated into movement by creating a “pull” towards this option. Thus, the more seriously the non-chosen option is considered during the decision process, the more the mouse trajectory will deviate more from an idealized direct trajectory. This difference between the idealized and the recorded trajectory can be used to draw inferences about the experienced conflict in this situation. Cognitive conflict as measured by mouse trajectories has already been related to prosocial behavior (Kieslich & Hilbig, 2014). Utilizing different economic games to measure cooperative behavior as one form of prosociality, their results indicate that participants experienced less cognitive conflict when cooperating compared to defecting. We add to this research by investigating in which way the experience of cognitive conflict in prosocial decision-making is related to willful ignorance.

Dispositional measures capturing interindividual heterogeneity

As emphasized by Dana et al. (2007), not everyone engages in willful ignorance. There is a consistent part of the population who behaves prosocially, regardless of whether or not they have the opportunity to ignore. Similarly, some people behave selfishly, independent of the situational circumstances. Thus, it is only a certain subset of people who exploit willful ignorance as an excuse for selfish behavior (Vu et al., 2023). By combining interindividual difference measures with behavioral measures and the implicit measure of cognitive conflict, we shed light on potential underlying factors for observed behavior. Who are the people who experience more conflict, and are more likely to engage in willful ignorance? With this, we aim for a better understanding of the drivers of willful ignorance, as well as implications on an individual level.

The general idea behind willful ignorance is that some people engage in ignorance in order to choose the selfish option without looking selfish, either to others or to themselves (Dana et al., 2007; Grossman & van der Weele, 2017; Vu et al., 2023). However, as Dana et al. (2007) already pointed out, there is a substantial share of participants who behave prosocially, independent of whether they have the option to ignore, while others consistently choose the selfish option. While for people who choose consistently selfish the hidden information can be regarded as irrelevant, people who want to choose the prosocial option would have to reveal the information. As a result, dispositional tendencies to behave prosocially or selfishly should be connected to willful ignorance. The Social Value Orientation (SVO) of a person captures interindividual differences in preferences for joint outcomes and cooperation (prosocial values) as compared to a pro-self orientation (Murphy et al., 2011; P. Van Lange, 1999). As such, it is highly predictive for dictator game giving (see Thielmann et al., 2020).

Another factor that can be conceptually related to willful ignorance is guilt. Guilt has been proposed as one reason for why people ignore in prosocial decision-making by several authors (Feiler, 2014; Thunström et al., 2014). The idea is that willful ignorance allows an agent to behave selfishly without feeling guilty about it. The Guilt Proneness subscale of the GASP (Cohen et al., 2011) captures people’s propensity to feel guilty for their actions. People high in Guilt Proneness should thus be more likely to engage in willful ignorance as to avoid the negative feeling of guilt when choosing selfishly.

Another prominent mechanism of willful ignorance in the literature is image concerns (Adena & Huck, 2020; Andreoni & Bernheim, 2009; Grossman, 2015; Grossman & van der Weele, 2017). The Brief Fear of Negative Evaluation scale (Reichenberger et al., 2015) captures a person's tolerance for the possibility of being judged negatively by others (i.e., a person’s social image concerns). As such, we expect people high in Fear of Negative Evaluation to be more likely to engage in willful ignorance.

Furthermore, Honesty-Humility has shown to predict between-participant variation in the relationship between prosociality and cognitive conflict: The difference in conflict between prosocial and selfish decisions was found to be more pronounced for participants high in Honesty-Humility, compared to participants low in Honesty-Humility (Kieslich & Hilbig, 2014). In this study, we aimed to conceptually replicate these findings. Honesty-Humility is one of the six personality factors of the HEXACO model and describes the dispositional tendency of a person to be “fair and genuine in dealing with others” (Ashton & Lee, 2007, p. 156). Honesty-Humility has been consistently found to predict prosociality in different contexts (see Thielmann et al., 2020).

Present study

In this study, we utilized mouse-tracking to implicitly and unobtrusively measure cognitive conflict while participants made simple donation decisions and related these measures with ignorance choices in a different decision context in order to understand the role of cognitive dissonance and willful ignorance. The study reported below first tested the hypothesis derived from the literature on willful ignorance combined with insights from the cognitive dissonance theory (Festinger, 1957; Konow, 2000) that participants who experience more conflict in these binary dictator decisions are more likely to engage in willful ignorance (H1). We predicted that selfish behavior elicited more cognitive conflict compared to prosocial choices, conceptually replicating the main finding of Kieslich et al. (2014; H2). We furthermore explored how the dispositional measures of SVO, Guilt Proneness, Fear of Negative Evaluation and Honesty-Humility related to willful ignorance, as well as whether these measures further specify the relationship between allocation decision and cognitive conflict.

Methods

Design and procedure

We tested these hypotheses in a fully incentivized study consisting of two parts.[10] In the first part, participants completed the online survey 48 to 24 hours prior to coming to the lab. In the online survey, we administered Honesty-Humility subscale (Ashton & Lee, 2009), SVO (Murphy et al., 2011), the Brief Fear of Evaluation Scale (Reichenberger et al., 2015), and Guilt Proneness subscale (Cohen et al., 2011). We also asked participants for their handedness, age and gender. The SVO measure was incentivized with €0.01 per point.

In the second part, participants were invited to the lab. They first read the instructions and answered several control questions. We did not exclude participants who gave incorrect answers, but ran robustness checks excluding those participants. The first incentivized decision participants made was the decision with the option to ignore the donations attached to each option. Participants then made 18 binary decisions, of which 12 trials contained unaligned payoffs and 6 trials aligned payoffs. Right before these binary donation decisions, participants completed two practice trials. In each trial, the distribution options were

[10] For all analyses as pre-registered, see Appendix A.

displayed in random order in the left and right upper corner on the screen respectively. It was also randomized which option would be displayed in which corner.

Materials

Ignorance choices

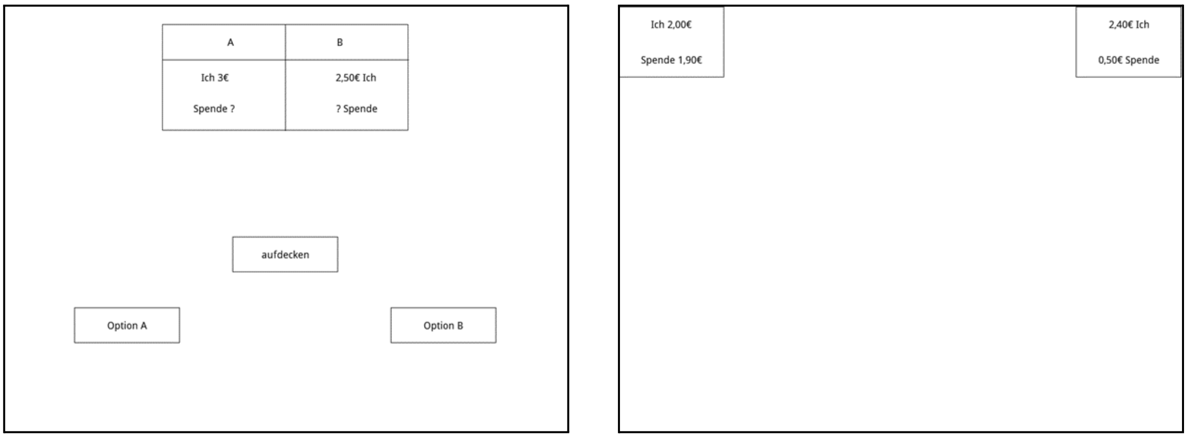
For the first decision context including the option to ignore, we replicated the setup of the hidden information treatment of Dana et al. (2007). Participants faced a decision between a potentially selfish option (A) and a potentially prosocial option (B). Participants were informed that it would be randomly decided whether they would face unaligned or aligned payoffs. In either case, choosing option A would result in a higher payoff for themselves (€3.00) than choosing option B (€2.50). In the unaligned scenario, the charity would receive a lower donation in option A (€0.50) compared to option B (€2.50). In the aligned scenario, the donations would be flipped, meaning that option A would be the more profitable option both for the participant and for the charity. Participants initially did not know which payoff distribution they faced, as the payoffs to the charity were hidden. However, participants could reveal the payoffs by clicking on a “reveal” button (see Figure 1a). As such, ignorance choices were measured as a binary variable (0 = reveal; 1 = ignore).

Cognitive conflict in simple donation decisions

The following 18 decision contexts in which participants mouse movements were recorded were designed in order to mirror the general features on this first distribution decision, though without hiding any payoffs (see Figure 1b). We designed the distribution options to represent a similar tradeoff compared to the initial decision. We “flipped” the payoffs to the charity in 6 trials (i.e., aligned trials), so that these trials had a mutually beneficial option. Though we were mainly interested in cognitive conflict in the 12 unaligned trials, we added 6 aligned trials to make the task less monotonous for our participants, as well as in order to test the assumption that self-other tradeoffs elicit more cognitive conflict than task without such a conflict. Participants could choose between options containing donations of €0.01 to €2.40, and additional payoffs to the participants themselves between €0.80 to €3.00.

Figure 1

Screenshots of a) the ignorance decision context (left), and b) one of the 18 mouse tracking decision contexts (right)



A	B
Ich 3€ Spende ?	2,50€ Ich ? Spende

Mouse-tracking specifications

Mouse-tracking was implemented using the Mousetrap plugin (Wulff et al., 2021) for OpenSesame (Mathôt et al., 2012). Participants were not told that their mouse movements were recorded, nor did they receive any specific instructions about moving the mouse. Participants began each mouse-tracking trial by clicking a start button. The mouse cursor position was then reset to the bottom center of the screen. Participants could indicate their response by clicking on one of the two distribution options in the top right and left corner of the screen. There were no time limits. The experiment was conducted full screen with a resolution of 1,920 × 1,080 pixels. Lab computers were running Windows 10, and mouse settings were left at their default values (mouse pointer speed of 10; medium). Cursor coordinates were recorded every 10 ms.

We calculated mouse trajectories by first mapping all trajectories on one side, time-normalizing trajectories into 101 time bins, and aggregating them first within and then across participants for the different decisions.

To capture the curvature of mouse trajectories, the literature has put forth various methods of how to calculate variables for statistical analyses (Freeman and Ambady in 2010; Koop and Johnson in 2011). One commonly used approach is the maximum absolute deviation (MAD), which calculates the highest perpendicular deviation between the real trajectory and the straight line that connects the trajectory's starting and ending points. We will use this measure to approximate the degree of conflict experienced by participants in each trial.

Participants

We collected data from 210 university students with an average age of 26.4 years (SD = 9.92), 66.19% of them identified as female. Participants earned €5.53 on average, consisting of an average of €0.84 for the incentivized SVO choice, €2.79 on average for the incentivized choice involving the option to ignore, and an average of €1.89 for the randomly chosen incentivized allocation decision from the mouse tracking part. Participants also donated an average of €3.59.

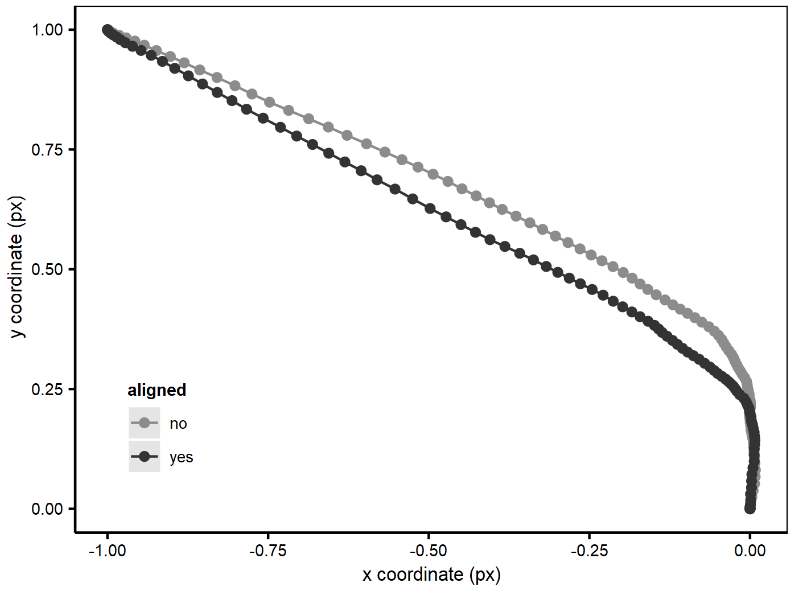
Results

Within the ignorance decision setting, we found that overall, 22.38% of our participants ignored the consequences of their own behavior. Of those participants who faced unaligned payoffs (independent of whether or not they revealed), 76.99% chose the prosocial option in the ignorance context. Within the 12 unaligned trials of the mouse-tracking part of the study, participants chose the prosocial option in 73.78% of the trials. In the 6 aligned trials, participants chose the mutually beneficial option in 96.24% of the trials.

To assess whether participants actually experience more conflict when facing a self-other tradeoff setting, we compared the trajectories of aligned to unaligned trials. As depicted in Figure 2, trajectories displayed a greater curvature in the unaligned than in the aligned trials, suggesting that participants experienced greater cognitive conflict in unaligned trials. To test this hypothesis statistically, we calculated two MAD scores per participant for unaligned and aligned trials respectively. Using a paired *t*-test, we saw that participants experienced significantly more conflict in unaligned trials compared to aligned trials, $t(209) = 4.951$, $p < .001$.

Figure 2

Average mouse trajectories in unaligend and aligned trials

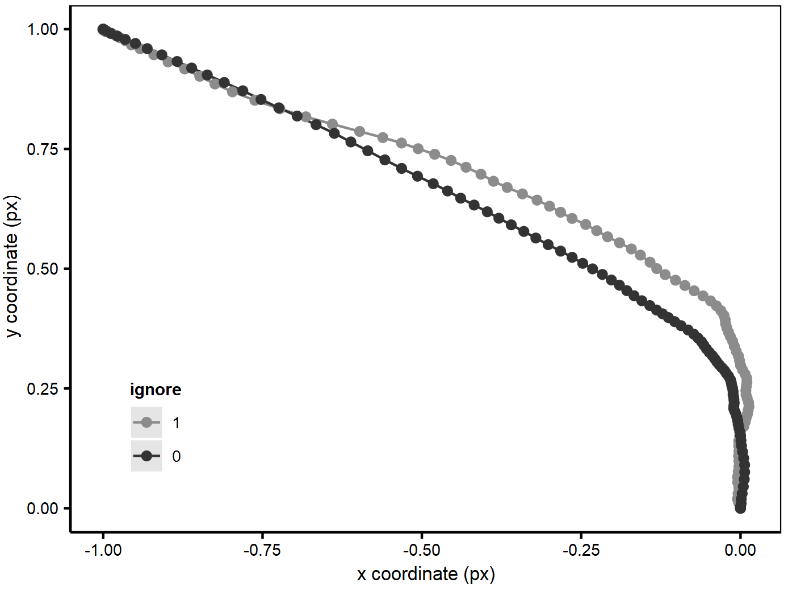


Ignorance and conflict

To test H1, we compared the MAD scores for unaligned trials between those participants who decided to ignore in the ignorance decision context, and those participants who revealed the information in that context. In line with our hypothesis, ignorant participants experienced significantly more conflict in unaligned trials than participants who had revealed the information, *r* = .133, *t*(208) = -1.936, *p* = .027 (see Figure 3). However, we also saw a similar pattern for MAD scores in aligned trials, *r* = .169, *t*(208) = 2.466, *p* = .007.

Figure 3

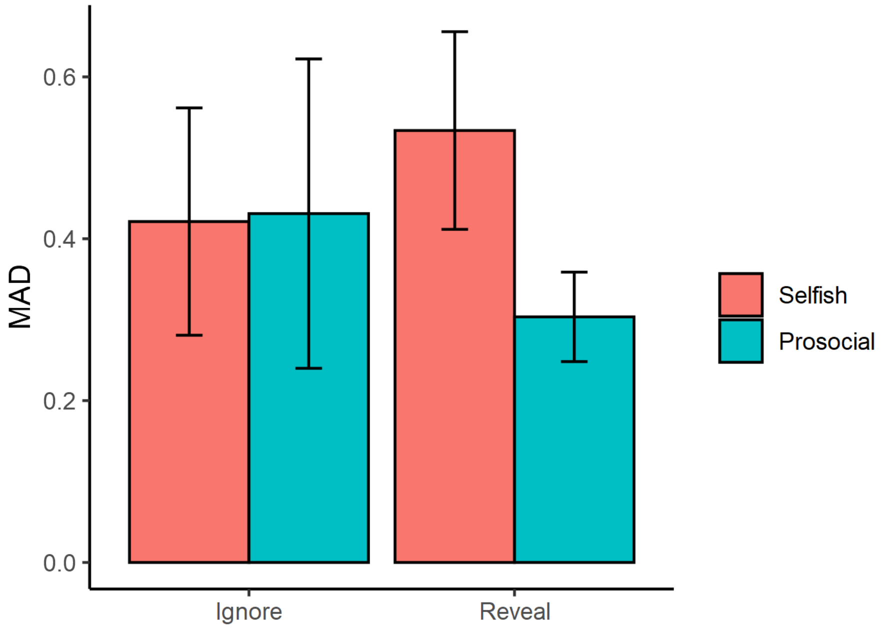
Mouse trajectories in unaligned trials between participants who ignore and participants who do not ignore



As participants who ignored chose the prosocial option in 46.63% of the unaligned trials, while participants who revealed chose the prosocial option in 82.16% of the unaligned trials, we also analyzed in which way this relationship between conflict and ignorance choices was further specified by participants' allocation choice. Indeed, there was an interaction of allocation decision and ignorance choice when predicting MAD scores, *t*(259) = 2.021, *p* = .044. Within the unaligned trials in which participants chose selfishly, MAD scores were not significantly correlated with participants' ignorance choices, *r* = -.124, *t*(86) = -1.159, *p* = .125. Within unaligned trials in which participants chose prosocially, MAD scores were significantly correlated with participants' ignorance choices, *r* = .127, *t*(173) = 1.685, *p* = .047. Put differently, participants who revealed experienced more conflict when choosing the selfish option as compared to choosing the prosocial option, while participants who ignored experienced similar levels of conflict independent of their allocation decision (see Figure 4).

Figure 4

Interaction of ignorance choices and allocation decisions in predicting MAD scores

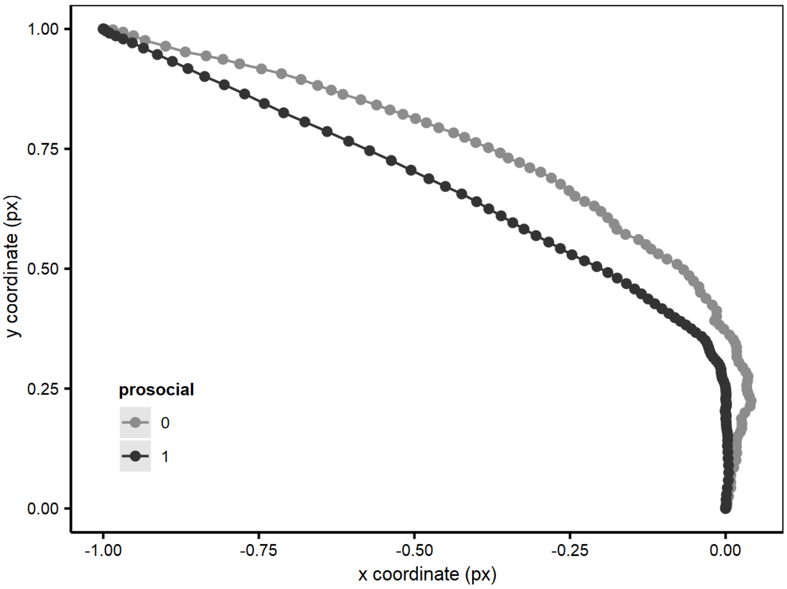


Allocation decisions

To investigate in which ways cognitive conflict was related to prosocial or selfish choices, we tested the relationship between MAD scores and allocation choices (H2). Using a repeated-measure linear regression indicated that participants experienced significantly more conflict when choosing prosocially than when choosing selfishly, β = -0.16, *t*(201.59) = -3.20, *p* = .002 (see Figure 5).

Figure 5

Average mouse trajectories for trials with prosocial and selfish choices in unaligned trials



Of our participants, 53 chose prosocially and selfishly at least once, 35 participants consistently chose the selfish option, and 122 participants consistently chose the prosocial option. A paired t-test only using the subsample of alternating participants did not reach significance, *t*(52) = 0.993, *p* = .163, *d* = .178. A post-hoc sensitivity analysis revealed that we had 80% power to find an effect as small as *d* = .346.

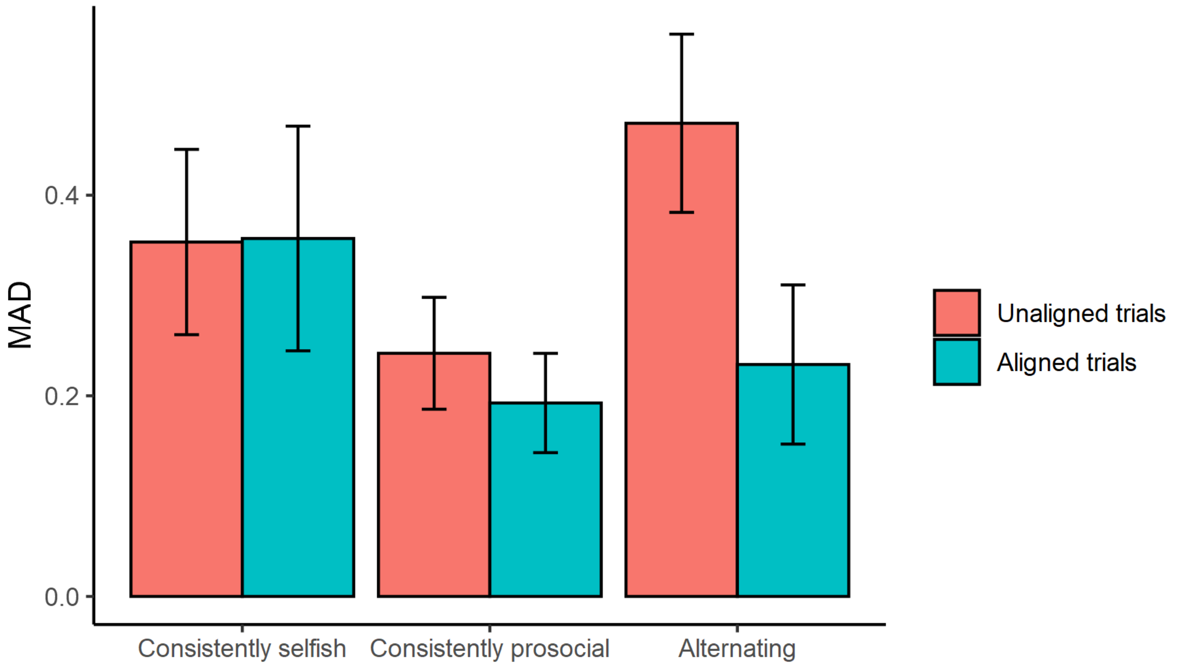
Utilizing the three types of consistently prosocial participants, consistently selfish participants and alternating participants allowed us to gain a deeper understanding of conflict in allocation decisions. A one-way anova revealed significant differences between these three types in terms of MAD scores, *F*(2, 207) = 10.54, *p* < .001. Post-hoc *t*-tests revealed that consistently prosocial participants experienced less cognitive conflict in unaligned trials than consistently selfish participants, *t*(62.43) = 2.017, *p* = .024. Interestingly, participants who alternated between selfish and prosocial choices experienced more conflict in unaligned trials independent of their choice compared to both consistent prosocials, *t*(95.861) = -4.370, *p* < .001, as well as consistently selfish participants, *t*(81.286) = -1.866, *p* = .033.

As each participant can be assigned to one of the three types, we can also compare MAD scores of these types in aligned trials. In fact, consistently prosocial participants also experienced less cognitive conflict than their consistently selfish counterparts in aligned trials, *t*(48.857) = 2.710, *p* = .005. Furthermore, consistently selfish participants experienced more conflict compared to alternating participants, *t*(66.49) = 1.852, *p* = .034, while there was no difference between consistent prosocials and alternating participants, *t*(95.356) =

-.821, *p* = .207 (see Figure 6). When comparing experienced conflict between aligned and unaligned trials, consistently selfish participants did not experience less cognitive conflict in aligned trials, as compared to unaligned trials, *t*(34) = -.097, *p* = .539. We did see a difference in MAD scores between aligned and unaligned trials for consistently prosocials, *t*(121) = 2.228, *p* = .014, as well as for alternating participants, *t*(52) = 6.674, *p* < .001.

Figure 6

Bar graph with MAD scores for consistent prosocials, consistent proselfs and alternating participants in unaligned vs. aligned trials



Next, we investigated the relationship of our interindividual difference measures to ignorance choices, allocation decisions, cognitive conflict, as well as their interactions. For all correlations between interindividual difference measures and behavioral outcomes, see Table 1. SVO was negatively correlated with MAD: More prosocial participants experienced less cognitive conflict averaged across all unaligned trials (Table 1). We also found a significant interaction effect of SVO and allocation decision on MAD: While participants low in SVO experienced a similar degree of conflict independent of what they chose, participants high in SVO felt more conflicted when choosing selfishly compared to choosing prosocially (see Table 1 and Figure 7a). The measures Honesty-Humility and Guilt Proneness followed the same pattern, though the interaction effect of Honesty-Humility and allocation decision on MAD did not reach significance (Table 1). The Fear of Negative Evaluation only correlated negatively with MAD, but had no relation to allocation decision or ignorance (Table 1).

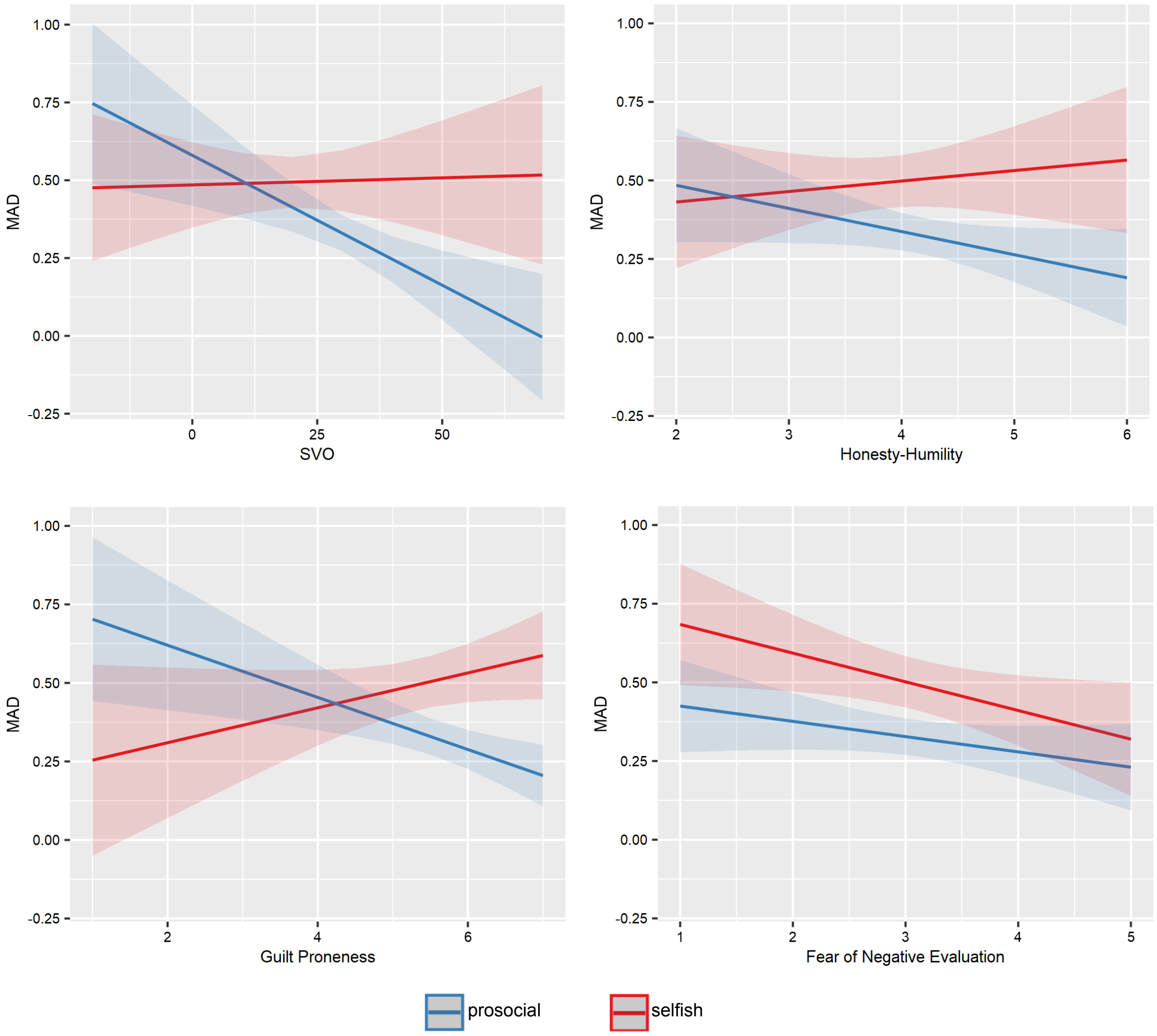
Table 1

Correlations of interindividual differences and behavior/cognitive conflict

	MAD (unaligned)	Allocation decision (unaligned)	Measure x Allocation decision	Ignorance
SVO	***r* = -.302**, *t*(208) = -4.5667, *p* < .001	***r* = .529**, *t*(208) = 8.9867, *p* < .001	*β* **= -.009**, *t*(259) = -2.323, *p* = .021	***r* = -.244**, *t*(208) = -3.629, *p* < .001
Honesty-Humility	***r* = -.116**, *t*(208) = -1.6837, *p* = .047	***r* = .255**, *t*(208) = 3.7959, *p* < .001	*β* = -.107, *t*(259) = -1.613, p = .108	***r* = -.279**, *t*(208) = -4.1965, *p* < .001
GP	***r* = -.168**, *t*(208) = -2.4612, *p* = .007	***r* = .199**, *t*(208) = 2.9222, *p* = .002	*β* **= -.319**, *t*(259) = -3.091, *p* = .002	***r* = -.246**, *t*(208) = -3.668, *p* < .001
FNE	***r* = -.117**, *t*(208) = -1.652, *p* = .050	*r* = -.011, *t*(208) = -0.165, *p* = .565	*β* = .043, *t*(259) = 0.787, *p* = .432	*r* = -.082, *t*(208) = -1.1829, *p* = .119

Figure 7

Predicted values of MAD by the interaction of allocation decision and a) SVO, b) Honesty-Humility, c) Guilt Proneness, and d) Fear of Negative Evaluation



Discussion

Using mouse-tracking as a measure for cognitive conflict, we explored the role of conflict for prosocial decision-making and its relation to willful ignorance. As predicted, cognitive conflict as measured by the curvature of mouse trajectories in unaligned trials was related to willful ignorance (H1). We found an interaction between ignorance and allocation choice in predicting conflict implying that those who ignored tended to experience similar levels of conflict, whereas those who revealed tended to experience less conflict when opting for the prosocial choice instead of the selfish one. We discovered that response trajectories were more curved in unaligned trials than in aligned trials, indicating that choices in unaligned trials entailed more conflict than in aligned trials. Moreover, we found that selfish behavior overall was associated with higher levels of conflict (H2). Finally, the interaction of SVO and allocation decision when predicting cognitive conflict showed that prosocial participants experienced less cognitive conflict when choosing prosocially compared to choosing selfishly, while selfish participants felt similarly conflicted independent of their choice.

Our results provide first correlative evidence that people who experience higher levels of cognitive conflict are more likely to engage in willful ignorance, potentially to avoid this conflict. On its own, this result supports the hypothesis that people exploit situational affordances (i.e., moral wiggle room) to evade the conflict between a should-self and a want-self (Bazerman et al., 1998; Dana et al., 2007; Konow, 2000; Matthey & Regner, 2011). The interaction of ignorance and allocation choice when predicting conflict suggests that people who ignore experience similar levels of conflict, while people who reveal experience less conflict when choosing the prosocial option compared to choosing the selfish option. As such, revealing participants could avoid conflict by choosing prosocially, while this was not the case for ignoring participants. This could suggest that these people ignore to avoid conflict, while others do not need to ignore to reduce cognitive conflict, as they can simply choose the prosocial option instead.

However, ignoring participants did not only experience more conflict compared to revealing participants in unaligned trials, but also in aligned trials. There are two potential reasons for this observation. First, mouse trajectories might not be a valid measure for interindividual differences in cognitive conflict. Mouse trajectories have been validated as a measure for conflict in cognitive processes, such as decision-making, language, social cognition, and learning (Dshemuchadse et al., 2013; Freeman et al., 2011; Koop, 2013; Koop & Johnson, 2011, 2013). Further support comes from neurophysiological evidence (Spivey, 2007). However, there is less evidence for the usefulness of mouse-tracking as a useful measure of interindividual differences in cognitive conflict. As such, mouse trajectories might be useful when comparing groups, but less so as an interindividual difference measure. Future research should conduct psychometric validations of the mouse-tracking measures.

Second, cognitives conflict in our experiment may reflect a more fundamental conflictedness or hesitation when making any kind of decision. For the interpretation of the phenomenon of willful ignorance more broadly, this means that it could be driven by a desire to simplify a decision setting more generally, rather than the desire to avoid a conflict between self-interest and moral standards. This would mean that it is people who generally struggle to make decisions engage in willful ignorance. Under ignorance, there is less information that one has to compare and integrate in order to get to a decision. Thus, ignorance simplifies the decision context, independent of the exact nature of the decision to be made. Research on willful ignorance often assumes that people ignore information to behave selfishly without appearing selfish (see Vu et al., 2023). However, agents might also ignore to simplify the decision setting, independent of its content. First evidence from behavioral data supports this interpretation: People ignore information not only when they make self-other tradeoffs, but also when making allocation decisions between other entities, or when there is no conflict of interests (Cerrone & Engel, 2019; Exley & Kessler, 2021). Our results suggest that large shares of ignorance are driven by a desire to avoid general conflictedness experienced when making any kind of decision.

We also contribute to a more nuanced understanding of the experience of conflict in simple sharing decisions. As mentioned in the introduction, prosociality can be viewed as a conflict within oneself, where one has to balance adhering to social and personal moral standards with self-interest. By comparing aligned and unaligned trials, our study suggests that this intrapersonal conflict can be measured using mouse trajectories, as we find that people are more conflicted in unaligned than in aligned trials. Furthermore, our results indicate that selfish behavior entailed more cognitive conflict than prosocial behavior in binary sharing decisions. However, note that the intraindividual comparison of this effect did not reach significance. This might be due to a large proportion of the sample consistently choosing either the prosocial or the selfish option in all trials, making an intraindividual comparison feasible for only a small fraction of the total sample. Therefore, our study was only powered to find an effect of *d* = .35. Taking the effect reported by Kieslich et al. (2014) of *d* = .32 as a reference suggests that our study is underpowered to find an effect of a similar size. Comparing participants who consistently chose the same option, we observed that consistently prosocial participants experienced less cognitive conflict than consistently selfish participants. This effect was also evident in aligned trials, which could suggest that these trajectories reflect a more generalized pattern that is not specific to self-other tradeoffs. Furthermore, consistently selfish participants experienced similar levels of conflict in aligned and unaligned trials, while consistently prosocials and alternating participants experienced more conflict in unaligned than in aligned trials. This might be considered tentative evidence for selfish participants having the same general decision strategy for aligned and unaligned trials (i.e., “Choose the own-payoff-maximizing option”). This interpretation would be supported by eye-tracking data showing that selfish participants often only process payoffs to the self (Fiedler et al., 2013). However, it remains an open question why selfish participants then experience more cognitive conflict than prosocial participants.

To further understand interindividual variation in the relationship of conflict and prosociality, we investigate their relationship to interindividual difference measures. We observed a similar pattern between sharing decisions, conflict and SVO, Honesty-Humility, as well as Guilt Proneness respectively: All three measures were positively related to prosocial choices, negatively related to willful ignorance, as well as showing a similar pattern with regards to the interaction of each of the measures with allocation decision in predicting conflict scores (however insignificant interaction term for Honesty-Humility). Generally, dispositionally prosocial individuals experienced more conflict when choosing selfishly than when choosing prosocially, while participants with a selfish disposition felt similarly conflicted regardless of their choice. These interactions are intriguing when considering the role of willful ignorance in our setup. Willful ignorance is generally negatively correlated with SVO, Guilt Proneness, and Honesty-Humility. Thus, it is typically less prosocial and less guilt-prone individuals who choose to remain ignorant. As dispositionally less prosocial and less guilt-prone individuals feel conflicted regardless of their choice, ignorance might be used as a means of reducing this conflict. Prosocial and guilt-prone participants, on the other hand, can avoid this conflict by selecting the prosocial option. The Fear of Negative Evaluation scale only showed a significantly negative correlation with conflict: Participants high in Fear of Negative Evaluation felt less conflicted both when choosing selfishly as well as when choosing prosocially, which is surprising. We did not find any support for image concerns as measured by the Fear of Negative Evaluation playing a role in prosociality or willful ignorance in our setup.

This study contributes to the empirical evidence on cognitive dissonance in prosociality (Konow, 2000). This theory is built upon the work of Leon Festinger (1957), who proposed that agents experience an unpleasant emotional reaction of cognitive dissonance when holding two conflicting desires. In the domain of prosociality, these conflicting desires can be the desire to be moral and fair, versus the opposing desire to maximize one's own payoffs. To reduce this tension, the agent can either engage in prosocial behavior or engage in self-deception (Konow, 2000). Our findings support and expand on this interpretation by adding a more nuanced understanding of how interindividual differences in disposition prosociality influence the relationship between cognitive conflict and prosociality. We found that dispositionally prosocial participants avoid potential negative emotional reactions brought about by cognitive dissonance by behaving prosocially, whereas dispositionally selfish participants experience conflict irrespective of their choice. In the context of cognitive dissonance theory, this suggests that only a subset of agents (i.e., the dispositionally prosocial) can decrease their cognitive dissonance by behaving prosocially, while another subset (i.e., the dispositionally selfish) must resort to self-deception strategies such as willful ignorance to avoid cognitive dissonance. One possible explanation for this phenomenon is that dispositionally prosocial individuals hold weak self-interests, so acting in accordance with their fairness norms generates little cognitive dissonance. On the other hand, dispositionally selfish individuals may have strong self-interest and strong fairness norms, resulting in cognitive dissonance regardless of their choice.

Limitations

While our study provides valuable insights into charitable giving and prosociality, as well as willful ignorance, there are limitations that must be acknowledged. One limitation is that our experimental design does not allow us to establish causality between cognitive conflict and willful ignorance. Our findings on conflict and prosociality, as well as their interaction with different interindividual differences measures suggest that it is the dispositionally selfish participants who experience more conflict. Because dispositionally selfish participants are more likely to ignore, participants' social dispositions may be a confounding factor that impacts both conflictedness and the propensity to ignore independently of each other. This could also explain why higher conflict is associated with a higher likelihood of ignoring when participants choose to act prosocially, but with a lower likelihood of ignoring when they choose to act selfishly: Dispositionally more selfish participants experience more conflict when choosing prosocially, but similar conflict levels as prosocials when choosing selfishly. Because of the correlation of dispositional prosociality and willful ignorance, this pattern also translates from interactions of allocation decision with dispositional prosociality to interactions with willful ignorance. Future research should experimentally manipulate conflictedness directly and observe its effects on willful ignorance to establish causality.

A second limitation is that a large share of our sample was very consistent in their choices, resulting in low power for intraindividual comparisons. The mouse-tracking items were designed to reflect a similar self-other conflict as in the ignorance setting, so as to answer the question of whether participants who experience more conflict in these kinds of tradeoffs are also the ones engaging in willful ignorance. As a result, the set of items was very low in variability. In future research, a more diverse set of items could be employed to increase the power to compare intra- and interindividual experiences of conflict and determine whether the experience of conflict is specific to self-other tradeoff situations or more broadly represents a fundamental conflictedness in decision contexts

Third, the level of willful ignorance observed in our study is significantly lower than reported in the literature. Only 22% of our sample ignored the consequences of their choice on the charity. A meta-analysis on willful ignorance found a mean of 40% of participants choosing to remain ignorant (Vu et al., 2023). Additionally, we observed high levels of prosocial behavior, both in the ignorance setting (77% prosocial choices) and in the mouse-tracking part (74% prosocial choices). These similar levels of prosociality suggest that the option to ignore may not have significantly impacted participants' allocation decisions. Further research should investigate the level of cognitive conflict after establishing the impact of willful ignorance on behavior in the respective setting.

Conclusion

In summary, our findings support the hypothesis that willful ignorance is linked to cognitive conflict in self-other tradeoffs. However, we also observed this relationship in decision settings where no conflict was present, indicating that participants may use willful ignorance as a means of simplifying decision-making more broadly. Regarding cognitive conflict and prosociality, we found that choosing prosocially generally entails less cognitive conflict. An interaction of SVO and allocation decision indicated that dispositionally selfish participants experienced conflict independent of their choice, while dispositionally prosocial participants experienced less conflict when choosing the prosocial option. Thus, willful ignorance could function as a self-deception strategy for dispositionally selfish participants in order to avoid the conflict of choosing either option.

Chapter 4:

Who ignores - and why?

This chapter is based on tho Pesch, F., Fiedler, S., & Baumert, A. (prepared for submission). Who ignores - and why?

Abstract

Our study investigated the motivations behind willful ignorance in prosocial decision-making, a behavior that has negative societal impacts. Willful ignorance has traditionally been assumed to be motivated by a desire to act selfishly without appearing so (wiggling-related motives), alternative motivations such as tradeoff aversion and inattention have not been thoroughly examined. In a study with 878 participants, we manipulated the context and accessibility of information within-subject, and administered dispositional measures to identify patterns of behavior and infer underlying motives. Results showed that 41% of ignorance was motivated by wiggling-related motives, 19% by tradeoff aversion, and 33% by inattention. These findings shed light on the motivations behind willful ignorance and suggest that the majority is not driven by a desire to act selfishly without appearing so.

Keywords: Willful ignorance, moral wiggle room, prosociality, interindividual differences

Authors' note: For all data, analysis scripts, pre-registration and materials see https://osf.io/p3a2g/?view_only=531aa932f3d94fe49bbc6447fba3cde1.

Introduction

Steering clear of the news after a natural disaster in order to avoid calls for donations, or neglecting the consequences of our shopping behavior for the environment: We all know situations in which we willfully ignore information that would otherwise impact our choices so that we do not feel compelled to adapt our behavior. Empirical research has corroborated the effect of willful ignorance on prosocial behavior in the domain of charitable giving: People ignore information that would suggest them to give more to charity and are more likely to choose the selfish option as a result (Dana et al., 2007; Exley, 2016). This form of willful ignorance predominantly has been interpreted as a desire to avoid the appearance of selfishness, either to oneself or others. In the present study, we aim to critically test this interpretation by exploring alternative mechanisms that may drive willful ignorance, namely tradeoff aversion and inattention.

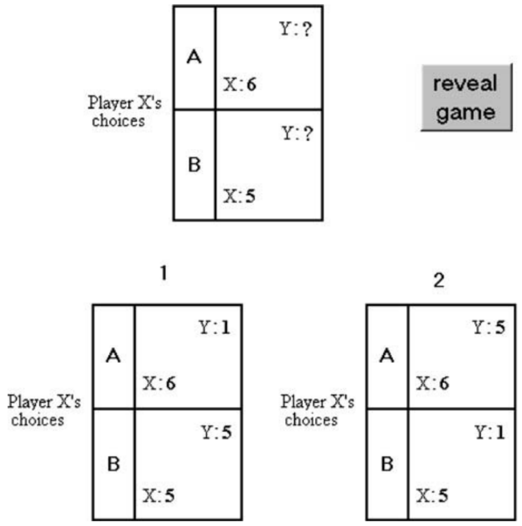
Willful ignorance

The idea that knowledge is power (Bacon, 1857) has been widely accepted, yet research has shown that people may ignore crucial and freely available information that would otherwise be instrumental in their decision-making (Hertwig & Engel, 2016). Rational choice theory suggests that people value information to the extent that it helps them make more informed decisions (Stiegler, 1961). Therefore, more information is generally seen as better, unless the costs of acquiring it outweigh its benefits. As a result, free information should always be sought out, especially when it is instrumental for a decision. However, empirical evidence on the concept of *willful ignorance* in social decision-making challenges this notion. Specifically, people might avoid information that otherwise would influence their decisions (Dana et al., 2007). The effect of willful ignorance on allocation decisions has been widely studied, with evidence indicating that it plays a role in prosociality and direct charitable giving (Dana et al., 2007; Exley, 2016; Grossman, 2014; Grossman & van der Weele, 2017; Larson & Capra, 2009; Matthey & Regner, 2011).

Most research on willful ignorance in social decision-making has used the *hidden information treatment* developed by Dana et al. (2007). In this seminal study, the authors challenged the assumption that prosocial behavior is driven purely by a preference for prosocial or fair outcomes, whether due to a "warm glow" (Andreoni, 1990) or an aversion to unfairness (Fehr & Schmidt, 1999). Using different versions of a binary dictator game, Dana et al. (2007) found that people behave less prosocially when the intentions behind their choices are not revealed by the observable outcomes. In the baseline treatment, participants made an allocation decision between themselves and a passive recipient. Choosing the prosocial option would result in the same payoff for both the participant and the recipient, while the selfish option would give a higher payoff to the participant and a lower payoff to the recipient. In the hidden information treatment, participants had the option to stay uninformed about the consequences of their choice for the recipient, remaining unaware of whether their choice was selfish or not (see Figure 1). Dana et al. (2007) found that participants who had the option to ignore were less likely to choose the prosocial option compared to those in the baseline treatment. This change in allocation decisions suggests that the ignored information would have influenced the decision if it had been known (for a review, see Vu et al., 2023).

Figure 1

The hidden information treatment of Dana et al. (2007)



The prevalent explanation for the effect of willful ignorance is that individuals wish to maximize their own benefits without appearing selfish. As such, people are thought to exploit the *moral wiggle room* offered by the option to ignore. In that sense, ignorance is seen as a strategic tool to maintain a certain self-concept (i.e., of being a moral and prosocial person), even while taking actions that suggest the opposite. The concrete hypothesized channels through which ignorance functions are diverse, but all engage the notion of wiggling (i.e., behaving selfishly while not wanting to appear selfish). Some authors proposed that people engage in willful ignorance for image reasons: By willfully ignoring the consequences of their behavior, agents can plausibly claim that they would have acted virtuously if they had been informed, meaning that they can still uphold a positive self- and social image (Adena & Huck, 2020; Andreoni & Bernheim, 2009; Exley, 2016; Feiler, 2014; Grossman, 2014; Grossman & van der Weele, 2017; Larson & Capra, 2009). Theoretical work on image and signaling assumes that individuals care about how they are perceived by others and how they perceive themselves, and that this desire is reflected in their actions (Bénabou & Tirole, 2006, 2011; Bodner & Prelec, 2003; Grossman, 2014; Grossman & van der Weele, 2017; Nyborg, 2011; Rabin, 1995). Social image or social signaling accounts proclaim that people infer other people's personality or intentions from observing their behavior (Bénabou & Tirole, 2011; Grossman, 2015). Similarly, self-image or self-signaling accounts surmise that people also infer their own type or motivations from observing their own behavior (Bem, 1972; Blasi, 1980; Grossman & van der Weele, 2017). As a result, any choice is not only a causal act leading to certain material outcomes in the world, but also sends a signal about the agent's motivation and intention. As attributes such as fair-mindedness and prosociality are not directly observable to outsiders and are difficult to introspect, social decision can be seen as an indicator for the agent's type (Grossman, 2015; Kelley, 1967). Within these frameworks, willful ignorance can function as a deception strategy to act selfishly without sending a strong signal about one's type, either to others or to oneself.

Also resonating with the concept of wiggling, some researchers have stressed the emotional responses as psychological channels through which willful ignorance affects prosocial behavior. For example, it has been hypothesized that ignorance allows agents to choose self-servingly while feeling less guilty (Feiler, 2014; Garcia et al., 2020; Thunström et al., 2014), or while avoiding to feel conflicted between one's own moral standards and a desire to maximize one's payoffs (Grossman, 2014; Lin & Reich, 2018; Matthey & Regner, 2011; Woolley & Risen, 2018). Yet other researchers have proposed that willful ignorance reduces the social pressure to give, or changes the social norm in a given situation (Dana et al., 2007). Though heterogeneous in their exact channels, these mechanistic accounts ultimately all have one element in common: They assume that the main driver for willful ignorance is related to a desire to choose the selfish option without having to admit selfish motivation (i.e., exploiting moral wiggle room). In the following, we refer to these mechanisms as the wiggling-related mechanisms.

There is empirical evidence pointing to wiggling-related motivations as drivers of the willful ignorance effect (Feiler, 2014; Grossman & van der Weele, 2017; Matthey & Regner, 2011). For example, van der Weele (2014) showed that in line with self-image accounts, people were less likely to ignore relevant information when prosocial behavior was inexpensive. This is presumably because when obtaining a positive self-image is cheaper, revealing the information is less threatening. In line with both self- and social image accounts, participants stated that they engaged in ignorance to avoid a bad conscience, or avoid having to be nice (van der Weele, 2014). With regards to the emotional channels, and the role of conflict in particular, (Matthey & Regner, 2011) found that participants who chose prosocially in the baseline conditions, but then ignored when possible, reported more cognitive conflict than consistently prosocial participants. Grossman and van der Weele (2017) showed that it is participants with a medium score on their Social Value Orientation (i.e., dispositional inclination to behave prosocially) who ignore, meaning that these people are not pronounced prosocials, but also not extremely selfish. This can be seen as further support for the role of conflict. Taking the perspective of an outsider observing the behavior of others, research corroborated that ignoring information reduced the social disapproval of selfish behavior (Krupka & Weber, 2013) and the respective punishment (Bartling et al., 2014; Conrads & Irlenbusch, 2013), thus lending plausibility that ignorance can indeed serve as an effective excuse.

While the wiggling-related mechanism is supported by empirical evidence, the literature lacks a thorough investigation of potential alternative explanations. However, there is some empirical evidence that other motivations might account for (part of) the willful ignorance effect. For example, some researchers have found that some people ignore their own payoffs when they are initially hidden and can be revealed with a click on a button (Kandul & Ritov, 2017; Moradi, 2018). Other evidence suggests that willful ignorance extends beyond self-other tradeoffs and also occurs when making decisions for others (Cerrone & Engel, 2019), or distributing money as a 3rd party (Exley & Kessler, 2021). Given the structure of the decision space, there are alternative mechanisms that could account for willful ignorance without alluding to prosocial versus selfish motivations. We identified two candidates in the literature: tradeoff aversion and inattention.

Alternative mechanisms

Tradeoff aversion refers to the idea that people may prefer not to make a decision in certain situations. While choice opportunities are generally assumed to be desirable because they increase the chances of finding an option that aligns with one's preferences (Kreps, 1979), they can also lead to higher cognitive load, as well as negative emotions such as regret, temptation, and fear of making a bad decision (Loewenstein, 1996). Empirical research has shown that people are decision averse in certain situations, and that this decision aversion is indeed related to anticipated regret and fear of blame for making bad decisions (Beattie et al., 1994; Le Lec & Tarroux, 2020). In the context of the hidden information treatment by Dana et al. (2007), participants can avoid having to make a tradeoff decision by choosing to stay ignorant. This is because under ignorance one option is strictly dominant to the other. In the hidden information treatment, the dominant option happens to be the selfish option, so a desire to avoid making a tradeoff would lead to more selfish choices, even if the individual is not motivated by wiggling. This might lead to an overestimation of wiggling-related mechanisms.

A second plausible mechanism is inattention. Inattention means that people remain ignorant due to a general lack of attention within the decision context. This may be due to being overwhelmed or uninterested. There is empirical evidence for this idea, indicating that a significant portion of ignorance is due to the specific choice architecture of the context: Changing the informational default in a study resulted in a significant decrease in ignorance (Grossman, 2014). This suggests that ignorance may largely disappear when it cannot be chosen passively, which can be seen as evidence for a strong default effect (although see Larson & Capra, 2009). As inaction automatically leads to choosing the default, inattention is one candidate for explaining why defaults have such a strong effect on choices. In the hidden information treatment, inattention might lead people to avoid revealing and simply to choose the option with the highest payoff under ignorance, which happens to be the selfish option in the classic setup of Dana et al. (2007). Therefore, inattention can increase selfish choices in this setup regardless of any selfish motivations, again leading to an overestimation of wiggling-related motivations as drivers of the willful ignorance effect.

In our study, we aim to trace the extent to which these three motivations (i.e., wiggling, tradeoff aversion and inattention) contribute to the effect of willful ignorance on social decision-making. To this end, we utilize the classic setup of Dana et al. (2007) and adaptations thereof. Understanding the drivers of willful ignorance within this paradigm allows us to derive implications for boundary conditions of prosocial and ethical behavior in everyday life. This investigation is crucial for understanding situational influences on prosocial behavior, and also allows us to draw conclusions about the motivational foundations of prosociality more generally.

Present study

To this end, we examined the motivational drivers of ignorance combining three different empirical approaches. First, we varied the decision context in which participants made allocation and ignorance decisions. Second, utilizing a within-subject manipulation of these contexts allowed us to determine typologies by combining intraindividual patterns of behavioral outcomes from allocation and ignorance choices in different contexts. Third, we measured interindividual differences in several predispositions which were conceptually linked to the motivations in question. In the following, each of these empirical approaches is explained in detail.

Decision contexts

We inferred in which way wiggling, tradeoff aversion and inattention contribute to the effect of willful ignorance on prosocial behavior by having participants make allocation decisions in three different types of binary dictator game contexts: the Self-Other (SO-) context, the Other-Other (OO-) context, and the No-Tradeoff (NT-) context. Each context had a baseline condition and an ignorance condition, in which some payoff information was initially hidden (i.e., parallel to the hidden information condition of Dana et al., 2007).

In the SO-context, participants were asked to distribute money between themselves and a charity. In the SO-baseline condition, they faced a decision between a selfish and a prosocial option. As SO-ignorance condition, we used the selfish ignorance condition, in which the payoffs to the charity were hidden and could be revealed by clicking a button (i.e., the hidden information treatment of Dana et al., 2007).

In the OO-context, participants again made binary allocation decisions, this time distributing money between two charities. In the OO-baseline, they made a decision between an efficient and a fair option. As OO-ignorance condition, we used the efficient ignorance condition, in which some payoff information was hidden, so that the efficient option was the one with the higher payoff under ignorance. As a result, willful ignorance should lead to an increase in efficient choices.

In the NT-context, participants again distributed money between two charities, but this time one charity received the same donation independent of the participant’s choice, while the other charity received more in one option than the other. This way, one option was always strictly dominant (i.e., better for both charities). In the NT-baseline condition, participants made a decision between this strictly dominant option and an antisocial option, in which the second charity received a lower payment. In the antisocial ignorance condition, the payoffs to this second charity were hidden.

With this design, we have different motivations plausibly driving ignorance in different contexts. In the selfish ignorance condition (SO-context), wiggling, tradeoff aversion, and inattention can all be seen as potential motivators for ignorance. By moving from the selfish (SO-) to the efficient ignorance condition (OO-context), we eliminate wiggling as a plausible motivation because the participant's own payoff is not at stake in this decision. While the choice in the efficient ignorance condition (OO-context) still involves a tradeoff between an efficient and a fair option, the NT-context removes tradeoff aversion as a plausible motivation, leaving only inattention as a potential plausible reason for ignorance in the antisocial ignorance condition. Therefore, the differences in ignorance levels between these different contexts can give us an indication of which motivations are likely at play.

Typologies

Utilizing the within-subject manipulation of our study design, we can identify exactly who ignores in which context, and how this affects their allocation choices. We came up with two typologies. The first typology identified who engaged in moral wiggling (i.e., allocation types): We categorized participants into moral wigglers, consistent prosocials and consistent selfish types by comparing their allocation choices in the SO-baseline with their allocation choices in the selfish ignorance conditions. The second typology determines different ignorance types, categorizing participants into consistent revealers, wiggling ignorers, tradeoff ignorers and inattentives. Consistent revealers did not engage in ignorance in any of the three ignorance conditions. Wiggling ignorers are participants who ignored exclusively in the selfish ignorance condition. Tradeoff ignorers ignored both in the selfish and the efficient ignorance condition, while inattentives ignored in all three ignorance conditions. Finally, we can combine both typologies to see which type of ignorers engaged in ignorance within the selfish ignorance condition, as well as which type of ignorers were most likely to engage in moral wiggling as defined above.

Dispositional measures

We further complemented our exploration of the motivational mechanisms of (patterns of) behavior, by linking differences in behavioral patterns across conditions with measures of dispositional characteristics. Specifically, we identified dispositional concepts that capture individual differences in those motivations, and how to measure them.

With regards to the motivation of *wiggling*, people’s dispositional tendencies to behave prosocially or selfishly are connected to willful ignorance. It has been proposed that it is people with moderate prosocial tendencies who exploit moral wiggle room, meaning that they are not pronounced prosocials, but also not extremely selfish (Grossman & van der Weele, 2017). As people who ignore are not all wiggling, but potentially also convinced selfish participants who are not interested in the hidden information, we would assume that the more dispositionally selfish a person, the more likely they are to ignore. For wiggling behavior itself, we would expect people with a moderate prosocial disposition to engage in moral wiggling. Social Value Orientation (SVO), which is a concept measuring stable preferences for joint outcomes and cooperation (prosocial values) as compared to a pro-self orientation (P. Van Lange, 1999), has been shown to reliably predict prosocial behavior

(Balliet et al., 2009; Murphy et al., 2011; Smith, 2012; Van Lange et al., 2007; see (Thielmann et al., 2020 for a review).

The motivation of *tradeoff aversion* describes a reluctance to make tradeoff decisions. People differ in how much they value making their own decisions (Beattie et al., 1994). These interindividual differences can be captured by the concept of Need for Cognitive Closure (Webster & Kruglanski, 1994). People higher (vs. lower) in Need for Closure tend to take decisions quickly (if necessary) and avoid them (if possible; Kruglanski, 1989; Kruglanski & Webster, 1996). A study relying on vignettes suggests that people who are high in Need for Closure are also more likely to be decision averse (Otto et al., 2016).

The motivational factor of *inattention* captures the extent to which people pay attention to a situation. It can be conceptually linked to the Need to Evaluate (Jarvis & Petty, 1996). The concept captures a personality trait that reflects a person's proclivity to create and hold attitudes (Bizer et al., 2004). People high in Need to Evaluate are especially likely to form attitudes towards all sorts of objects. If willful ignorance is partly driven by any form of inattention, people who have a stronger attitude towards a topic should be less likely to ignore information due to inattention, as they should be more engaged in the decision.

Furthermore, we explored additional dispositional measures related to the three motivations above. The factor of wiggling can also be conceptually linked to image concerns, or how much people care about what others think of them. Image concerns have long been thought to drive wiggling (Adena & Huck, 2020; Andreoni & Bernheim, 2009; Grossman, 2015; Grossman & van der Weele, 2017). Image concerns can be operationalized by the Brief Fear of Negative Evaluation Scale (Reichenberger et al., 2015). The scale captures a person's tolerance for the possibility of being judged negatively by others. Furthermore, there are alternative ways of measuring selfishness. Instead of using SVO as a measure for selfishness, we could also utilize the factor Honesty-Humility of the HEXACO model (Ashton & Lee, 2007), or the factor Guilt Proneness of the Guilt and Shame Proneness scale (Cohen et al., 2011) to capture people's propensity to engage in prosocial behaviors (see Thielmann et al., 2020 for a review). Another way to capture tradeoff aversion is the Desirability of Control scale (Burger & Cooper, 1979), while inattention could also be captured by the factor Conscientiousness of the HEXACO model (Ashton & Lee, 2007). We thus explored the relationship of our behavioral outcomes with these dispositional measures to be able to identify which concept is best at capturing the underlying motives of behavior.

Additional decision context: prosocial ignorance condition

We further consolidated the wiggling motivation for willful ignorance in the selfish ignorance condition by comparing ignorance and allocation choices in the selfish ignorance condition to a second ignorance condition within the SO-context: the prosocial ignorance condition. The setup of this condition was similar to the selfish ignorance condition, only here the payoffs to the agent themselves were initially hidden instead of the donations. This condition allowed us to further specify in which way ignorance in the selfish ignorance condition was due to a desire to avoid the self-other tradeoff more generally, or whether it was driven by wiggling motives. If ignorance in the selfish ignorance condition was driven by wiggling-related motives, participants ignoring in this condition should not ignore in the prosocial ignorance condition. If participants, however, were motivated to avoid the self-other tradeoff decision in the SO-context independent of which information was hidden, we expect that people who ignored in the selfish ignorance condition should also ignore in the prosocial ignorance condition. The prosocial ignorance condition also allowed us to add to the literature on prosocial ignorance (Kandul & Ritov, 2017; Moradi, 2018).[11]

Hypotheses

Based on our design, we hypothesized that overall, participants would be more likely to choose the option with the higher payoff when the option to stay ignorant was present (rather than absent, H1). This means that in the selfish ignorance condition, participants would be more likely to choose the selfish option, while in the efficient ignorance condition, they would be more likely to choose the efficient option, and in the antisocial ignorance condition, they would be more likely to choose the antisocial option as compared to their respective baselines.

We predicted that there was an interaction of the ignorance cognition and the decision context on choice behavior: The impact of the introduction of the ignorance manipulation on behavior should decrease from the SO- to the OO- to the NT-context (H2).

We also predicted that the frequency of ignorance would be higher in the selfish ignorance condition compared to the efficient ignorance condition and again lower compared to the antisocial ignorance condition (H3). We expected this pattern because, as the motivation for ignorance should decrease between these settings: While in the selfish ignorance condition,

[11] We also added a second ignorance condition to the OO-context. In the fair ignorance condition, the fair option has the higher payoff under ignorance, and thus ignorance should increase the share of fair choice. We originally planned to collapse the fair and the efficient ignorance condition into one OO-ignorance condition, which was not possible. For all analyses on the fair ignorance condition, see Appendix C.

all three motivations (wiggling, tradeoff aversion and inattention) should promote ignorance, the wiggling motivation is eliminated in the efficient ignorance condition, and for the antisocial ignorance condition, only inattention should play out.

In hypothesis 4, we specified the relationship of our three main dispositional measures in predicting ignorance choices: SVO should predict ignorance in the selfish ignorance condition better than in the other ignorance conditions (H4a); Need for Closure should predict ignorance in the selfish and efficient ignorance condition better than in the antisocial ignorance condition (H4b); and Need to Evaluate should predict ignorance across all ignorance conditions (H4c).

We furthermore explored the intraindividual patterns of behavior, and connect these patterns back to the different motivations we suspect to drive behavior in the different contexts. If people ignore for wiggling reasons, they should only ignore in the selfish, but not in the efficient or antisocial ignorance conditions. If tradeoff aversion is the reason for ignorance, participants should ignore in the selfish and efficient, but not in the antisocial ignorance condition. If participants ignore in all three conditions, these participants are likely motivated by aspects related to inattention.

Methods

Transparency statement

All data, materials, analysis scripts, and pre-registration of design, hypotheses, and statistical analyses for this study are available at https://osf.io/p3a2g/?view_only=531aa932f3d94fe49bbc6447fba3cde1. We excluded the fair ignorance condition from the remainder of this paper, for all analyses on this condition see Appendix C. In the results, we will partly report different analyses than the pre-registered ones for illustrative purposes. All analyses also hold when using the pre-registered models. For all pre-registered analyses including pre-registered exploratory analyses, see Appendix A.

Participants

In four data collection waves in March 2022, we collected data from 878 participants on Amazon Mechanical Turk (53.48% identified as female). We started with a total sample of 1110 participants in wave 1, thus having an overall attrition rate of about 20% (11 dropouts after wave 1, 136 dropouts after wave 2, 44 dropouts after wave 3, and 41 dropouts after wave 4). After the first wave, we excluded three participants who failed the attention check and four participants who failed control questions. All other dropouts were participants who failed to participate in the next wave within 48 hours. Participants who failed the control questions in wave 2 to 4 were excluded from making a decision in the respective context, but were nonetheless invited to further participate in the study, thus generating missing values in the data.

Design and procedure

In a within-subject design over four data collection waves spread over ten days, participants faced eight different allocation decisions in three Self-Other conditions (including the less central prosocial ignorance condition), three Other-Other conditions (including the fair ignorance condition, see Appendix C), and two No-Tradeoff conditions. Five of these contexts also involved the decision of whether or not to ignore. We counterbalanced the order in which participants faced the eight different allocation decisions among three counterbalancing groups (see Table 1). In the first wave, all participants answered to a battery of dispositional measures, and then made their choice in the baseline condition of the No-Tradeoff context. In the following three waves, participants always made one of the three Self-Other context decisions before being exposed to an Other-Other context. In the fourth wave, participants eventually made their choice in the ignorance condition of the No-Tradeoff context. After each baseline condition, participants answered a short post-decision questionnaire about how they felt when making the decision.

Participants were informed that the first survey was part of a series of four surveys that would be sent over the next ten days. On the first page of the survey, we asked participants only to take the survey if they were willing to participate in the following three surveys. The surveys were sent out at three-day intervals, and participants had 48 hours to complete the survey after receiving it. To minimize attrition we offered a $3 bonus payment for participants who completed all four surveys. Participants were informed that the first survey would take approximately 20 minutes to complete, while the other three surveys would take about 6 minutes each.

Table 1

Counterbalanced order of eight decisions context for the three counterbalancing groups

	1. Group (SO-baseline)	2. Group (selfish ignorance)	3. Group (prosocial ignorance)
Wave 1	Dispositions NT-baseline	Dispositions NT-baseline	Dispositions NT-baseline
Wave 2	SO-baseline OO-fair ignorance	SO-selfish ignorance OO-baseline	SO-prosocial ignorance OO-efficient ignorance
Wave 3	SO-selfish ignorance OO-efficient ignorance	SO-prosocial ignorance OO-fair ignorance	SO-baseline OO-baseline
Wave 4	SO-prosocial ignorance OO-baseline NT-antisocial ignorance	SO-baseline OO-efficient ignorance NT-antisocial ignorance	SO-selfish ignorance OO-fair ignorance NT-antisocial ignorance

Note: Abbreviations: SO = Self-Other; OO = Other-Other; NT = No-Tradeoff

Incentivation

Participants earned an average of $6.98 and donated $4.75 on average throughout the four waves of data collection. In the first wave, participants earned a flat fee of $0.70, and received additional bonus payment according to one randomly drawn incentivized SVO decision. In wave two to three, participants received a flat fee of $0.10, and a decision-contingent bonus payment of between $0.50 and $0.60. In all waves, donations were for the charities Feeding America, The American Red Cross, and Direct Relief.

Material

Pre-study

In a pre-study, we first wanted to conceptually replicate the results of Dana et al. (2007) in an online setting using charities as recipients in order to confirm that the setup will be suitable for our purposes. Thus, we recruited 100 MTurkers, two of which failed the comprehension questions and were therefore excluded from the experiment before entering the decision stage. We fully incentivized all choices. Participants received a flat fee of $0.35 and a bonus payment of $0.50 to $0.60.

The data revealed a significant main effect of our manipulation on prosocial behavior: participants in the baseline condition were more likely to choose the prosocial option (69.4%) compared to those in the ignorance treatment (34.4%), Chi^2(1) = 8.363, *p* = .004, Cramer's V = -.351. 56.5% of participants ignored the information on the consequences of their behavior if they could. These results were very similar to the original results of Dana et al. (2007) in which prosociality dropped from 74% to 38%, while 44% of participants chose to ignore the information.

At the end of the experiment, we pre-tested the five most popular charities in the US to identify charities that are similar to one another. For all results, see Appendix A.

Decision contexts

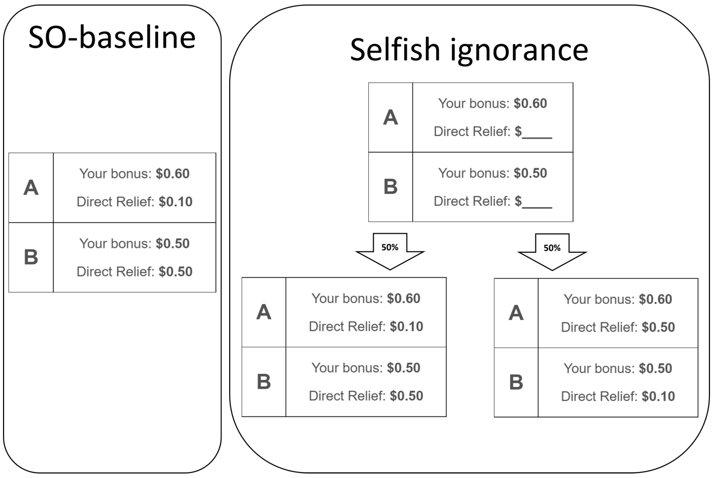
Participants faced eight different allocation decisions in three different decision contexts: (1) The Self-Other contexts, (2) the Other-Other contexts, and (3) No-Tradeoff contexts. In all decision contexts, participants made a binary allocation decision between two options, A and B.

Self-Other (SO-) context

In the SO-context, participants chose how to distribute money between themselves and the charity "Direct Relief" (see Figure 2). In the SO-*baseline condition*, all payoffs were visible when choosing between a selfish option (A) and a prosocial option (B). Option A would result in a payoff of $0.60 for the recipient, and a donation of $0.10. Option B would mean $0.50 for both the participant and the charity. In the *selfish ignorance condition*, participants faced the same options, but the payoffs of the charity were initially hidden. Participants had the option of either choosing option A or B without knowing the donation attached to these options, or revealing the payoffs of the charity before the choice. Participants were informed that it would be randomly decided whether participants would face unaligned payoffs (i.e., the payoff structure of the SO-baseline condition), or aligned payoffs (i.e., a payoff structure in which the payoffs for the charity would be flipped). In the latter case, option A would be better for both the participant and the charity.

Figure 2

Schematic representation of the choice context in the SO-baseline and the selfish ignorance condition

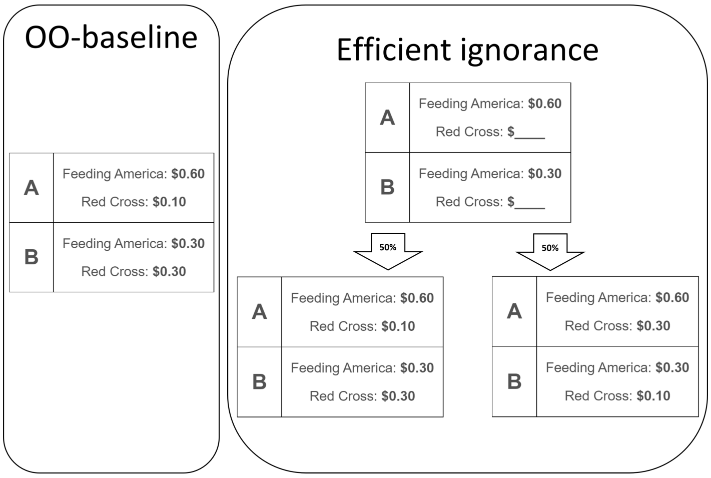


Other-Other (OO-) context

In the OO-context, participants distributed money between the two charities “American Red Cross” and “Feeding America”, facing a tradeoff between a fair and an efficient option (see Figure 3). We counterbalanced between subjects which charity would be associated with option A or B. In the OO-*baseline condition*, participants could choose between an efficient option (A) and a fair option (B). Option A would result in a donation of $0.60 to charity 1, and a donation of $0.10 to charity 2. Option B would mean $0.30 for each charity. In the *efficient ignorance condition*, payoffs of the charity with the lower payoffs were hidden. This meant that the efficient option was the one with the higher payoff under ignorance. Similar to the SO-condition again, the payoffs of both charities were aligned in half of the cases and unaligned in the other half.

Figure 3

Schematic representation of the choice context in the OO-baseline and the efficient ignorance condition

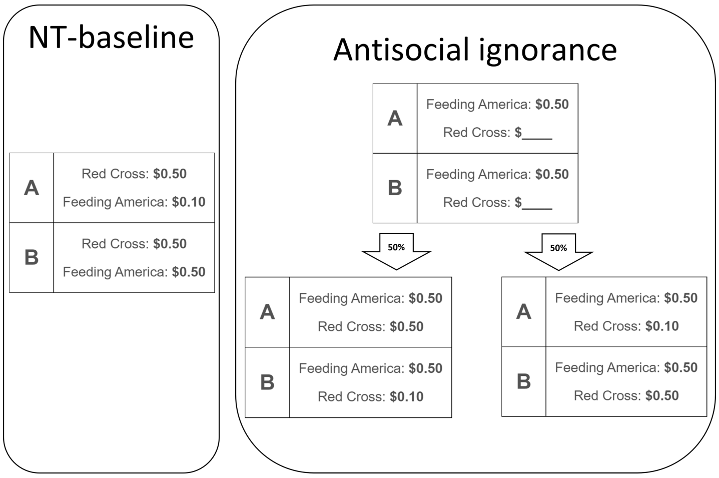


No-Tradeoff (NT-) context

In the NT-context, participants again distributed money between the two charities, “American Red Cross” and “Feeding America”, counterbalancing between subjects which charity would be associated with option A or B (see Figure 4). In the NT-*baseline condition*, participants could choose between an antisocial option (A) and a fair and efficient option (B). Option A would result in a donation of $0.50 to charity 1, and a donation of $0.10 to charity 2. Option B would mean $0.50 for each charity. In the *antisocial ignorance condition*, payoffs to one charity were hidden, so that participants did not know which option was the antisocial option. Again, participants could inform themselves about the full payoff structure by clicking a “reveal” button.

Figure 4

Schematic representation of the choice context in the NT-baseline and the antisocial ignorance condition

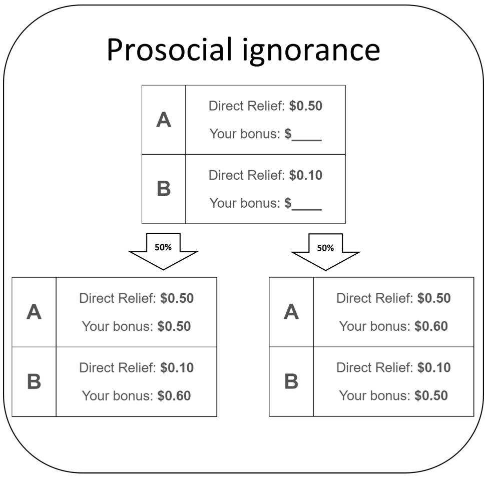


Prosocial ignorance

The *prosocial ignorance condition* was part of the SO-context. Allocation choices in this condition were compared to choices in the SO-baseline condition in order to test whether prosocial choices increased when ignorance was possible. In the prosocial ignorance condition, option A was the prosocial option, while option B was the selfish option (see Figure 5). While in the selfish ignorance condition payoffs to the charity were hidden, this time the payoffs of the participants themselves were hidden. Again, participants could decide for either option without knowing the payoffs to themselves or reveal the payoffs. Participants were told they would be randomly assigned to either aligned or unaligned payoffs.

Figure 5

Schematic representation of the choice context in the prosocial ignorance condition



Dispositional measures

In wave 1, all participants filled out a questionnaire with eight dispositional measures. We measured SVO using the slider measure (Murphy et al., 2011), which involves a series of 15 scenarios, in which participants have to make dictator game decisions, distributing money between themselves and another participant. We furthermore used 8 items from the Brief Fear of Negative Evaluation Scale (Leary, 1983), which were selected based on intercorrelations and face validity. We used the decisiveness facet (6 items) of the Need for Closure scale (Kruglanski et al., 1993) as our measure for dispositional tradeoff aversion. To gauge people's dispositional tendency to be inattentive, we used the 16-item Need to Evaluate Scale (Jarvis & Petty, 1996). For an overview over all dispositional measures, see Table 2.

Table 2

All dispositional measures administered in the study (wave 1)

Measure	Citation	Scale	Number of items	Example item	Internal consistency (McDonald's omega)	Note
Wiggling-related						
Social Value Orientation	Murphy et al., 2011	n.a.	15	n.a.		
Brief HEXACO Inventory - Honesty-Humility	De Vries, 2013	1 (strongly disagree) to 5 (strongly agree)	4	"I would like to know how to make lots of money in a dishonest manner."	.642	
Guilt And Shame Proneness scale	Cohen et al., 2011	1 (Very unlikely) to 7 (Very likely)	8	"You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?"	.853	Guilt subscale only
Brief Fear of Negative Evaluation Scale	Leary, 1983	1 (Not at all characteristic of me) to 5 (Extremely characteristic of me)	8	"I am afraid that others will not approve of me."	.957	Original scale has 12 items, item selection based on face validity and correlations.
Tradeoff aversion						
Need for Closure	Kruglanski et al., 1993	1 (completely disagree) to 5 (completely agree)	6	"When I am confronted with a problem, I'm dying to reach a solution very quickly. "	.879	Decisiveness facet only
Desirability of Control	Burger & Cooper, 1979	1 (Doesn't apply to me at all) to 7 (Always applies to me)	5	"I enjoy having control over my own destiny."	.777	Original scale has 20 items, item selection based on face value and factor loadings.
Inattention						
Need to Evaluate Scale	Jarvis & Petty, 1996	1 (extremely uncharacteristic of me) to 5 (extremely characteristic of me)	16	"I form opinions about everything."	.887	
Brief HEXACO Inventory - Conscientiousness	De Vries, 2013	1 (strongly disagree) to 5 (strongly agree)	4	"I work very precisely."	.647	

Data analysis

The results section is split into two parts. In the first part, we present our main analyses, only using data from the selfish, efficient and antisocial ignorance conditions, as well as their respective baselines. In the second part of the results, we present the results of the prosocial ignorance condition separately.

For our main analyses (Hypotheses 1 to 3), we used repeated measurement logistic regressions, clustering the error term on the subject level. We then identified different allocation types as well as ignorance types by combining choice information from different contexts. Finally, we investigated in which way different dispositional measures predicted ignorance and allocation choices using repeated measurement logistic regressions (Hypothesis 4). We employed anovas and *t*-tests to check for differences in these dispositional measures between the different types.

The variable *option with the higher payoff under ignorance* described the selfish option in the selfish ignorance condition, the efficient option in the efficient ignorance condition, and the antisocial option in the antisocial ignorance condition. For all dispositional measures, we first calculated participants' average scores (after potential item reversing), and then z-standardized these scores. The data for this study were analyzed using STATA version 17.0.

Carry-over and time effects, missing values and selectivity in dropouts

As pre-registered, we checked for order effects in a total of 13 Chi^2-tests (one per behavioral outcome). We found no significant effects between our counterbalanced groups, all *p*s > critical *p* (Bonferroni corrected), and thus used the whole dataset. We also checked for time effects in the SO- and the OO-context decisions, as we counterbalanced the order in which participants were exposed to these contexts over the four data collection waves. Indeed, we found that the participants were less likely to ignore in later waves compared to earlier waves, OR = 0.74, 95% CI [0.65, 0.84]. We did not find time effects in participants' allocation decisions, OR = 1.05, 95% CI [0.98, 1.12].

We employed Little's test for randomness in missing values (Little, 1988), and found no suspicious pattern, Chi^2(10) = 7.501, *p* = .678. As noted above, we had a total dropout of about 20% from wave 1 to wave 4, meaning that 878 of the initial 1110 participants took part in all four data collection waves. We explored differences between participants who dropped out and participants who took part in all four waves. When comparing participants who took part in all four waves to participants who dropped out at some point, we found significant differences in all dispositional measures except for SVO, with dropout participants showing lower scores on Honesty-Humility, Guilt Proneness, Desirability of Control, and Conscientiousness, but higher scores in the Need to Evaluate, Need for Closure, and the Brief Fear of Negative Evaluation, all *p*s < . 024.

Results

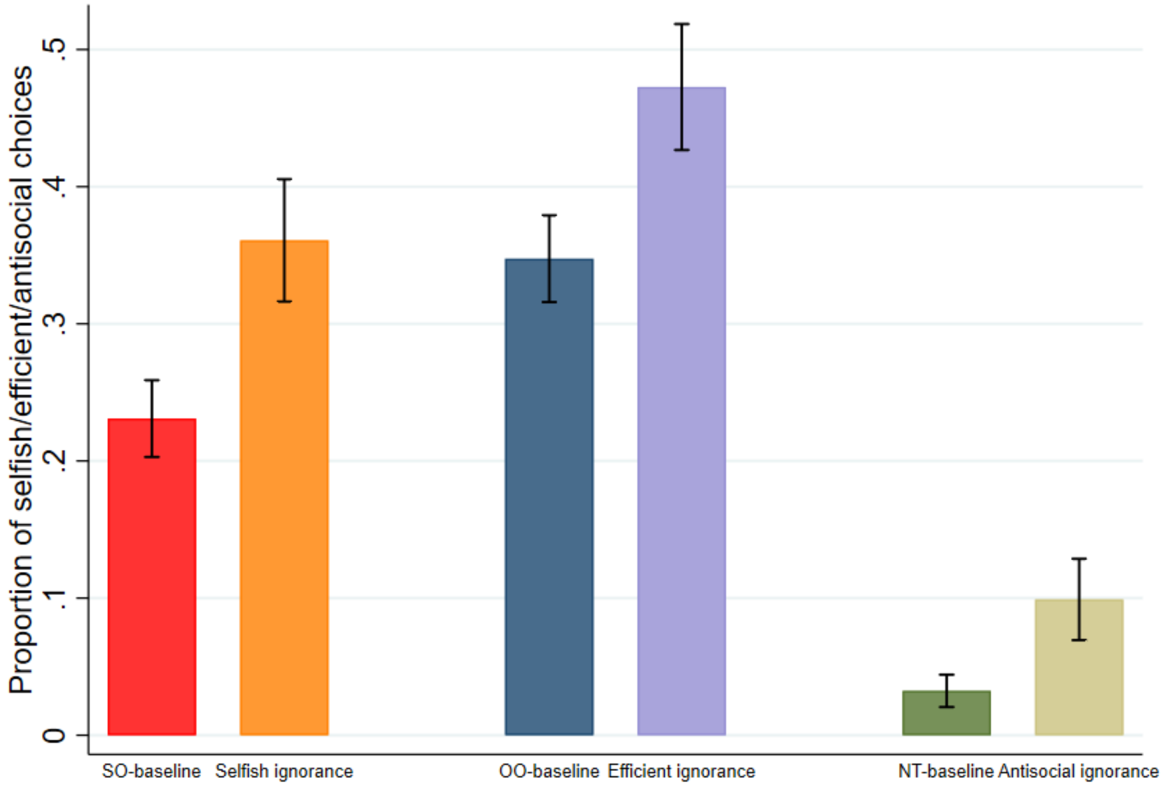
Selfish, efficient and antisocial ignorance

Allocation decisions

To investigate the effect of the option to ignore on allocation decisions across decision-making contexts, we utilized a repeated measurement logistic regression to predict choices by condition (H1). We predicted allocation choices (0 = option with lower payoff under ignorance, 1 = option with higher payoff under ignorance) with whether participants could ignore in the context (0 = baselines, 1 = ignorance conditions). We also included one categorical variable for the decision context (SO, OO, and NT), and the interaction of decision context (OO as reference category) and ignorance condition (yes/no) to be able to investigate potential differences in the impact of ignorance on choice behavior across decision contexts (H2). The results showed that participants generally were more likely to choose the option with the higher payoff (i.e., the selfish, efficient, or antisocial choice, respectively) when given the opportunity to ignore parts of the payoffs, compared to their respective baselines, OR = 2.00, 95% CI [1.52, 2.63], as well as in each context individually, all *p*s < .001(see Figure 6). The effect of ignorance condition on choice behavior was more pronounced in the OO-context (choice difference 12.52%) compared to the NT-context (choice difference 5.74%), OR = 1.98, 95% CI [1.09. 3.61], while choice changes due to the opportunity to ignore were of similar size in the SO-context compared to OO-context (12.09% vs. 12.52%), OR = 1.12, 95% CI [0.75, 1.66]. Running the same repeated measurement logistic regression with the SO-context as a reference category revealed also no significant interaction between the SO- and the NT-context, OR = 1.77, 95% CI [0.97, 3.26].

Figure 6

Allocation decisions in the six different conditions within the SO-context (left), the OO-context (center), and the NT-context (right)



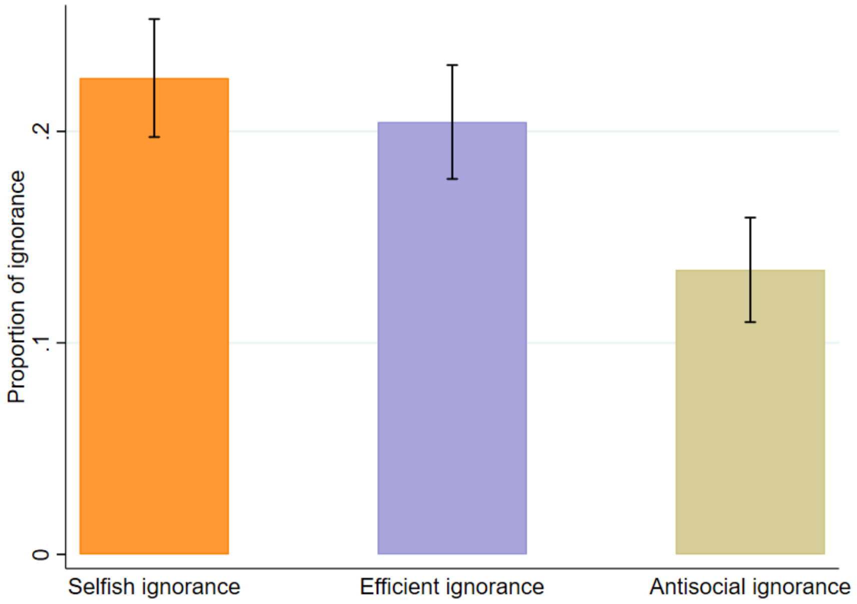
Taking advantage of the within subject design we identified different choice patterns across decision contexts, and combined them to specify several allocation types. 60.71% of our sample were consistently prosocial within the SO-baseline and the selfish ignorance condition, while 18.97% were consistently selfish. 16.96% of our sample behaved as *moral wigglers*. Specifically, they chose prosocially in the SO-baseline, and selfish in the selfish ignorance condition (3.35% showed the reverse pattern). Of these moral wigglers, 73.68% ignored in the selfish ignorance condition.

Decision to ignore

To better understand the decision to ignore relevant payoff information, we specifically analyzed conditions allowing for ignorance. Utilizing a repeated measurement logistic regression, we tested for context effects (SO, OO and NT, NT as reference category) on ignorance decisions (H3). The results showed that participants were least likely to ignore in the antisocial condition (NT, 13.45%) compared to the selfish condition (SO, 22.52%), OR = 3.55, 95% CI [2.40, 5.24, and the efficient condition (OO, 20.44%), OR = 2.78, 95% CI [1.89, 4.09], (see Figure 7). Using the OO-context as a reference category in a second repeated measurement logistic regression, there was no significant difference in the likelihood of ignoring between the selfish and efficient ignorance condition, OR = 1.28, 95% CI [0.92, 1.76], (see Figure 7).

Figure 7

Ignorance decision across different contexts



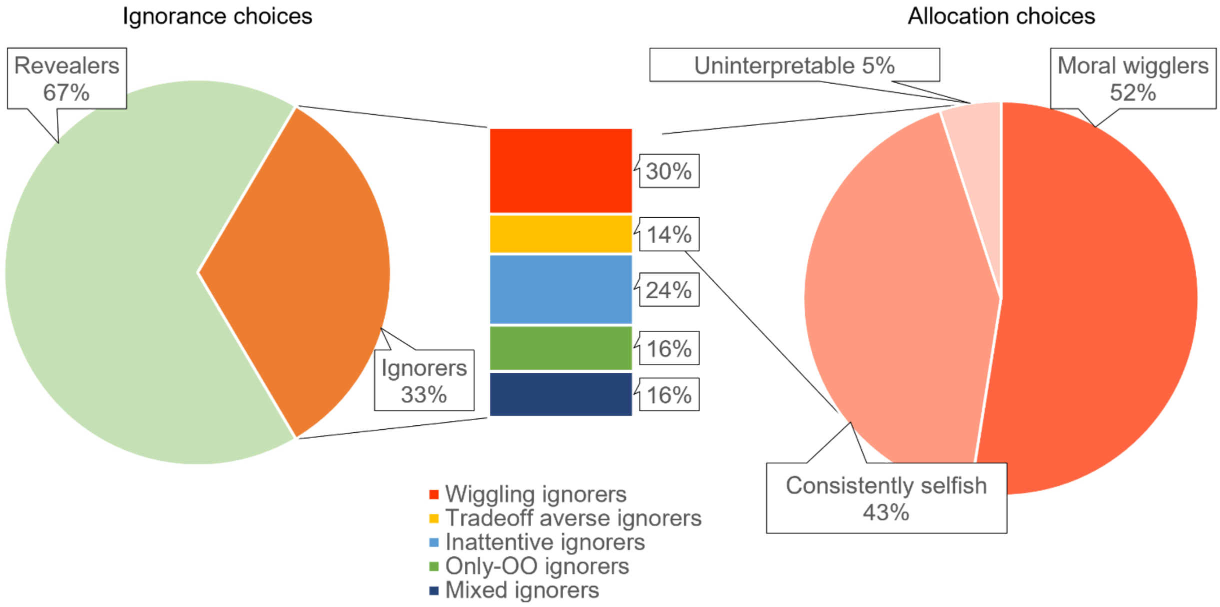
Analyzing participants' ignorance choice across the different contexts, ignorance choices were significantly correlated across all three conditions, all *r*s > .44, all *p*s < .001, meaning that participants who ignored in one context were also more likely to ignore in the other two contexts.

When investigating intraindividual patterns of ignorance choices, we identified several different ignorance types. 66.94% of participants revealed all payoffs in all three contexts (*consistent revealers*) while 33.06% of participants ignored at least once (*ignorers*, see Figure 8). Amongst these ignorers, 29.83% ignored only in the selfish ignorance condition (*wiggling ignorers*), 15.97% only in the efficient ignorance condition (*only-OO ignorers*); 13.87% in the selfish and the efficient ignorance condition (*tradeoff ignorers*), while 24.37% ignored in all three conditions (*consistent ignorers*). 15.97% of the participants who chose to ignore did not fit into any of the previous patterns (*unclassified*). When further investigating the choices of the *wiggling ignorers*, we see that 52% chose selfishly only in the ignorance condition, while 43% were consistently selfish (5% chose the low payoff option under ignorance, see Figure 8).

In a next step, we investigated how much ignorance in the selfish ignorance condition is attributable to each of these ignorance types. We found that 40.80% of ignorance in the selfish information condition comes from participants classified as wiggling ignorers, 18.87% from tradeoff ignorer and 33.33% from inattentives.

Figure 8

Proportion of different types



Dispositional measures

To gain a deeper understanding of the motivational basis for the observed choices, we combined choice data with measures of inter-individual differences in pre-dispositions that are conceptually linked to the motivations in question.

Decision to ignore

We tested Hypotheses 4a to 4c by employing a repeated measurement logistic regression with ignorance choices predicted by each of the three dispositional measures, the decision context, and the interaction term of dispositional measure and decision context. We find no support for any of the three hypotheses (see Appendix B).

Exploring in which way our dispositional measures are related to the decision to ignore, we extended our analyses by using a total of three separate logistic regressions (one for each ignorance condition) predicting ignorance choices by all eight dispositional measures (see Figure 9). The results of the first regression showed that selfish ignorance was significantly related to lower scores on SVO, OR = 0.78, 95% CI [0.66, 0.92], and Honesty-Humility, OR = 0.76, 95% CI [0.63, 0.92].

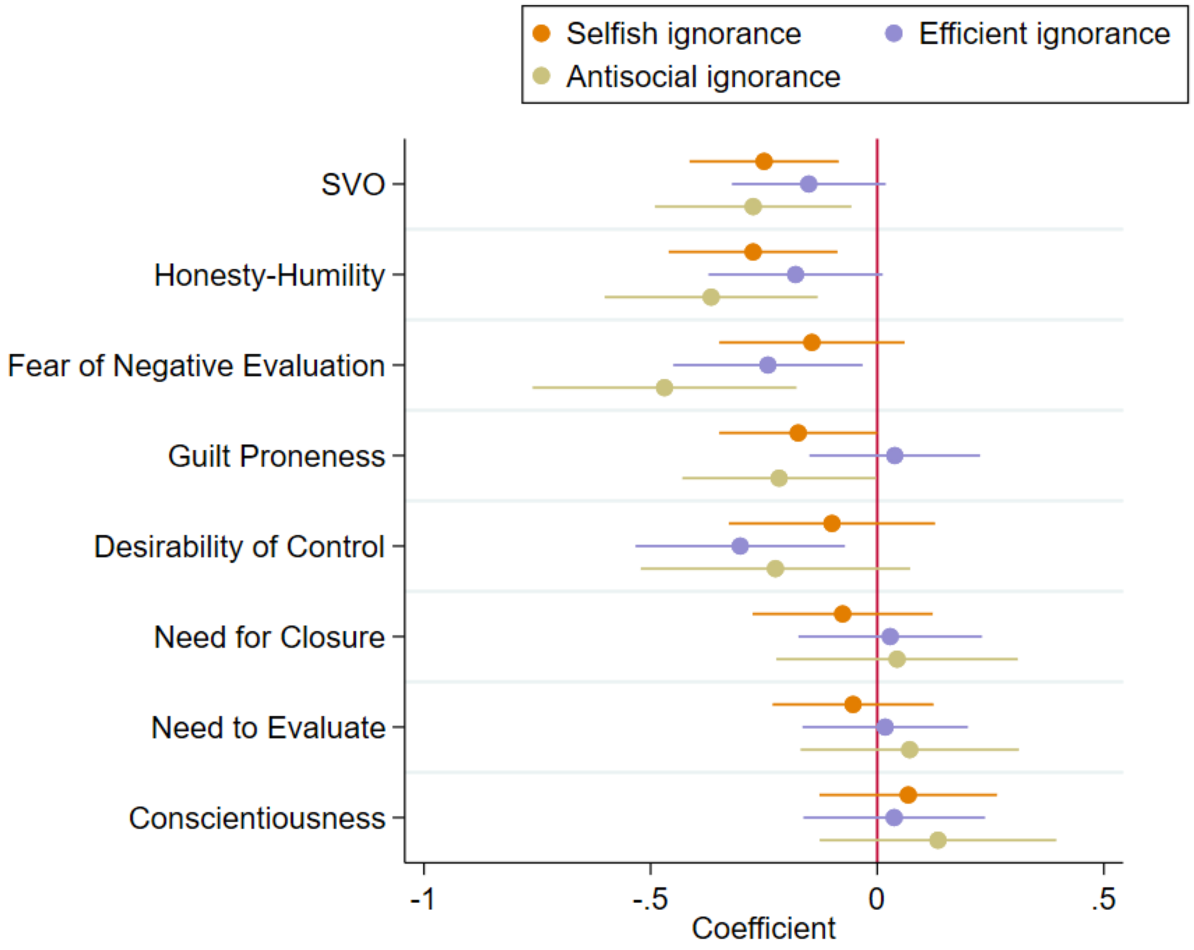
In the second model, efficient ignorance choices showed a negative relation to the Brief Fear of Negative Evaluation Score, OR = 0.79, 95% CI [0.64, 0.97], and the Desirability of Control, OR = 0.74, 95% CI [0.59, 0.93].

The third model showed that antisocial ignorance was related to lower scores on SVO, OR = 0.76, 95% CI [0.61, 0.94], Honesty-Humility, OR = 0.69, 95% CI [0.55, 0.88], Guilt Proneness, OR = 0.81, 95% CI [0.65, 1.00], and the Brief Fear of Negative Evaluation scale OR = 0.63, 95% CI [0.47, 0.84]. The measures of Need for Closure, Need to Evaluate and Conscientiousness did not significantly predict any of the ignorance choices (see Figure 9, for all statistics, see Appendix C).

Linking inter-individual differences to the identified ignorance types allows us to better understand the motive and value structure of each group. To this end, we ran several anovas combined with post-hoc t-tests investigating differences in our dispositional measures between ignorance types. Interestingly, both *consistent revealers* as well as *only-OO ignorers* were more prosocial as measured by SVO and Honesty-Humility compared to *wiggling ignorers*, all *p*s < .045 (for all related analyses, see Appendix C).

Figure 9

Standardized beta coefficients of logistic regressions predicting ignorance choices with dispositional measures



Allocation decisions

In order to explore the association of our eight dispositional measures to allocation decisions in the different contexts, we calculated six different logistic regressions, separate for each condition, with all eight dispositional measures predicting allocation choice (see Figure 10).

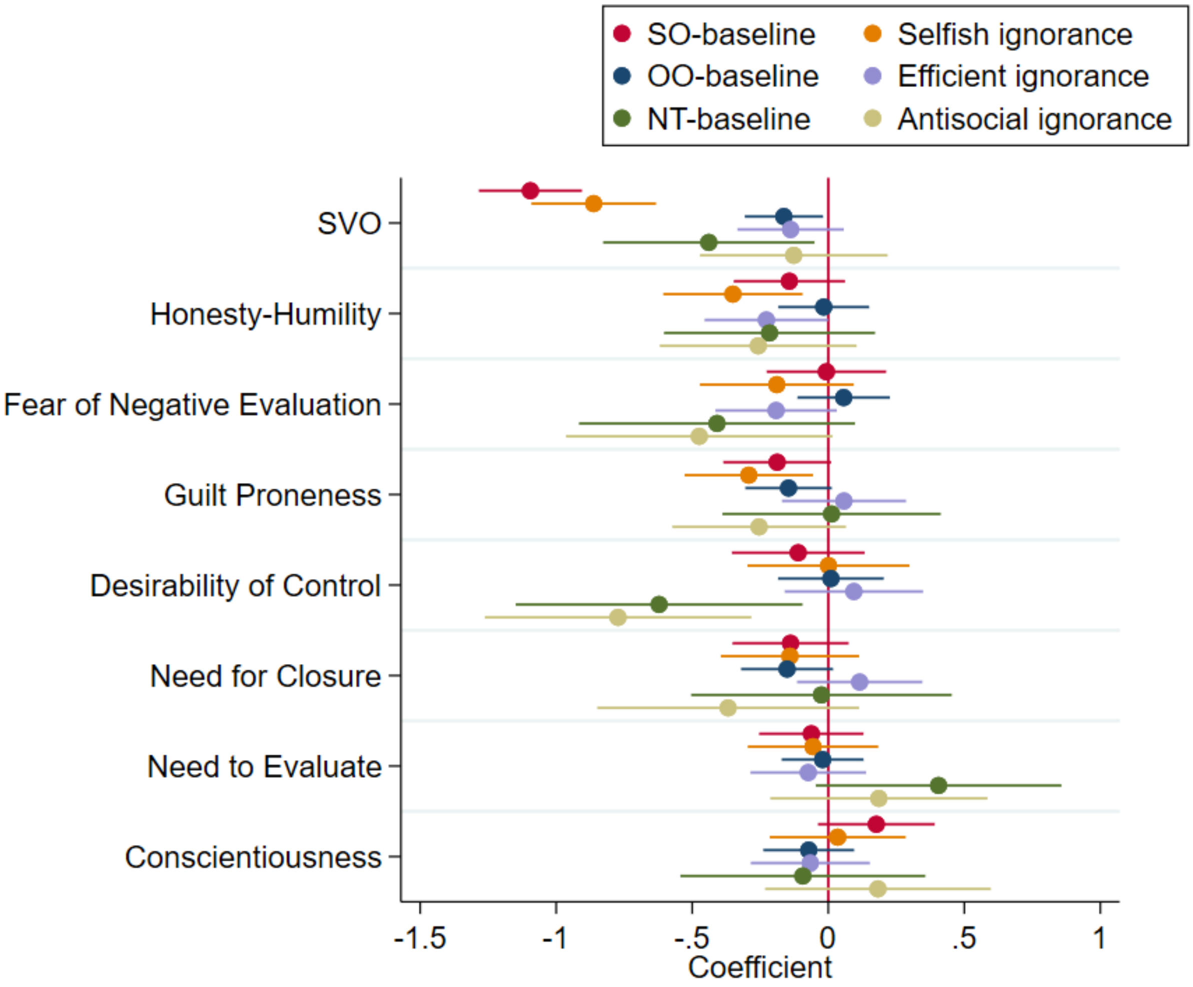
In the SO-baseline, we replicated the well-established link between SVO and prosocial decisions, OR = 0.33, 95% CI [0.28, 0.40]. In the selfish ignorance condition, SVO, OR = 0.42, 95% CI [0.34, 0.53], Honesty-Humility, OR = 0.70, 95% CI [0.55, 0.91], and Guilt Proneness, OR = 0.75, 95% CI [0.59, 0.95], all predicted selfish choices negatively.

In the OO-baseline, only SVO predicted allocation choice significantly, OR = 0.85, 95% CI [0.74, 0.98]. In the efficient ignorance condition, only Honesty-Humility significantly predicted efficient choice, OR = 0.80, 95% CI [0.63, 1.00].

In the NT-baseline condition, SVO, OR = 0.64, 95% CI [0.44, 0.95], and Desirability of Control, OR = 0.54, 95% CI [0.32, 0.91], negatively predicted antisocial choice. In the antisocial ignorance condition, only Desirability of Control significantly predicted choice, OR = 0.46, 95% CI [0.28, 0.75].[12]

Figure 10

Standardized beta coefficients of logistic regressions predicting allocation choices with dispositional measures



Using an ANOVA, there was a significant difference between the three groups of *consistent prosocials*, *consistent selfish* and *moral wigglers* in terms of their SVO scores, *F*(2, 429) =

[12] For all results on the post-decision questionnaire, see Appendix C.

71.05, *p* < .001. Participants classified as *moral wigglers* showed lower SVO scores compared to the consistent prosocials, *t*(345) = 4.04, *p* < .001, but higher SVO score than the consistently selfish participants, *t*(159) = -5.81, *p* < .001.[13] We also saw differences between these groups with regards to Honesty-Humility: Consistent prosocials showed higher scores than moral wigglers, *t*(345) = 4.89, *p* < .001, but there was no significant difference between consistent selfish types and moral wigglers, *t*(159) = 1.20, *p* = .231. The same pattern showed for Guilt Proneness, all *p*s < .001. We found no significant differences between these groups in the Brief Fear of Negative Evaluation scale, Desirability of Control, Need for Closure, Need to Evaluate, or Conscientiousness, all *p*s > .44.

Prosocial ignorance

In order to see in which way the option to ignore potentially can also increase prosociality, we investigated choice behavior of participants in the prosocial ignorance condition.

Running a repeated measurement logistic regression with ignorance condition (prosocial vs. selfish) predicting ignorance choices, there was no significant difference in the proportion of people who ignored in the prosocial (22.57%) and the selfish (22.52%) ignorance condition, OR = 1.00, 95% CI [0.78, 1.29].

Comparing allocation choices between prosocial ignorance condition and SO-baseline, we ran a repeated measurement logistic regression with condition (SO-baseline vs. prosocial ignorance) predicting allocation choice. Participants were not more likely to select the prosocial option in the prosocial ignorance condition (79.12%) compared to the SO-baseline condition (76.91%), OR = 0.97 95% CI [0.86, 1.10].

When investigating ignorance types in the SO-context, the majority of participants consistently revealed the information in both the selfish and the prosocial ignorance condition (63.50%). While 13.97% of all participants ignored only in the selfish ignorance condition, 14.08% ignored only in the prosocial ignorance condition; 8.45% ignored in both the selfish and the prosocial ignorance condition.

Looking at how choice behavior differs between the SO-baseline and the prosocial ignorance condition showed that 72.41% of all participants consistently chose the prosocial option, while 14.35% consistently chose the selfish option. 6.62% of participants chose the

[13] Note that these analyses resemble the results of Grossman and van der Weele (2017) on the relationship between SVO and willful ignorance. However, they used engaging in ignorance as a proxy for exploiting moral wiggle room, as they did not employ a within-subject design. We can also replicate their results directly, showing that participants who ignored in the selfish ignorance condition had higher SVO scores than participants who revealed, but behaved selfishly, *t*(167) = 3.02, *p* = .003, but a lower SVO-score compared to participants who revealed and chose the prosocial option, *t*(365) = -6.17, *p* < .001.

selfish option in the SO-baseline, and the prosocial option in the prosocial ignorance condition (i.e., *prosocial wigglers*), while 6.62% of participants showed the reverse pattern.

Of those participants who ignored the prosocial ignorance condition, the majority (74.07%) were consistently prosocials, and only a small minority (7.41%) chose the selfish option in the SO-baseline, but the prosocial option in the prosocial ignorance condition. The remaining 18.52% chose the option with the lower payoff under ignorance, which was not compatible with any of our hypothesized motivations.

Discussion

Our study investigated the underlying motivation of willful ignorance in prosocial decision-making. We used variations of the classic hidden information treatment (Dana et al., 2007), in a large-scale online study and manipulated the context and accessibility of information within-subject. As such, this study is the first of its kind to explore within-subject choice patterns across different contexts, allowing us to make precise inferences about what motivated behavior. The results showed that (1) the option to ignore impacted allocation decisions in the selfish, efficient, and antisocial ignorance condition. The option to ignore had the strongest influence on behavior in the selfish (12.09% choice shift) and the efficient ignorance condition (12.52% choice shift), but much less in the antisocial ignorance condition (5.74% choice shift). Exploring dispositional tendencies of ignoring participants, we showed that (2) selfish and antisocial ignorance was predicted by lower scores on SVO; while efficient ignorance was predicted by low scores on Desirability of Control and Fear of Negative Evaluation. Finally, (3) we estimate that 41% of ignorance was motivated by wiggling-related motivations, 19% by tradeoff aversion and 33% by inattention. Next, we turn to interpreting each finding separately, and what it means for our research question of who ignores, and why.

(1) Context effects

In our study, participants made decisions in three different contexts, designed to eliminate different motivations driving ignorance choices. While in the selfish ignorance condition, participants could be motivated by wiggling, tradeoff aversion and inattention, we removed the wiggling motivation in the efficient ignorance condition. In the antisocial ignorance condition, finally, only inattention should motivate ignorance. As such, we hypothesized a steady decrease of ignorance from the selfish, to the efficient, to the antisocial ignorance condition. The option to ignore influenced allocation decisions in all three contexts, but participants were less likely to ignore in the antisocial ignorance condition compared to the other two ignorance conditions. Equivalently, we see that the option to ignore had a weaker effect on allocation choices in the antisocial ignorance condition, but was similar in strength in the selfish and efficient ignorance conditions.

(2) Typologies

Our within-subject design allowed for more complex analyses by determining allocation and ignorance types. Our data showed that 30% of all ignorant participants (those who ignored at least once) ignored exclusively in the selfish ignorance condition, suggesting wiggling-related concerns (wiggling ignorers). 14% of ignorant participants ignored in both the selfish and efficient ignorance condition, and 24% of ignorant participants ignored in all three ignorance conditions, implying tradeoff aversion and inattention, respectively.

We further identified different allocation types by investigating allocation choice patterns within the SO-context. We identified 16.96% of our sample to exploit moral wiggle room, meaning that these people chose the prosocial option in the baseline condition with transparent payoffs, while choosing selfishly when they could ignore the consequences of their choice for the charity. This relates to prior work, in which the amount of wiggling has been inferred from subtracting the level of prosocial behavior in the baseline condition from the level of prosocial behavior in the selfish ignorance condition. In their meta-analysis, Vu et al. (2023) conclude to find 15.6% of moral wigglers, which is similar in size to our estimate.

(3) Dispositional measures

We also administered several dispositional measures to understand who ignored and their underlying motivation for doing so. We identified three motives and measured them with their respective dispositional measures. Social Value Orientation (SVO) was matched to the motivation of wiggling, predicting that more selfish participants should be more likely to engage in willful ignorance. Though we did not find the hypothesized interaction effect of SVO and condition (H4a), we did find that SVO significantly predicted selfish ignorance, as well as antisocial ignorance, but not efficient ignorance. For the antisocial ignorance condition, we can speculate that participants were not necessarily inattentive in this setting, but rather did not care about the decision as it did not involve any payoffs to themselves. We furthermore showed that it was the participants with medium SVO-scores who exploit moral wiggle room in the selfish ignorance condition, meaning that they chose the prosocial option in the transparent setting (i.e., the SO-baseline), but the selfish option when given the option to ignore the consequences of their choices for the charity (i.e., the selfish ignorance condition). By this, we conceptually replicated Grossman and van der Weele (2017), who used ignorance as a proxy for wiggling. This indicates that it is people who are not convinced prosocials, but also not convinced selfish participants who exploit moral wiggle room, but participants who are somewhere in between. As selfish ignorance is negatively predicted by SVO, not only participants who exploit moral wiggle room by behaving more selfishly, but also convinced selfish participants seem to engage in ignorance in the selfish ignorance condition. This is further supported by the fact that 43% of wiggling ignorers consistently chose the selfish option in the SO-context.

For Need for Closure as a measure for tradeoff aversion, we did not find significant relationships without behavioral measures (H4c). However, Desirability of Control as an alternative measure did show significant effects. First, participants high in Desirability of Control were less likely to ignore in the efficient ignorance condition. Also, participants high in Desirability of Control were less likely to choose the antisocial option in the NT-baseline, as well as in the antisocial ignorance condition. Together, this suggests that Desirability of Control plays a role in both the OO- as well as the NT-context, suggesting that it might be a better measure for tradeoff aversion in our setup.

The Need to Evaluate as a measure for inattention did not show significant relations to our behavioral measures (H4c), nor did Conscientiousness as an alternative measure. Antisocial ignorance is overall related to more selfish dispositions (as measured by SVO, Honesty-Humility, as well as Guilt Proneness). As such, ignorance in this setting might be less connected to inattention than to indifference, as more selfish people did not have any stakes in this setup because it did not involve payoffs to themselves.

For the Brief Fear of Negative Evaluation scale, we found negative relations to efficient ignorance, as well as antisocial ignorance, but no relation to selfish ignorance. This suggests that image concerns as measured by the Brief Fear of Negative Evaluation play no role in prosocial decision-making. Simultaneously, participants might fear to be negatively evaluated when ignoring in settings in which one distributes money between the two charities.

How much ignorance is wiggling-related?

We originally set out to understand how much (if any) of the ignorance observed in the classic hidden information treatment of Dana et al. (2007) was actually driven by a desire to behave selfishly while avoiding negative consequences from doing so. So far, we have discussed in which ways ignorance over all three of our contexts has been driven by different motivations. We will now turn to understanding how much ignorance in the selfish ignorance condition specifically (which is equivalent to the classic hidden information treatment) is driven by the respective motivations. There are different approaches to answer this question.

First, we can follow the approach of Exley and Kessler (2021) who based their estimates of how much ignorance was wiggling-related on comparing levels of ignorance between a context involving a self-other tradeoff, and a context involving an other-other tradeoff, removing wiggling-related motivations as drivers for potential ignorance. They concluded that a majority of ignorance (66% to 81%) remained when wiggling-related motives were removed. If we were to follow Exley and Kessler's approach, we would divide 2.08 (difference between selfish and efficient ignorance levels) by 22.52 (level of selfish ignorance), resulting in 9.23% of ignorance which could be attributed to wiggling-related motives in the selfish ignorance condition. Similarly, we could estimate the share of ignorance that could be attributed to tradeoff aversion by dividing 7.01 (difference between efficient and antisocial ignorance levels) by 22.52 (selfish ignorance level), resulting in 31.13% of ignorance in the selfish ignorance condition to be attributed to tradeoff aversion. The remaining 59.64% of ignorance could be attributed to inattention of related motivations. This approach, however, comes with a conceptual drawback. By only comparing level effects, one relies completely on the logic that one moving from one context to the other removes one form of motivation (in our case, wiggling-related motivation) without adding any other forms of motivation. This logic would assume that participants who ignored in the antisocial ignorance condition should also ignore in the other two ignorance conditions, and participants who ignored in the efficient ignorance condition should also ignore in the selfish ignorance condition. However, by only observing level effects, one cannot empirically confirm this basic prerequisite for the analysis.

Our data allowed us to test this assumption by utilizing our within-subject design to investigate patterns of behavior in different contexts. The results showed that a large majority of our sample falls into the pattern predicted by this logic, with 24% of all ignoring participants engaging in ignorance in all three contexts, 14% ignoring in the selfish and efficient ignorance condition, and 30% ignoring only in the selfish ignorance condition. However, we also identified a prevalent pattern in our data that did not fit into the predefined logic: 16% of our ignoring participants engaged in ignorance only in the efficient ignorance condition, but not in the other two. Interestingly, these participants showed similar patterns in terms of interindividual differences in dispositional prosociality to people who consistently revealed, meaning that both groups were dispositionally more prosocial. As such, some ignorance in the efficient ignorance condition might be driven by an indifference towards choosing the efficient or the fair option, as both these options can be regarded as appropriate. As such, though moving from the SO-context to the OO-context arguably removes wiggling-related motivations, it might also introduce other motivations. In our data, we found a considerable share of participants only ignored in the OO-context. As Exley and Kessler (2021) only compared levels of ignorance, one might have underestimated the share of ignorance that can be attributed to wiggling.

Our second approach to estimate the share of ignorance that can be attributed to wiggling addresses this drawback by combining behavioral outcomes in different contexts from the same participant, sorting them into types. We investigated how many of the participants who ignored in our selfish ignorance condition would fall into the category of wiggling ignorers, i.e., who ignored exclusively in the selfish ignorance condition. Our data revealed that 40.80% of ignorance within the selfish ignorance condition is attributable to these wiggling ignorers. As they revealed in both the efficient, and the antisocial ignorance condition, we would suspect that they were not motivated by tradeoff aversion or inattention when ignoring in the selfish ignorance condition. When investigating allocation choices of these wiggling ignorers in the SO-context, we saw that 52% of these participants indeed did *wiggle*, meaning that they chose the prosocial option in the SO-baseline, but the selfish option in the selfish ignorance condition. Another 43% of the wiggling ignorers chose the selfish option in both the SO-baseline and the selfish ignorance condition. There are two plausible explanations for these participants to ignore: Either they ignore as to avoid feeling bad for choosing the selfish option, or the information they could acquire is irrelevant for them, as they would choose the selfish option in any case.

18.97% of ignorance in the selfish ignorance condition stemmed from tradeoff ignorers, meaning that these participants ignored in both the selfish and the efficient ignorance condition, but not in the antisocial ignorance condition. It is thus plausible that these participants avoided the tradeoff decision they had to face if revealing in the selfish or efficient ignorance condition.

Another 33.33% of ignorance in the selfish ignorance condition can be attributed to consistent ignorers, meaning that these participants ignored in all three contexts. This result supports the idea that some ignorance is driven by inattention or disinterest, adding to research showing that a significant proportion of ignorance is due to the informational default in the paradigm (Grossman, 2014). In the classic setup of Dana et al. (2007), the agent has to actively choose to reveal, and stays ignorant if failing to do so. Changing this default decreases ignorance, showing that willful ignorance is highly susceptible to such default effects (Grossman, 2014). Inattention might be one driving factor for this susceptibility to default effects.

Prosocial ignorance

In our prosocial ignorance condition, we replicate prior findings showing that though people ignore in the prosocial ignorance condition, there is no significant impact on sharing behavior (Kandul & Ritov, 2017; Moradi, 2018). Indeed, the large majority of participants who engaged in prosocial ignorance (74.07%) chose the prosocial option both in the SO-baseline and the prosocial ignorance condition. Only 7.41% of these ignoring participants showed the hypothesized behavioral change, meaning that they chose the selfish option in the SO-baseline but the prosocial option in the prosocial ignorance condition. This suggests that ignorance in this condition is not used strategically by a large proportion of people.

The prosocial ignorance condition can also shed light on whether ignorance in the selfish ignorance condition is due to the self-other tradeoff without any wiggling motivation. Only 8.45% of all participants ignored in both the prosocial or selfish ignorance condition, 77.19% of which fall into the category of *consistent ignorers*. About 14% ignored only in the selfish and prosocial ignorance condition respectively. Thus, simply avoiding the self-other tradeoff does not seem to be the driving motivation behind selfish ignorance.

Limitations

One limitation of our study is that we observed lower levels of ignorance (22.52%) and slightly higher levels of baseline prosociality (76.91%) in our main study as compared to the original study of Dana et al. (2007) with ignorance levels of 44% and baseline prosociality levels of 74%, as well as compared to our replication in the same subject pool within our pre-study.[14] The main difference of our study to other designs is its within subject multi-wave data collection over ten days. As typical for multi-wave data collection, we had a total dropout of 20.90% from wave 1 to wave 4. Investigating differences in our dispositional measures between participants who dropped out and participants who continuously participated in all four waves showed selectivity in dropouts: participants who dropped out were less prosocial and less conscientious compared to their counterparts. This could be one explanation for the rather high rates of prosociality. We also observed time effects, meaning that people were less likely to ignore in later waves of the study. Both the selection and the time effects might explain why we observed low ignorance levels and high prosociality levels. If anything, this should make it harder to find the main effect of willful ignorance, meaning that in other samples, we would expect even more pronounced effects

[14] Using the same participant pool, we found strikingly similar rates of ignorance (56.5%), as well as baseline prosociality (69.4%) in our pre-study compared to the original results of Dana et al. (2007).

(see Vu et al., 2023). For the estimates of how much ignorance is indeed wiggling-related, this means that we can probably not translate the results for our sample to the general population at large. Future studies should make more accurate estimations using representative samples.

Implications

The insights of this study not only speak to the literature on willful ignorance in prosocial decision-making, but also have implications for other forms of ignorance. People have been shown to ignore in all kinds of situations, and for all kinds of reasons (for reviews see Golman & Loewenstein, 2018; Hertwig & Engel, 2016; Sweeny et al., 2010). In this paper, we have demonstrated the usefulness for comparing behavior in one context not only to a baseline, but also to behavior in other contexts, systematically varying the predicted motivation that should lead to behavioral change. Other domains for ignorance can surely profit from such an approach. For example, researchers have found a consistent tendency for people to avoid information about their HIV status due to a failure to return after getting tested (Hightow et al., 2003; Molitor et al., 1999; Tao et al., 1999). In order to better understand this failure to return, it might be useful to investigate in which way these people also fail to return to other kinds of test results, and in which way this failure to return is specific to HIV results.

From a policy perspective, our results indicate that simply providing more information about the social benefits of certain actions may not be sufficient. Crucially, the motives behind ignorance are diverse, and as such, interventions have to tackle multiple barriers to making informed choices at once. Our results indicate that one promising avenue might be to make people genuinely care about the information in question. Much willful ignorance is driven by selfishness, however, not necessarily in the way we thought: only some people ignore in order to behave selfishly without appearing so. A large group of people ignore because they do not care enough. Interventions aiming at making people care and empathize might be a fruitful avenue (Batson et al., 1981). Another approach would be to incentivize people to take up certain information. Our study showed that people do incorporate information they received, but sometimes they do not bother seeking out the information when it is hidden. If we want people to incorporate certain information into their decisions, policy-makers should consider incentivizing the uptake of information, as the benefits might outweigh the costs (D. Cain & Dana, 2012).

Conclusion

Willful ignorance in prosocial decision-making refers to the phenomenon that people avoid information about the consequences of one's behavior on others, leading to more selfish behavior. Our study systematically investigated the extent to which ignorance is motivated by wiggling-related concerns and tested two alternative motivations: tradeoff aversion and inattention. We relied on the classic hidden information treatment of Dana et al. (2007) as an operationalization to investigate willful ignorance. First, we replicated the effect that introducing the option to ignore decreases prosociality in a binary dictator game. Next, we showed that ignorance also changes behavior in other contexts, notably increasing efficient and antisocial choices as well when specific information is initially hidden. By investigating patterns of behavior across different contexts, we showed that 41% of ignorance is attributable to wiggling-related motivations, 19% to tradeoff aversion and 33% to inattention. This means that though some willful ignorance is motivated by wiggling, the majority of ignorance is not. Our results furthermore indicate that the option to ignore impacts behavior in all kinds of settings. As such, informational defaults are not only exploited as moral wiggle room, but change behavior more generally. When designing choices settings, such as consumer products or charity advertisements, one should keep in mind that people are highly sensitive to these informational defaults.

General discussion

Prosociality is essential for humans whenever they live and interact together, making it a crucial factor for civilizations at large. All societies depend on prosocial behavior of their members. These behaviors can range from minor acts like helping elderly people cross the stress or offering them a seat on the bus, to more large-scale issues such as redistributive social policies and taking in refugees in times of crisis. Unsurprisingly, then, researchers from all across the social sciences have taken an interest in studying the preconditions, constraints, and mechanisms of prosociality (e.g., Batson & Powell, 2003; Boyd & Richerson, 2009; Henrich et al., 2006; Ohtsuki & Iwasa, 2006; Pfattheicher et al., 2022; Rand & Nowak, 2013). While some researchers suggest that people feel a “warm glow” when engaging in prosocial behavior (Andreoni, 1990), others argue that people in general dislike unfairness (Fehr & Schmidt, 1999), and experience an aversive state of cognitive dissonance when behaving selfishly (Festinger, 1957; Konow, 2000).

Despite the prevalence of prosocial behavior, selfish and antisocial behavior is also common. Research on moral wiggle room has identified several situational factors that decrease prosocial behavior (Dana et al., 2007; Exley, 2016; Grossman & van der Weele, 2017; Matthey & Regner, 2011). For instance, selfish choices tend to increase when the agent's intentions are not clearly inferable from their actions. Moral wiggle room has been investigated using a variety of set-ups, such as giving participants the option to ignore the consequences of their behavior for others (Dana et al., 2007; Grossman & van der Weele, 2017), or the possibility of having their choices overwritten by a computer (Dana et al., 2007; Regner, 2021). Additional choice attributes (Chapter 2), or risk involved in the choice options (Exley, 2016) can also function as moral wiggle room. Although the concept of moral wiggle room has gained considerable attention in academia, its underlying mechanisms remain unclear. Why does selfish behavior increase under moral wiggle room? Do people feel less guilty when choosing selfishly, or do they expect less harsh judgment from others? Do they convince themselves that they chose selfishly for non-selfish reasons? Answering these questions can provide insights into the motivational underpinnings of prosocial behavior more generally. Against this backdrop, the current dissertation aimed to understand when, why, and by whom moral wiggle room is exploited, thereby investigating the motivational foundations of prosociality.

Chapter-by-chapter summary

In the first chapter of this dissertation, we laid out the theoretical foundation by verbally specifying the theory of moral wiggle room. We defined the concept of moral wiggle room as “situational characteristics that obfuscate the signal which the outcome of an own-payoff-maximizing (i.e., potentially selfish) behavior sends to others about one's intention to be selfish". With this definition, we deviated from the original definition of Dana et al. (2007) who talked about a reduction in the "commonly known one-to-one mapping between the [agent's] actions and the outcomes to both parties" (Dana et al., 2007, p. 69). This original definition emphasized the importance of not being able to infer an agent's actions or behavior. However, in some forms of moral wiggle room, the behavior is clearly observable (Dana et al., 2007; Exley, 2016), but it is the intentions behind behavior that are unclear. This new definition captures the concept more holistically. We also verbally formalized potential mechanisms through which moral wiggle room impacts behavior. We identified three psychological mechanisms in the literature that contribute to the effect: the anticipation of image damage, the perception of social norms, and anticipatory emotions. Additionally, we suggest that individual differences in other-regarding preferences and social image concerns may impact susceptibility to moral wiggle room. Ultimately, we hope to encourage more rigorous and efficient research on the effect of moral wiggle room and to demonstrate the benefits of formal verbal theory specification. By fully defining all concepts and operationalizations, we offer a more comprehensive and testable theory of moral wiggle room. In the remaining three chapters of this dissertation, I tested several elements of the theory, including the mechanisms and interindividual differences.

In Chapter 2, we introduced a novel form of moral wiggle room. Building on the classic social psychological research of Snyder et al. (1979), we investigated the effect of attributional ambiguity on charitable giving in binary dictator decisions. To manipulate attributional ambiguity, we varied the setup such that one group of participants was given a prosocial and selfish donation option to the same charity, while the other group had the two options (prosocial, selfish) attached to different charities. Participants were less likely to choose the prosocial option when different charities were used, regardless of which charity was associated with the more prosocial option. As such, we argue that attributional ambiguity represents yet another form of moral wiggle room, as choosing the selfish option does not necessarily signal selfish motivations, but could also reveal genuine preference for one specific charity on the individual level. Only by observing choices from a larger sample and by varying which charity is associated with which option can we reveal a hidden preference for the selfish option under attributional ambiguity. We indeed discovered no indication that individuals displayed a liking towards the charity linked with the self-profiting choice, which would have aligned with the hypothesis that individuals are concerned about their self-image. Instead, our findings indicated that self-profiting decisions were perceived as less selfish under attributional ambiguity. Consequently, in situations where attributional ambiguity existed, the shared expectations that not giving is socially unacceptable were diminished, leading to a reduction in the amount of donations via a decrease in the social pressure to give.

In Chapter 3, we focused on cognitive conflict as a mechanism behind willful ignorance, one prominent form of moral wiggle room. The study found that people who experience cognitive conflict were more likely to engage in willful ignorance. Moreover, results showed that the respective interactions between allocation choice on one hand and interindividual differences in Guilt Proneness, Social Value Orientation, and Honesty-Humility on the other, predicted cognitive conflict. The interactions show that dispositionally selfish individuals feel equally conflicted regardless of their choice, while prosocial individuals feel less conflicted when choosing the prosocial option. These findings speak to the possibility that willful ignorance can be viewed as a strategy for dispositionally selfish individuals to avoid conflict that arises from either choice.

Chapter 4 critically investigated the proclaimed mechanism behind willful ignorance (Dana et al., 2007; Grossman & van der Weele, 2017; Vu et al., 2023) by examining the extent to which wiggling-related motivations drive willful ignorance, and whether other motivations, such as tradeoff aversion and inattention, might play a role. To this end, we conducted a study using a within-subject manipulation of the context in which participants made their allocation decisions, varying the receiving parties, the payoff structure, and the possibility of engaging in ignorance. First, we replicated the finding that introducing the option to ignore decreases prosocial behavior in a binary dictator game (Dana et al., 2007; Vu et al., 2023). Next, we demonstrated that ignorance also influences behavior in other contexts, increasing both efficient and antisocial choices when specific information is initially hidden. Overall, by analyzing patterns of behavior across different contexts, we found that 41% of willful ignorance can be attributed to wiggling-related motivations, 19% to tradeoff aversion, and 33% to inattention.

What we learned about moral wiggle room

Overall, this dissertation sheds light on the mechanisms behind moral wiggle room. While we replicated the behavioral effect of several moral wiggle room paradigms, our work also stresses the importance of more closely investigating the proposed mechanisms, as well as ruling out alternative explanations. Moral wiggle room indeed does not specify the definite form it can assume, ranging from risk involved in the different options to the possibility to remain ignorant about the consequences of one's own behavior for others. There is an abundance of research showing that situational characteristics impact all kinds of behaviors through different channels. What sets moral wiggle room apart from other situational characteristics influencing behavior is its mechanism through wiggling-related motives. As a result, it is not enough to show the effect of moral wiggle room on prosocial behavior when one wants to claim that participants exploit moral wiggle room. Defaults, for example, impact behavior, making it more likely that people enroll in retirement schemes (Madrian & Shea, 2001) or become organ donors (Johnson & Goldstein, 2003). If particular defaults affect prosocial behavior, it might be that this is due to other reasons that have nothing to do with wiggling-related motivations, but rather with inattention, in which case we would not talk about moral wiggle room.

There are two ways to approach this conundrum. First, one can show that the behavioral effect is specific to contexts in which one could wiggle, i.e., that the specific paradigm does not influence other types of behavior besides selfish choices. Second, one can demonstrate that a particular motivation drives the behavior change, for example by showing that it relieves social pressure or that it is connected to interindividual differences that would suggest wiggling-related motives. In the last chapter of this dissertation, we used both these approaches to show that the effect of willful ignorance on prosocial behavior is at least partially driven by wiggling-related motives. We utilized the first approach by investigating whether and how the option to ignore influences behavior, not only in self-other tradeoffs (i.e., situations in which individuals might want to exploit willful ignorance for wiggling-related motives), but also in other contexts. Our results from comparing these different behavioral outcomes between three contexts provided inconclusive evidence: while participants were more likely to ignore in the self-other context compared to a decision context without any tradeoff to be made, we saw no difference between the self-other and the other-other tradeoff contexts. Furthermore, the option to ignore showed an effect on behavior in all three contexts. As such, this form of willful ignorance did not seem to be specific to self-other contexts. Our within-subject design allowed us to dig deeper into the data to identify patterns of behavior instead of comparing simple level effects. By determining different types of ignoring participants, we showed that a considerable share of participants exclusively ignored in the self-other tradeoff contexts. The allocation choices and interindividual difference measures of this type of participants supports the interpretation that they were motivated by wiggling. Our estimates suggest that about 41% of ignorance in the classic setup of Dana et al. (2007) was indeed driven by wiggling-related motives. The discrepancy between the conclusions we would draw from observing level effects and the conclusions from analyzing patterns of behavior stems from the fact that people seem to ignore in the other-other tradeoff context for different reasons than in the self-other tradeoff context. We identified another type of ignorers who exclusively ignored in the other-other tradeoff context.

Comparing this type to other participants suggests that they are more dispositionally prosocial, and less interested in having control over their choices. It is thus likely that they ignored because they are indifferent between the two choice options in the other-other tradeoff context. When only comparing level effects, one cannot identify whether ignorance was driven by similar motives in the different contexts.

This is to say that we as researchers have to be careful when identifying suitable baseline conditions to which we compare behavior under moral wiggle room. Situational characteristics change behavior for all kinds of reasons. To claim that behavior changes due to a desire to behave selfishly without looking selfish, one needs to show that the behavior change can actually be attributed to this mechanism. Not all willful ignorance we observe seems to be driven by wiggling-related concerns. The same might hold true for other forms of moral wiggle room.

Mechanisms

It is crucial to demonstrate that the behavioral effects observed when moral wiggle room is introduced are connected to wiggling-related motives in order to accurately refer to it as moral wiggle room. In our theoretical framework, we have identified three possible mechanisms: a normative mechanism, an emotional mechanism, and an image mechanism. In the following section, I will summarize and analyze the evidence for each of these mechanisms, connect the findings of this dissertation with previous research, and discuss their implications.

Normative mechanism

In our theory specification, we proposed that moral wiggle room takes effect by changing social norms, making selfish behavior less socially inappropriate. With this, we build on the original proposition of Dana et al. (2007) who suggest that a change in norms and constraints would make agents feel less compelled to give. In Chapter 2, we test this claim empirically in the domain of attributional ambiguity as moral wiggle room. In two studies, we measured social norms using the incentivized elicitation method proposed by Krupka and Weber (2013). We did so first after participants made the allocation decision in a binary dictator game themselves, and then again in a naive sample of participants who did not make the decision in question before. In both studies, we manipulated the context (with vs. without attributional ambiguity) between subjects. We found that choosing the selfish option was seen as less socially inappropriate when under attributional ambiguity in both samples. In the former sample, we saw that this change in norms mediated the relationship between experimental condition and allocation choices. We dug deeper, trying to understand how exactly this change in norms came about. We believed there were two obvious candidates: Either participants thought that it would be simply more appropriate to behave selfishly under attributional ambiguity, or choosing the selfish option was perceived as less selfish under moral wiggle room to begin with. We found empirical support for the latter: When asked how selfish each choice was, participants found choosing the selfish option to be less selfish under attributional ambiguity. These results are in line with the theoretical account of attributional ambiguity derived from the correspondent inference theory (Jones & Davis, 1965). In their theory, the authors postulate that the strength of an inference one can make from observing someone's behavior depends on the number of dimensions this choice option differs on compared to the other available choice options (i.e., noncommon effects). In our baseline treatment, choice options only differed on one dimension, namely how much to share with the charity. In our manipulation, we introduced a second noncommon effect by varying the receiving charity. Because this second noncommon effect created attributional ambiguity as to which dimension drove the agent's choice (i.e., the specific payoff distribution vs. the specific charity), behavior could not be clearly attributed to selfish intentions. As a result, choosing the selfish option was perceived as less selfish, since the agent could have acted according to a charity preference, and not in order to maximize their own payoffs.

With this experiment, we add to existing research showing how moral wiggle room reduces social disapproval of selfish behavior. So far, this relationship has mostly been shown in the domain of willful ignorance. For example, Krupka and Weber (2013) report that choosing the self-profiting option after deciding to ignore relevant information about others' payoffs was perceived as less socially inappropriate by observers compared to knowingly choosing the selfish option. Similarly, sharing less than the equal split leads to less ultimatum game rejections (Conrads & Irlenbusch, 2013), and lower third-party punishment (Bartling et al., 2014) when payoffs could be partly ignored. While this could reflect a more lenient social norm perception, our results suggest that choosing the self-profiting option is itself seen as less selfish. In that sense, selfish behavior does not become more permissible, but rather maximizing one's own outcomes is seen as less selfish under moral wiggle room. Future research should investigate whether and how this also holds for other forms of moral wiggle room, such as willful ignorance.

Emotional mechanism

A second mechanism we put forward in our theory specification is the anticipation of negative emotional reactions following selfish behavior. Specifically, we proposed that agents might anticipate feeling less guilty after behaving selfishly under moral wiggle room as opposed to transparent settings. Chapter 2 speaks to this idea, showing that participants who did not actually make an incentivized decision, but rather were asked to imagine choosing a selfish option, anticipated feeling less guilty under moral wiggle room. We also investigated post-decision emotional reactions in Chapter 2, but did not find the predicted differences between conditions with and without moral wiggle room. Similarly, in Chapter 4, we used self-reported emotional measures assessed after each baseline choice to predict ignorance choices in these contexts, but found no significant relationship. This discrepancy can be explained by the fact that people's feelings after they made a decision depend on the decision itself so the behavior represents self-selection into one or the other group. As such, a person that would feel very guilty if choosing a particular option will simply not choose that option. Asking all participants to anticipate how they would feel if choosing the selfish option gets rid of this problem and reveals that attributional ambiguity, and potentially moral wiggle room more generally, indeed can function as a buffer for emotional response in case of selfish behavior. These findings confirm the hypotheses brought forward by several authors that moral wiggle room decreases anticipatory guilt in case of selfish behavior, making selfishness more prevalent (Feiler, 2014; Garcia et al., 2020; Thunström et al., 2014).

Chapter 3 also speaks to an emotional mechanism of moral wiggle room. In this chapter, I related the concept of moral wiggle room, or willful ignorance to be specific, to the Cognitive Dissonance Theory of Leon Festinger (1957). Prosocial decision-making can be conceptualized as an intrapersonal conflict between what an agent wants (selfishness), and what they consider appropriate (prosociality). The theoretical model of Konow (2000) assumes that people experience cognitive dissonance when they share less than their regard as fair. Cognitive dissonance is generally conceptualized as a negative emotional state that people want to avoid. Applied to simple sharing decisions, this means that people avoid this negative emotional state of cognitive dissonance either by aligning their moral standards with their actual behavior by behaving prosocially, or they engage in self-deception strategies (Konow, 2000). Utilizing mouse-tracking to directly and unobtrusively measure this intrapersonal conflict, the study reported in Chapter 3 suggests that willful ignorance potentially functions as a self-deception strategy. Participants who experienced more cognitive conflict when making self-other tradeoff decisions were more likely to ignore in an independent decision context. These findings add to research showing that ignorance is related to a “negative drive state” (Festinger, 1957), operationalized by longer decision times, as well as higher self-reported choice difficulty in transparent binary dictator decisions (Matthey & Regner, 2011). Though the correlation between conflict and ignorance observed in our study is a first indicator, we cannot infer causality due to our study design. Furthermore, we assume that participants anticipate cognitive conflict before actually experiencing it. Future research would do well to investigate both the causal nature of the relationship between willful ignorance and cognitive conflict, as well as the ways in which the experience of cognitive conflict is anticipated by participants.

Image mechanism

In our theory specification of moral wiggle room (Chapter 1), we proposed that the main psychological mechanism driving the behavioral moral wiggle room effect was the agent’s reduced anticipation of image damage. We furthermore specified that the image mechanism is interrelated with the other two mechanisms of social norms and anticipated emotions. Adhering to social norms can help us achieve a positive social image, while not adhering to norms can damage our social image. As such, a change in social norms can also mean that our social image is less damaged in case of selfish behavior. Our results in Chapter 2 on the change in social norms under moral wiggle room thus suggest that participants also anticipated less social image damage in case they chose the self-profiting option. This interpretation is further supported by the finding that choosing the self-profiting option was perceived to be less selfish by the decision-makers and an independent sample.

Chapter 2 also adds to the discussion on self- vs. social image as drivers for moral wiggle room. Within the literature, several authors have suggested that self-image concerns are a main driver for the effect of moral wiggle room (Grossman & van der Weele, 2017; Matthey & Regner, 2014). In our theory specification, we decided to include social image, but not self-image concerns as a mechanism behind the effect. We did so in order to pacify the original conception of moral wiggle room with the hypothesized psychological mechanism we proposed. Recall that moral wiggle room was originally conceptualized as a reduction in the “*commonly* known one-to-one-mapping” (Dana et al., 2007, p. 69). Results from Chapter 2 indeed support the idea that it is mainly social, and not self-image driving the effect of moral wiggle room.Specifically, we found that people did not try to fool themselves into thinking that they chose the selfish option for non-selfish motives. Rather, they expected others to simply find selfish behavior less socially impermissible, speaking for the relevance of social image in this setting. We therefore add to the literature stressing the importance of social image for the effect of moral wiggle room (Andreoni & Bernheim, 2009; Grossman, 2015). It is important to note that during our experiments, no one else was present to witness the sharing decisions made. Nevertheless, the participants' behavior seemed to adapt to evolving norms, supporting the idea of internalized norm-following (Bicchieri, 2005; Conte et al., 2010). Future research should explore whether the effect of attributional ambiguity on behavior is even more pronounced when sharing decisions are made in the presence of others.

Interindividual heterogeneity in moral wiggle room

As early as their first paper on the subject, Dana et al. (2007) acknowledged that not everyone exploits moral wiggle room. While there are some people who always behave prosocially, including in situations with moral wiggle room, others consistently choose the selfish option, even in situations with full transparency. So, do people who exploit moral wiggle room systematically differ from those who do not? In our theory specification, we proposed that the effect of moral wiggle room depends on relatively stable interindividual differences in other-regarding preferences and image concerns. Specifically, we propose that individuals who exploit moral wiggle room should have moderate dispositional other-regarding preferences. For individuals who already have a strong inclination to act prosocially or selfishly, moral wiggle room should be less important for their decision, compared to those with more moderate other-regarding preferences. Furthermore, the moral wiggle room effect should increase with agents' dispositional image concerns.

Chapter 4 of this dissertation speaks to these propositions. With regards to how other-regarding preferences specify the impact of moral wiggle room on behavior, we confirm the proposition from our theory specification. Moral wigglers scored higher on Social Value Orientation than consistently selfish participants (i.e., are more other-regarding), but lower than consistent prosocials (i.e., are less other-regarding). With this we add to existing evidence, in which ignorance choices have been used as a proxy for exploiting moral wiggle room, showing a similar pattern (Grossman & van der Weele, 2017).

We also measured participants' fear of negative evaluation, which we argue captures how much people care about what others think of them (i.e., their social image concerns; Leary et al., 2015). Interestingly, we found no differences between moral wigglers and other participants on this measure. Similarly, in Chapter 3, we found no relationship between participants' fear of negative evaluation and allocation or ignorance choices. Pulling these findings together, and although we find indirect evidence for a social image-related mechanism by showing a change in social norms, we did not find any direct support for the role of image-related concerns. This could be due to the fact that the scale we used is unsuitable to capture social image concerns in this context. Leary and colleagues (2015) suggest nine scales with different foci to capture social image concerns. Future research should systematically study whether and how these measures may be better suited to capture social image in moral wiggle room situations.

What we learned about prosociality

As laid out in the introduction, the phenomenon of moral wiggle room is not only interesting for understanding the situational conditionalities of prosociality, but also gives us a better understanding of what drives prosociality at large. I introduced several factors that could potentially affect prosociality, and I will discuss the ways in which this dissertation adds to these respective literatures.

Social preferences

Social preference theories advance that agents care not only about their pure material outcomes, but also about the consequences of their actions on others (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Fischbacher, 2002; Fehr & Schmidt, 1999), and that this in turn results in prosocial behavior. This approach is therefore outcome-oriented, suggesting that people have preferences for certain outcomes of social interactions, including a desire to avoid unequal outcomes (Fehr & Schmidt, 1999). By contrast, the evidence presented in this dissertation adds to the literature arguing that it is not only preferences over certain distributional outcomes that drives prosocial behavior. Research on moral wiggle room has shown that people behave more selfishly when given an excuse to do so even in anonymous one-shot interactions, which speaks against social preference accounts. If people were motivated by distributional outcomes only, moral wiggle room should not impact behavior. This does not mean that people do not differ on how much they value their own payoffs vs. payoffs to others. Interindividual differences in Social Value Orientation, which can be seen as one measure for social preferences, is a valid and strong predictor for prosocial behavior, in situations with and without moral wiggle room (see Chapters 3 and 4). Bringing these ideas together, we can conclude that social preference, though limited in its capacity to explain all variance in prosociality, is one piece of the puzzle for understanding prosociality.

Social norms

One important factor influencing prosociality are social norms (Andreoni & Bernheim, 2009; Bénabou & Tirole, 2006; De Groot & Steg, 2009; Krupka & Weber, 2009, 2013). Indeed both in the presence and absence of others, it seems that expectations as to what others do or find appropriate influences people's prosocial choices. As such, norms have been identified to be decisive for prosocial behavior (Bicchieri & Chavez, 2010), and changes in norms match changes in behavior (Krupka & Weber, 2013). We add to this literature by showing that social norm perceptions mediate the effect of moral wiggle room on behavior. Crucially, we show that it is not a more lenient norm that leads to this change, but rather the perception of the same behavior as less selfish when enacted under moral wiggle room. We therefore demonstrate that not only the norm can be influenced by situational characteristics, as suggested by Krupka and Weber (2013), but also whether a certain behavioral expression is perceived as selfish.

Cognitive dissonance

In Chapter 3, we supported the idea that prosocial decision contexts elicit cognitive conflict as measured by mouse-tracking. Participants showed a greater curvature of their mouse trajectories in trials with a self-other tradeoff as opposed to trials in which there was one mutually beneficial option. As such, prosocial decision contexts can be conceptualized as an intrapersonal conflict situation. Our study furthermore shows that it is the selfishly choosing participants who experience more cognitive conflict in the process. Interestingly, we found an interaction of Social Value Orientation and allocation choice when predicting cognitive conflict: The dispositionally selfish participants felt conflicted independent of what they chose in each trial, while the dispositionally prosocials felt less conflicted when choosing prosocially. This finding adds to our understanding of cognitive dissonance in prosocial decision-making. Festinger (1957) proposed that individuals experience discomfort or dissonance when their beliefs, attitudes, or behaviors are inconsistent, leading them to change their attitudes or behaviors to reduce the dissonance. Applied to prosocial behavior, this means that people have conflicting desires for both self-interest and fairness, thus leading to the experience of cognitive dissonance. Konow (2000) proposes that people reduce this tension by either behaving prosocially and thus aligning their moral standards with their behavior or by engaging in self-deception. We add to this theory by specifying systematic interindividual differences that shape or influence how people deal with cognitive dissonance in prosociality. Our results indicate that some people (i.e., the dispositionally prosocials) can reduce cognitive dissonance by behaving prosocially, while others (i.e., dispositionally selfish individuals) cannot reduce tension this way. Rather, they have to engage in self-deception in order to reduce cognitive dissonance, as they still feel conflicted if behaving prosocially. Potentially, this means that while dispositionally selfish participants feel torn between their desire to maximize their own payoffs and their moral standards, dispositionally prosocial people seem to be more at peace with their decision, suggesting that they simply might have weaker self-profiting motives, but similar moral standards.

Person perception/image

As discussed in the general introduction, prosociality is thought to be driven by a concern for being seen as a moral and good person, either by others or oneself (Andreoni & Bernheim, 2009; Ariely et al., 2009; Bagwell & Bernheim, 1996; Glazer & Konrad, 1996). As noted above, we showed that social norms changed with the introduction of moral wiggle room, making selfish behavior more prevalent (Chapter 2). When choosing the self-profiting option is seen as less socially inappropriate, this means that selfish behavior is less damaging for one's social image. As social norms affect choice behavior even though participants' choices are not observable by others, agents seem to have internalized the prevailing social norm, and acted accordingly. In the same study setup, we find no evidence for self-image effects. As such, social image in the form of internalized social norms seem to be an important driver for prosociality, even in anonymous, non-strategic, one-shot interactions.

In summary, the results of this dissertation support the idea that some prosociality indeed is reluctant, meaning that people feel pressured into behaving prosocially. This dissertation also adds to the literature on the importance of social norms for prosocial behavior. Importantly, the effect of social norms on behavior seems highly internalized, which means that social norms impact behavior also in anonymous settings. Furthermore, we learned that cognitive conflict plays an important role in prosocial behavior: While dispositionally prosocial people can avoid the aversive emotional experience of conflict by behaving prosocially, dispositionally selfish people cannot, as they feel similarly conflicted independent of their choice. This suggests that some proportion of the population indeed values prosociality.

Open questions and future directions

In addition to the questions raised by each chapter of this dissertation, there are several open questions on how moral wiggle room impacts behavior, that go beyond the scope of this dissertation. First, prosociality has been shown to be contingent on prior prosocial behavior. The theory of moral balancing (Nisan & Horenczyk, 1990), for instance, suggests that individuals strive to maintain a level of moral status that they find acceptable. The theory outlines how individuals cope with situations where their actions deviate from their moral self-image. When their moral self-image falls below a certain standard, individuals engage in moral cleansing to make up for it. On the other hand, when their moral self-image is above an ideal level, individuals may be more likely to engage in immoral behavior as a form of moral license. As such, prior (im)moral behavior has been shown to impact the likelihood of future (im)moral behavior (Blanken et al., 2015). In the domain of moral reasoning, research has shown that people are aware of and condone their own motivated reasoning (Cusimano & Lombrozo, 2023). Similarly, people might be aware that they exploit moral wiggle room when they do. As a result, exploiting moral wiggle room might also be subject to time-contingent effects such as moral balancing.

Second, the effect of moral wiggle room might be culturally specific. There are several strands of research pointing into this direction. Prosociality generally seems to be highly culturally specific. Based on anthropological research and recent findings from intercultural psychology, there appear to be significant differences in how prosocial behavior is understood and integrated into social interactions across cultures (Köster et al., 2015). Also in simple dictator games, differences between cultures can be observed: A meta-analysis revealed that non-Western participants tend to be more generous than their Western counterparts (Engel, 2011). Another strand of research suggesting that moral wiggle room might have a culturally-specific component is research on agency and responsibility (Miller et al., 1990; Miller & Bersoff, 1992; Savani et al., 2010). As these two concepts are central to the effect of moral wiggle room, investigating cultural specificity of the effect of moral wiggle room seems a promising way forward.

Implications

We live today in a highly connected world. Our economies are globally intertwined, and we collaborate with people from all over the world. We also face equally interconnected challenges, ranging from rising inequality (Chancel et al., 2022), to climate change (Rama et al., 2022), as well as migration (International Organization for Migration, 2022). Addressing these global challenges requires large-scale cooperation, and in parts the willingness to share and empathize with others.

This large-scale cooperation already starts in small and mundane choice behavior. Many of our daily decisions have a (direct or indirect) impact on others, from how we spend our free time, and what we eat and consume, to how we travel. In our daily lives, we often experience that people forgo their own profits or give up resources for others or for the greater good, such as when giving to charity or opting for more sustainable consumption. In this dissertation, I argued that at least some of this prosociality is reluctant, meaning that people would prefer the selfish outcome, but feel pressured into being prosocial. Especially, situations in which the intentions behind choices cannot be clearly inferred promote selfish behavior. This means that by carefully designing situations in which intentions are clear to everyone, one might be able to foster prosocial interactions. Research has shown that besides the obvious societal benefits of prosociality, being nice to others can also enhance well-being, both for those who engage in it and those who receive it (Aknin et al., 2013). By promoting prosocial behavior, society can enhance the well-being of individuals and communities, as well as improve relationships, increase trust and reduce conflict.

Overall, better understanding the drivers of prosociality can have numerous positive implications for society, from promoting positive social interactions to addressing social issues. We have numerous global challenges ahead of us in which cooperation will be key. In this dissertation, I hope to contribute one small piece to the puzzle of understanding prosociality.

References

Adena, M., & Huck, S. (2020). Online fundraising, self-image, and the long-term impact of ask avoidance. *Management Science*, *66*(2), 722–743. https://doi.org/10.1287/mnsc.2018.3232

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Aknin, L. B., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Burns, J., Biswas-Diener, R., Kemeza, I., Nyende, P., Ashton-James, C. E., & Norton, M. I. (2013). Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology*, *104*, 635–652. https://doi.org/10.1037/a0031578

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477. https://doi.org/10.2307/2234133

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636. https://doi.org/10.3982/ECTA7384

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653. https://doi.org/10.1086/691703

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037//0022-3514.83.6.1423

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555. https://doi.org/10.1257/aer.99.1.544

Aronson, E. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry*, *3*(4), 303–311.

Asendorpf, J. B., Conner, M., de Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2016). *Recommendations for increasing replicability in psychology*. American Psychological Association. https://doi.org/10.1037/14805-038

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166.

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340–345.

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Bacon, F. (1857). *Meditationes sacrae*. Excusum impensis Humfredi Hooper.

Bagwell, L. S., & Bernheim, B. D. (1996). Veblen effects in a theory of conspicuous consumption. *American Economic Review*, *86*(3), 349–373.

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations*, *12*(4), 533–547.

Bandura, A. (1999). Social cognitive theory of personality. *Handbook of Personality*, *2*, 154–196.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*(2), 364.

Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C., & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of Personality and Social Psychology*, *80*(1), 125.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the “tragedy of the commons.” *Evolution and Human Behavior*, *25*(4), 209–220. https://doi.org/10.1016/j.evolhumbehav.2004.04.002

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Bartling, B., Engl, F., & Weber, R. A. (2014). Does willful ignorance deflect punishment? An experimental study. *European Economic Review*, *70*, 512–524. https://doi.org/10.1016/j.euroecorev.2014.06.016

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Batson, C. D., Ahmad, N., Lishner, D. A., Tsang, J., Snyder, C. R., & Lopez, S. J. (2002). Empathy and altruism. *The Oxford Handbook of Hypo-Egoic Phenomena*, 161–174.

Batson, C. D., Bolen, M. H., Cross, J. A., & Neuringer-Benefiel, H. E. (1986). Where is the altruism in the altruistic personality? *Journal of Personality and Social Psychology*, *50*(1), 212.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*(2), 290.

Batson, C. D., & Powell, A. A. (2003). Altruism and prosocial behavior. *Handbook of Psychology*, 463–484.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*(2), 107–122.

Baumeister, R. F. (1998). The self. In *The handbook of social psychology, Vols. 1-2, 4th ed* (pp. 680–740). McGraw-Hill.

Baumeister, R. F., & Alghamdi, N. G. (2015). Role of self-control failure in immoral and unethical actions. *Current Opinion in Psychology*, *6*, 66–69.

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*, *30*(3), 178–192. https://doi.org/10.1027/1015-5759/a000183

Bazerman, M. H., Tenbrunsel, A. E., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review*. https://doi.org/10.5465/amr.1998.533224

Beattie, J., Baron, J., Hershey, J. C., & Spranca, M. D. (1994). Psychological determinants of decision attitude. *Journal of Behavioral Decision Making*, *7*(2), 129–144. https://doi.org/10.1002/bdm.3960070206

Beauvois, J. L., & Joule, R. V. (1996). A radical theory of dissonance. *European Monographs in Social Psychology. Taylor and Francis, New York, NY*.

Bell, E., Norwood, F. B., & Lusk, J. L. (2017). Are consumers willfully ignorant about animal welfare? *Animal Welfare*, *26*(4), 399–402. https://doi.org/10.7120/09627286.26.4.399

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 6, pp. 1–62). Academic Press. https://doi.org/10.1016/S0065-2601(08)60024-6

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, *126*(2), 805–855. https://doi.org/10.1093/qje/qjr002

Berman, J. Z., & Small, D. A. (2012). Self-interest without selfishness: The hedonic benefit of imposed self-interest. *Psychological Science*, *23*(10), 1193–1199.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*(2), 161–178.

https://doi.org/10.1002/bdm.648

Bicchieri, C., Dimant, E., Gächter, S., & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72. https://doi.org/10.1016/j.geb.2021.11.012

Bicchieri, C., Dimant, E., & Sonderegger, S. (2019). *It's not a lie if you believe it: On norms, lying, and self-serving belief distortion* (Working Paper No. 2019–07). CeDEx Discussion Paper Series. https://www.econstor.eu/handle/10419/228354

Birkelund, J., & Cherry, T. L. (2020). Institutional inequality and individual preferences for honesty and generosity. *Journal of Economic Behavior & Organization*, *170*, 355–361. https://doi.org/10.1016/j.jebo.2019.12.014

Bizer, G. Y., Krosnick, J. A., Holbrook, A. L., Christian Wheeler, S., Rucker, D. D., & Petty, R. E. (2004). The impact of personality on cognitive, behavioral, and affective political processes: The effects of need to evaluate. *Journal of Personality*, *72*(5), 995–1028. https://doi.org/10.1111/j.0022-3506.2004.00288.x

Bizer, G. Y., Magin, R. A., & Levine, M. R. (2014). The Social-Norm Espousal Scale. *Personality and Individual Differences*, *58*, 106–111. https://doi.org/10.1016/j.paid.2013.10.014

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, *72*(2), 321–338. https://doi.org/10.1016/j.geb.2010.09.008

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, *41*(4), 540–558. https://doi.org/10.1177/0146167215572134

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, *88*, 1–45. https://doi.org/10.1037/0033-2909.88.1.1

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In *The Psychology of Economic Decisions. Vol. 1: Rationality and well-being* (Vol. 1, pp. 105–126). Oxford University Press.

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, *65*(11), 5371–5391. https://doi.org/10.1287/mnsc.2018.3191

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Bonner, J. M., Greenbaum, R. L., & Quade, M. J. (2017). Employee unethical behavior to shame as an indicator of self-image threat and exemplification as a form of self-image protection: The exacerbating role of supervisor bottom-line mentality. *Journal of Applied Psychology*, *102*(8), 1203–1221. https://doi.org/10.1037/apl0000222

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Boyce, C. J., Wood, A. M., & Ferguson, E. (2016). Individual differences in loss aversion: Conscientiousness predicts how life satisfaction responds to losses versus gains in income. *Personality and Social Psychology Bulletin*, *42*(4), 471–484. https://doi.org/10.1177/0146167216634060

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), 3281–3288. https://doi.org/10.1098/rstb.2009.0134

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875), 20180116. https://doi.org/10.1098/rspb.2018.0116

Burger, J. M., & Cooper, H. M. (1979). The desirability of control. *Motivation and Emotion*, *3*(4), 381–393. https://doi.org/10.1007/BF00994052

Burger, J. M., Sanchez, J., Imberi, J. E., & Grande, L. R. (2009). The norm of reciprocity as an internalized social norm: Returning favors even when no one finds out. *Social*

Influence, *4*(1), 11–17. https://doi.org/10.1080/15534510802131004

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*(1), 131–153. https://doi.org/10.1146/annurev-economics-063016-103625

Cain, D., & Dana, J. (2012). *Paying people to look at the consequences of their actions*. Citeseer.

Cain, D. M., Dana, J., & Newman, G. E. (2014). Giving versus giving in. *The Academy of Management Annals*, *8*(1), 505–533. https://doi.org/10.1080/19416520.2014.911576

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, *7*(5), 225–231. https://doi.org/10.1016/S1364-6613(03)00094-9

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827. https://doi.org/10.1257/aer.97.3.818

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., & Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, *118*(2), 280–283. https://doi.org/10.1016/j.econlet.2012.10.030

Cerrone, C., & Engel, C. (2019). Deciding on behalf of others does not mitigate selfishness: An experiment. *Economics Letters*, *183*, 108616. https://doi.org/10.1016/j.econlet.2019.108616

Chancel, L., Piketty, T., Saez, E., & Zucman, G. (2022). *World Inequality Report 2022*. Harvard University Press.

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1-2, 4th ed* (pp. 151–192). McGraw-Hill.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology*, *100*(5), 947.

Conrads, J., & Irlenbusch, B. (2013). Strategic ignorance in ultimatum bargaining. *Journal of Economic Behavior & Organization*, *92*, 104–115. https://doi.org/10.1016/j.jebo.2013.05.010

Conte, R., Andrighetto, G., & Campennì, M. (2010). Internalizing Norms: A Cognitive Model of (Social) Norms' Internalization. *International Journal of Agent Technologies and Systems (IJATS)*, *2*(1), 63–73. https://doi.org/10.4018/jats.2010120105

Cramwinckel, F. M., De Cremer, D., & van Dijke, M. (2013). Dirty hands make dirty leaders?! The effects of touching dirty objects on rewarding unethical subordinates as a function of a leader's self-interest. *Journal of Business Ethics*, *115*(1), 93–100. https://doi.org/10.1007/s10551-012-1385-4

Cullis, J., Jones, P., & Savoia, A. (2012). Social norms and tax compliance: Framing the decision to pay tax. *The Journal of Socio-Economics*, *41*(2), 159–168. https://doi.org/10.1016/j.socec.2011.12.003

Cusimano, C., & Lombrozo, T. (2023). People recognize and condone their own morally motivated reasoning. *Cognition*, *234*, 105379. https://doi.org/10.1016/j.cognition.2023.105379

D'Adda, G., Gao, Y., Golman, R., & Tavoni, M. (2018). *It's so hot in here: Information avoidance, moral wiggle room, and high air conditioning usage* (Working Paper ID 3149330). Social Science Research Network.

https://papers.ssrn.com/abstract=3149330

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don’t know won’t hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383. https://doi.org/10.1037/h0025589

De Groot, J. I. M., & Steg, L. (2009). Morality and Prosocial Behavior: The Role of Awareness, Responsibility, and Norms in the Norm Activation Model. *The Journal of Social Psychology*, *149*(4), Article 4. https://doi.org/10.3200/SOCP.149.4.425-449

De Vries, R. E. (2013). The 24-item brief HEXACO inventory (BHI). *Journal of Research in Personality*, *47*(6), 871–880.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, *127*(1), 1–56. https://doi.org/10.1093/qje/qjr050

Deutscher Spendenrat e.V. (2023, February 1). *Bilanz des Helfens 2022*. Deutscher Spendenrat e.V. https://www.spendenrat.de/bilanz-des-helfens-2022/

Donnellan, M. B., Lucas, R. E., & Fleeson, W. (2009). Introduction to personality and assessment at age 40: Reflections on the legacy of the person–situation debate and the future of person–situation integration. *Journal of Research in Personality*, *43*(2), 117–119. https://doi.org/10.1016/j.jrp.2009.02.010

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge University Press.

Dovidio, J. F. (1984). Helping behavior and altruism: An empirical and conceptual overview. *Advances in Experimental Social Psychology*, *17*, 361–427.

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General*, *142*, 93–100. https://doi.org/10.1037/a0028499

Duncan, B. (2009). Secret santa reveals the secret side of giving. *Economic Inquiry*, *47*(1), 165–181. https://doi.org/10.1111/j.1465-7295.2008.00145.x

Dunning, D. (2007). Self-image motives and consumer behavior: How sacrosanct self-beliefs sway preferences in the marketplace. *Journal of Consumer Psychology*, *17*(4), 237–249. https://doi.org/10.1016/S1057-7408(07)70033-5

Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*(2), 181–191.

Ehrich, K. R., & Irwin, J. R. (2005). Willful ignorance in the request for product attribute information. *Journal of Marketing Research*, *42*(3), 266–277. https://doi.org/10.1509/jmkr.2005.42.3.266

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), Article 4. https://doi.org/10.1007/s10683-011-9283-7

Erlandsson, A., Jungstrand, A. Å., & Västfjäll, D. (2016). Anticipated Guilt for Not Helping and Anticipated Warm Glow for Helping Are Differently Impacted by Personal Responsibility to Help. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01475

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*(2), 587–628. https://doi.org/10.1093/restud/rdv051

Exley, C. L., & Kessler, J. B. (2021). *Information avoidance and image concerns* (Working Paper No. 28376). National Bureau of Economic Research. https://doi.org/10.3386/w28376

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter – The impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, *112*(478), 1–33. https://doi.org/10.1111/1468-0297.00027

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and*

Human Behavior, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *Journal of Economic Psychology*, *45*, 253–267. https://doi.org/10.1016/j.joep.2014.10.003

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fiedler, S., Glöckner, A., Nicklisch, A., & Dickert, S. (2013). Social Value Orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational Behavior and Human Decision Processes*, *120*(2), 272–284. https://doi.org/10.1016/j.obhdp.2012.07.002

Fiedler, S., Hellmann, D. M., Dorrough, A. R., & Glöckner, A. (2018). Cross-national in-group favoritism in prosocial behavior: Evidence from Latin and North America. *Judgment and Decision Making*, *13*(1), 42–60. https://doi.org/10.1017/S1930297500008810

Fishbach, A., & Woolley, K. (2015). Avoiding ethical temptations. *Current Opinion in Psychology*, *6*, 36–40. https://doi.org/10.1016/j.copsyc.2015.03.019

Fishburn, P. C. (1970). *Utility theory for decision making*. Research analysis corp McLean VA.

Fiske, S. T. (2009). *Social beings: Core motives in social psychology*. John Wiley & Sons.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*. https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00059

Fricke, K., & Vogel, S. (2020). How interindividual differences shape approach-avoidance behavior: Relating self-report and diagnostic measures of interindividual differences to behavioral measurements of approach and avoidance. *Neuroscience & Biobehavioral Reviews*, *111*, 30–56. https://doi.org/10.1016/j.neubiorev.2020.01.008

Friedrichsen, J., & Engelmann, D. (2013). *Who Cares for Social Image? Interactions between Intrinsic Motivation and Social Image Concerns* (SSRN Scholarly Paper No. 2371250). https://doi.org/10.2139/ssrn.2371250

Funder, D. C. (2008). Persons, situations, and person-situation interactions. In *Handbook of personality: Theory and research, 3rd ed* (pp. 568–580). The Guilford Press.

Gächter, S., Gerhards, L., & Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, *97*, 72–86. https://doi.org/10.1016/j.euroecorev.2017.06.001

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), Article 7595. https://doi.org/10.1038/nature17160

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the External Validity of Social Preference Games: A Systematic Lab-Field Study. *Management Science*, *65*(3), 976–1002. https://doi.org/10.1287/mnsc.2017.2908

Garcia, T., Massoni, S., & Villeval, M. C. (2020). Ambiguity and excuse-driven behavior in charitable giving. *European Economic Review*, *124*, 103412. https://doi.org/10.1016/j.euroecorev.2020.103412

Gärtner, M., & Sandberg, A. (2017). Is there an omission effect in prosocial behavior? A laboratory experiment on passive vs. active generosity. *PLOS ONE*, *12*(3), 1–21. https://doi.org/10.1371/journal.pone.0172496

Gavrilets, S., & Richerson, P. J. (2017). Collective action and the evolution of social norm internalization. *Proceedings of the National Academy of Sciences*, *114*(23), 6068–6073. https://doi.org/10.1073/pnas.1703857114

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191–203. https://doi.org/10.1016/j.obhdp.2011.03.001

Giving USA Foundation. (2021). *Giving USA 2021: The annual report on philanthropy for the year 2020*. https://givingusa.org/

Glazer, A., & Konrad, K. A. (1996). A signaling explanation for charity. *American Economic Review*, *86*(4), 1019–1028.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*(8), 711–721.

Goffman, E. (1959). The Moral Career of the Mental Patient. *Psychiatry*, *22*(2), 123–142. https://doi.org/10.1080/00332747.1959.11023166

Gollwitzer, M., & Schwabe, J. (2020). *Context dependency as a predictor of replicability* [Working Paper]. PsyArXiv. https://doi.org/10.31234/osf.io/53yhg

Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, *5*, 143–164. https://doi.org/10.1037/dec0000068

Gotowiec, S., & van Mastrigt, S. (2019). Having versus doing: The roles of moral identity internalization and symbolization for prosocial behaviors. *The Journal of Social Psychology*, *159*(1), 75–91. https://doi.org/10.1080/00224545.2018.1454394

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Grossman, Z. (2010). *Strategic Ignorance and the Robustness of Social Preferences*. https://escholarship.org/uc/item/60b93868

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, *60*(11), 2659–2665. https://doi.org/10.1287/mnsc.2014.1989

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, *117*, 26–39. https://doi.org/10.1016/j.jebo.2015.05.008

Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, *15*(1), 173–217. https://doi.org/10.1093/jeea/jvw001

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Haesevoets, T., Folmer, C. R., & Van Hiel, A. (2015). Cooperation in mixed–motive games: The role of individual differences in selfish and social orientation. *European Journal of Personality*, *29*(4), 445–458. https://doi.org/10.1002/per.1992

Haley, K. J., & Fessler, D. M. T. (2005). Nobody’s watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*(3), 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413. https://doi.org/10.1177/0146167206291006

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, *54*, 35–43. https://doi.org/10.1016/j.joep.2016.03.001

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123–135. https://doi.org/10.1002/evan.10110

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.

C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359–372. https://doi.org/10.1177/1745691616635594

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*(3), 319.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280.

Hightow, L. B., Miller, W. C., Leone, P. A., Wohl, D., Smurzynski, M., & Kaplan, A. H. (2003). Failure To Return for HIV Posttest Counseling in an STD Clinic Population. *AIDS Education and Prevention*, *15*(3), 282–290. https://doi.org/10.1521/aeap.15.4.282.23826

International Organization for Migration. (2022). *World Migration Report 2022*. https://publications.iom.int/books/world-migration-report-2022

Jacobson, R. P., Mortensen, C. R., & Cialdini, R. B. (2011). Bodies obliged and unbound: Differentiated response tendencies for injunctive and descriptive social norms. *Journal of Personality and Social Psychology*, *100*(3), 433–448. https://doi.org/10.1037/a0021470

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*, 172–194. https://doi.org/10.1037/0022-3514.70.1.172

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). Elsevier.

Jordan, J., Leliveld, M. C., & Tenbrunsel, A. E. (2015). The moral self-image scale: Measuring and understanding the malleability of the moral self. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01878

Kandul, S., & Ritov, I. (2017). Close your eyes and be nice: Deliberate ignorance behind pro-social choices. *Economics Letters*, *153*, 54–56. https://doi.org/10.1016/j.econlet.2017.02.010

Kawamura, Y., Ohtsubo, Y., & Kusumi, T. (2021). Effects of cost and benefit of prosocial behavior on reputation. *Social Psychological and Personality Science*, *12*(4), 452–460. https://doi.org/10.1177/1948550620929163

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*(1), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*, 23–34. https://doi.org/10.1037/0003-066X.43.1.23

Kieslich, P. J., & Hilbig, B. E. (2014). Cognitive conflict in social dilemmas: An analysis of response dynamics. *Judgment and Decision Making*, *9*(6), 510–522.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608–638. https://doi.org/10.1111/jeea.12152

Kohlberg, L. (1981). *The philosophy of moral development*. Harpers.

Kohlberg, L., & Kramer, R. (1969). Continuities and discontinuities in childhood and adult moral development. *Human Development*, *12*(2), 93–120.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, *90*(4), 1072–1091.

https://doi.org/10.1257/aer.90.4.1072

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making*, *8*(5), 527–539. https://doi.org/10.1017/S1930297500003636

Koop, G. J., & Johnson, J. G. (2011). Response dynamics: A new window on the decision process. *Judgment and Decision Making*, *6*(8), 750–758. https://doi.org/10.1017/S1930297500004186

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive Psychology*, *67*(4), 151–185. https://doi.org/10.1016/j.cogpsych.2013.09.001

Köster, M., Schuhmacher, N., & Kärtner, J. (2015). A cultural perspective on prosocial development. *Human Ethology Bulletin*, *30*(1), 71–82.

Kreps, D. M. (1979). A representation theorem for “preference for flexibility.” *Econometrica*, *47*(3), 565–577. https://doi.org/10.2307/1910406

Kruglanski, A. W. (1989). The psychology of being “right”: The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*(3), 395.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: “Seizing” and “freezing.” *Psychological Review*, *103*, 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology*, *65*(5), 861.

Krupka, E. L., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), 307–320. https://doi.org/10.1016/j.joep.2008.11.005

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006

Krysowski, E., & Tremewan, J. (2021). Why does anonymity make us misbehave: Different norms or less compliance? *Economic Inquiry*, *59*(2), 776–789. https://doi.org/10.1111/ecin.12955

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lakatos, I. (1968). Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, *69*, 149–186.

Larson, T., & Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? A comment. *Judgment and Decision Making*, *4*(6), 467–474.

Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, *4*(1), 136–163. https://doi.org/10.1257/app.4.1.136

Le Lec, F., & Tarroux, B. (2020). On attitudes to choice: Some experimental evidence on choice aversion. *Journal of the European Economic Association*, *18*(5), 2108–2134. https://doi.org/10.1093/jeea/jvz036

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*(3), 371–375. https://doi.org/10.1177/0146167283093007

Leary, M. R., Jongman-Sereno, K. P., & Diebels, K. J. (2015). Measures of concerns with public image and social evaluation. In *Measures of personality and social psychological constructs* (pp. 448–473). Elsevier.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, *25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: A meta-analytic review. *Social Influence*, *11*(3), 199–215. https://doi.org/10.1080/15534510.2016.1216891

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives*, *21*(2), Article 2. https://doi.org/10.1257/jep.21.2.153

Lin, S. C., & Reich, T. (2018). To give or not to give? Choosing chance under moral conflict. *Journal of Consumer Psychology*, *28*(2), 211–233. https://doi.org/10.1002/jcpy.1008

Lin, S. C., Schaumberg, R. L., & Reich, T. (2016). Sidestepping the rock and the hard place: The private avoidance of prosocial requests. *Journal of Experimental Social Psychology*, *64*, 35–40. https://doi.org/10.1016/j.jesp.2016.01.011

Lindsey, L. L. M., Yun, K. A., & Hill, J. B. (2007). Anticipated guilt as motivation to help unknown others: An examination of empathy as a moderator. *Communication Research*, *34*(4), 468–480.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482–493. https://doi.org/10.1086/519249

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Locey, M. L., Jones, B. A., & Rachlin, H. (2013). Self-control and altruism. In *APA handbook of behavior analysis, Vol. 1: Methods and principles* (pp. 463–481). American Psychological Association. https://doi.org/10.1037/13937-020

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272–292. https://doi.org/10.1006/obhd.1996.0028

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, *64*(1), 237–267. https://doi.org/10.1016/j.geb.2007.10.009

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324.

Matthey, A., & Regner, T. (2011). Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior. *Games*, *2*(1), 114–135. https://doi.org/10.3390/g2010114

Matthey, A., & Regner, T. (2014). *More than outcomes: The role of self-image in* *other-regarding behavior* (Working Paper No. 2014–036). Jena Economic Research Papers. https://www.econstor.eu/handle/10419/108543

McAuliffe, W. H. B., Forster, D. E., Pedersen, E. J., & McCullough, M. E. (2019). Does cooperation in the laboratory reflect the operation of a broad trait? *European Journal of Personality*, *33*(1), Article 1. https://doi.org/10.1002/per.2180

McClintock, C. G. (1972). Social motivation—A set of propositions. *Behavioral Science*, *17*(5), 438–454. https://doi.org/10.1002/bs.3830170505

Miller, J. G., & Bersoff, D. M. (1992). Culture and moral judgment: How are conflicts between justice and interpersonal responsibilities resolved? *Journal of Personality and Social Psychology*, *62*(4), 541.

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*, 33–47. https://doi.org/10.1037/0022-3514.58.1.33

Molitor, F., Bell, R. A., Truax, S. R., Ruiz, J. D., & Sun, R. K. (1999). Predictors of failure to return for HIV test result and counseling by test site type. *AIDS Education and Prevention*, *11*(1), 1–13.

Momsen, K., & Ohndorf, M. (2020). When do people exploit moral wiggle room? An experimental analysis of information avoidance in a market setup. *Ecological Economics*, *169*, 106479. https://doi.org/10.1016/j.ecolecon.2019.106479

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. *Personality, Identity, and Character: Explorations in Moral Psychology*, 341–354.

Monin, B., & Norton, M. I. (2003). Perceptions of a fluid consensus: Uniqueness bias, false consensus, false polarization, and pluralistic ignorance in a water conservation crisis. *Personality and Social Psychology Bulletin*, *29*(5), 559–567.

Moradi, H. (2018). *Selfless ignorance: Too good to be true* (Working Paper SP II 2018-208). WZB Discussion Paper. https://www.econstor.eu/handle/10419/194000

Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both? A behavioral test of impression management and overclaiming. *Psychological Assessment*, *31*(2), 181–191. https://doi.org/10.1037/pas0000657

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13–41. https://doi.org/10.1177/1088868313501745

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Narvaez, D., & Lapsley, D. K. (2009). Moral identity, moral functioning, and the development of moral character. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 50, pp. 237–274). Academic Press. https://doi.org/10.1016/S0079-7421(08)00408-8

Nisan, M., & Horenczyk, G. (1990). Moral balance: The effect of prior behaviour on decision in moral conflict. *British Journal of Social Psychology*, *29*(1), 29–42. https://doi.org/10.1111/j.2044-8309.1990.tb00884.x

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*(6), 817–831. https://doi.org/10.1037/0022-3514.87.6.817

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Nyborg, K. (2011). I don’t want to hear about it: Rational ignorance among duty-oriented consumers. *Journal of Economic Behavior & Organization*, *79*(3), 263–274. https://doi.org/10.1016/j.jebo.2011.02.004

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Ockenfels, A., & Werner, P. (2012). ‘Hiding behind a small cake’ in a newspaper dictator game. *Journal of Economic Behavior & Organization*, *82*(1), 82–85. https://doi.org/10.1016/j.jebo.2011.12.008

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 4716.

Otto, A. S., Clarkson, J. J., & Kardes, F. R. (2016). Decision sidestepping: How the motivation for closure prompts individuals to bypass decision making. *Journal of Personality and Social Psychology*, *111*, 1–16. https://doi.org/10.1037/pspa0000057

Penner, L. A., & Finkelstein, M. A. (1998). Dispositional and structural determinants of volunteerism. *Journal of Personality and Social Psychology*, *74*, 525–537. https://doi.org/10.1037/0022-3514.74.2.525

Petty, R. E., Brinol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In *Handbook of individual differences in social behavior* (pp. 318–329). The Guilford Press.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms5939

Pfattheicher, S., Nielsen, Y. A., & Thielmann, I. (2022). Prosocial behavior and altruism: A review of concepts and definitions. *Current Opinion in Psychology*, *44*, 124–129. https://doi.org/10.1016/j.copsyc.2021.08.021

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

Popper, K. (1934). *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Verlag von Julius Springer.

Rabin, M. (1995). *Moral preferences, moral constraints, and self-serving biases*. Berkeley Department of Economics Working Paper No. 95-241.

Rachlin, H. (2002). Altruism and selfishness. *Behavioral and Brain Sciences*, *25*(2), 239–250.

Rama, H.-O., Roberts, D., Tignor, M., Poloczanska, E. S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B., & Ayanlade, S. (2022). *Climate Change 2022: Impacts, Adaptation and Vulnerability Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. https://doi.org/10.1017/9781009325844

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Regner, T. (2021). What's behind image? Toward a better understanding of image-driven behavior. *Frontiers in Psychology*, *12*, 614575. https://doi.org/10.3389/fpsyg.2021.614575

Reichenberger, J., Schwarz, M., König, D., Wilhelm, F. H., Voderholzer, U., Hillert, A., & Blechert, J. (2015). Angst vor negativer sozialer Bewertung: Übersetzung und Validierung der Furcht vor negativer Evaluation–Kurzskala (FNE-K). *Diagnostica*, *62*(3), 169–181. https://doi.org/10.1026/0012-1924/a000148

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104

Rothmund, T., & Baumert, A. (2014). Shame on me: Implicit assessment of negative moral self-evaluation in shame-proneness. *Social Psychological and Personality Science*, *5*(2), 195–202.

Russo, J. E. (2019). Eye fixations as a process trace. In *A Handbook of Process Tracing Methods* (2nd ed.). Routledge.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92. https://doi.org/10.1177/1043463195007001004

Sassenberg, K., & Hansen, N. (2007). The impact of regulatory focus on affective responses to social discrimination. *European Journal of Social Psychology*, *37*(3), 421–444. https://doi.org/10.1002/ejsp.358

Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S., & Berlia, N. (2010). What counts as a choice? U.S. Americans are more likely than indians to construe actions as choices. *Psychological Science*, *21*(3), 391–398. https://doi.org/10.1177/0956797609359908

Schulte-Mecklenbeck, M., & Huber, O. (2003). Information search in the laboratory and on the Web: With or without an experimenter. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 227–235. https://doi.org/10.3758/BF03202545

Schulte-Mecklenbeck, M., Kuehberger, A., & Johnson, J. G. (Eds.). (2019). *A handbook of process tracing methods: 2nd edition* (2nd ed.). Routledge. https://doi.org/10.4324/9781315160559

Schwartz, S. H. (1977). Normative Influences on Altruism. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 10, pp. 221–279). Academic Press. https://doi.org/10.1016/S0065-2601(08)60358-5

Sezer, O., Gino, F., & Bazerman, M. H. (2015). Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology*, *6*, 77–81.

Shaw, L. L., Batson, C. D., & Todd, R. M. (1994). Empathy avoidance: Forestalling feeling for another in order to escape the motivational consequences. *Journal of Personality and Social Psychology*, *67*(5), 879–887. https://doi.org/10.1037/0022-3514.67.5.879

Sheldon, O. J., & Fishbach, A. (2015). Anticipating and resisting the temptation to behave unethically. *Personality and Social Psychology Bulletin*, *41*(7), 962–975. https://doi.org/10.1177/0146167215586196

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Smith, A. (1759). *The theory of moral sentiments*. Liberty Fund: Indianapolis.

Smith, J. (2012). The endogenous nature of the measurement of social preferences. *Mind & Society*, *11*(2), 235–256. https://doi.org/10.1007/s11299-012-0110-4

Smith, J. R., Louis, W. R., Terry, D. J., Greenaway, K. H., Clarke, M. R., & Cheng, X. (2012). Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *Journal of Environmental Psychology*, *32*(4), 353–361.

https://doi.org/10.1016/j.jenvp.2012.06.001

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979a). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*(12), 2297–2306. https://doi.org/10.1037/0022-3514.37.12.2297

Spivey, M. J. (2007). *The continuity of mind*. Oxford University Press.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. https://doi.org/10.1073/pnas.0503903102

Stiegler, G. J. (1961). The economics of information. *Journal of Political Economy*, *69*(3), 213–225.

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology*, *14*(4), 340–353. https://doi.org/10.1037/a0021288

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tao, G., Branson, B. M., Kassler, W. J., & Cohen, R. A. (1999). Rates of Receiving HIV Test Results: Data From the U.S. National Health Interview Survey for 1994 and 1995. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *22*(4), 395.

Tasimi, A., Dominguez, A., & Wynn, K. (2015). Do-gooder derogation in children: The social costs of generosity. *Frontiers in Psychology*, *6*, 1036. https://doi.org/10.3389/fpsyg.2015.01036

Teoh, Y. Y., & Hutcherson, C. A. (2022). The games we play: Prosocial choices under time pressure reflect context-sensitive information priorities. *Psychological Science*, *33*(9), 1541–1556. https://doi.org/10.1177/09567976221094782

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), Article 1. https://doi.org/10.1037/bul0000217

Thøgersen, J. (2008). Social norms and cooperation in real-life social dilemmas. *Journal of Economic Psychology*, *29*(4), 458–472. https://doi.org/10.1016/j.joep.2007.12.004

Thunström, L., Veld, K. van 't, Shogren, J. F., & Nordström, J. (2014). On strategic ignorance of environmental harm and social norms. *Revue d'economie Politique*, *Vol. 124*(2), 195–214. https://doi.org/10.3917/redp.242.0195

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061. https://doi.org/10.2307/2937956

van der Weele, J. J. (2014). Inconvenient truths: Determinants of strategic ignorance in moral dilemmas. *Available at SSRN 2247288*.

Van Lange, P. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337

Van Lange, P. A., Bekkers, R., Schuyt, T. N., & Vugt, M. V. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384.

Vu, L., Soraperra, I., Leib, M., van der Weele, J. J., & Shalvi, S. (2023). *Willful ignorance: A meta-analytic review*. Manuscript submitted for publication.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral Judgment. In *The Oxford Handbook of Thinking and Reasoning* (p. 364).

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049.

West, R., Godinho, C. A., Bohlen, L. C., Carey, R. N., Hastings, J., Lefevre, C. E., & Michie,

S. (2019). Development of a formal system for representing behaviour-change theories. *Nature Human Behaviour*, *3*(5), 526–536. https://doi.org/10.1038/s41562-019-0561-2

West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, *32*(4), 231–262.

Willemsen, M. C., & Johnson, E. J. (2019). (Re)Visiting the decision factory: Observing cognition with MouselabWEB. In *A Handbook of Process Tracing Methods* (2nd ed.). Routledge.

Willer, R. (2009). Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review*, *74*(1), 23–43. https://doi.org/10.1177/000312240907400102

Winterich, K. P., Aquino, K., Mittal, V., & Swartz, R. (2013). When moral identity symbolization motivates prosocial behavior: The role of recognition and moral identity internalization. *Journal of Applied Psychology*, *98*(5), 759–770. https://doi.org/10.1037/a0033177

Woolley, K., & Risen, J. L. (2018). Closing your eyes to follow your heart: Avoiding information to protect a strong intuitive preference. *Journal of Personality and Social Psychology*, *114*(2), 230–245. https://doi.org/10.1037/pspa0000100

Woolley, K., & Risen, J. L. (2021). Hiding from the truth: When and how cover enables information avoidance. *Journal of Consumer Research*, *47*(5), 675–697. https://doi.org/10.1093/jcr/ucaa030

Wulff, D. U., Kieslich, P. J., Henninger, F., Haslbeck, J. M., & Schulte-Mecklenbeck, M. (2021). *Movement tracking of cognitive processes: A tutorial using mousetrap*. PsyArXiv. https://doi.org/. https://doi.org/10.31234/osf.io/v685r.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., Inukai, K., Tanida, S., Kiyonari, T., Takagishi, H., & Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality.

Organizational Behavior and Human Decision Processes, *120*(2), 260–271. https://doi.org/10.1016/j.obhdp.2012.06.002

Zhang, L., & Ortmann, A. (2014). The effects of the take-option in dictator-game experiments: A comment on Engel's (2011) meta-study. *Experimental Economics*, *17*(3), 414–420. https://doi.org/10.1007/s10683-013-9375-7