How what goes on in others' minds affects our choices and well-being

A dissertation submitted to the
Department of Social and Decision Sciences
in partial fulfillment of the requirements

for the degree of
Doctor of Philosophy in Behavioral Decision Research
by

Andras Molnar

Dissertation Committee:
George Loewenstein (SDS, chair)
Shereen Chaudhry (U Chicago)
Silvia Saccardo (SDS)

Carnegie Mellon University
Pittsburgh, Pennsylvania
November 2020

This page intentionally left blank

ABSTRACT

WHY do we care about what others think and believe? How does what happens in other people's minds affect our well-being? When are we motivated to take actions, such as attempts to change another's mind, or to reveal harmful information to others, just to make sure that others believe what we want them to believe? How can these insights about people's preference for what others believe inform theories of decision-making and policy? These are the main research questions that I focus on in my dissertation. The overarching theme of the present work is the idea that we inherently monitor and care about what goes on in other people's minds—and not necessarily because doing so benefits us in any way. As I highlight in Chapter I, most previous work has hypothesized that such preferences over others' mental states serve as intermediate steps towards an ultimate goal (e.g., to outwit an opponent or to foster social relations). By contrast, my work demonstrates that individuals' well-being and choices can be *directly* affected by consideration for others' beliefs. My work also expands our understanding of *belief-based preferences* to include preferences over second-order beliefs, i.e., the beliefs of others. While previous theories of belief-based motives have examined individuals' preferences over *their own* non-instrumental cognitive states, my dissertation demonstrates that such belief-based motives can be extended *beyond the individual*. That is, people have an intrinsic preference for what others (should) believe, and this desire has important implications to their well-being and behavior in a multitude of domains. In Chapter II, I demonstrate that people inherently dislike when they think that others hold incorrect beliefs—as opposed to different beliefs *per se*—and argue that this finding puts prior literature on belief-homophily in a new light. In the subsequent chapters I investigate behavior in two domains, in which people take costly actions to correct others' misunderstandings: resource allocation (Chapter III) and moral punishment (Chapter IV). I conclude by discussing the limitations of the present work in Chapter V. In addition, I provide an outline for future research and discuss possible applications of belief-based motives in various domains. Taking into account an intrinsic preference over others' mental states can help us to better understand a plethora of contemporary societal issues: the polarization of political beliefs and belief-based geographic sorting; the dramatic deterioration of public trust in democratic institutions and the media (and the emergence of "fake news"); the psychological effects of the rapid acceleration of automatization and the increasing prevalence of human-computer interactions; and the worsening mental health conditions due to people feeling misunderstood, isolated, and "left behind" by society, which might also contribute to the recent surge in anti-establishment and extremist sentiments across the globe.

This page intentionally left blank

Dedicated to Granny Irene, the strongest woman I know.

Irén Maminak, a legerősebb nőnek, akit ismerek.

This page intentionally left blank

Acknowledgements

WRITING a doctoral dissertation is typically seen as the epitome of individual scholarly work; a lonely endeavor that lasts four, five, six years, or—as in my case—even longer. However, nothing could be farther from the truth. In reality, I could not have completed this dissertation without the tremendous support, motivation, and guidance I received from my family, friends, and colleagues. First of all, I would like to thank my loving parents, who served as wonderful role models: my Mom as a teacher and my Dad as a writer. It is no coincidence that I ended up pursuing a career in academia. I consider myself extremely lucky for having met the Marvellous Ms. Ma, who not only supported me during the most challenging part of my journey, but also helped me discover that even though life is often crispy on the outside, it is always soft on the inside. I am also immensely grateful for the amazing friends I have made since I started my new "American life": Varun, my brother from another mother; and the other three members of the *phantastic phour*: Ali, Nik, and Steph. Without you, and the hundreds (thousands?) of hours of quality time spent together, this journey would have been much less enjoyable and meaningful—see you in the introduction! I also thank my best friends in Hungary: Endre, Lilla, Sari, and Zsófi, who kept supporting me in my new life, even though there was (and still is) an ocean and 4,600 miles between us.

Of course, none of this would have been possible without the consistent support of my advisor, George, who not only offered me a once-in-a-lifetime opportunity (of transferring to SDS), but also helped me to discover my true potential, for which I am extremely grateful. I also thank my wonderful committee: Shereen and Silvia; as well as my other co-authors and mentors who all helped me to become a better academic: Alex, Eric, Julie, and Russell. I am also grateful for the tremendous support I have received from the SDS administrative staff, especially from Mary Anne and Rosa—they made my journey as smooth as possible, without having to worry about administrative issues. I would also like to thank Dr. Bram, M. Eberhart, and P. Durban for helping me to get through the most difficult parts of this journey. Last but not least, I want to thank the staff of the following fine establishments, where I spent countless hours brainstorming with friends, designing research, and writing papers: 61C Cafe, Café Phipps, Dobra Tea Pittsburgh, Pamela's Diner (Squirrel Hill), Squirrel Hill Cafe ("The Cage"), and the Carnegie Library of Pittsburgh (Oakland).

— Pittsburgh, November 2020.

This page intentionally left blank

Contents

List of Figures

List of Tables

This page intentionally left blank

"Belief is nearly the whole of the universe
whether based on truth or not."

— *Kurt Vonnegut* (1987)

Chapter I

Thoughts and Players

We are social creatures, and as such, require a complex set of cognitive tools to navigate in a complex social world. We not only have to constantly monitor the behavior of others and pay attention to the consequences of their actions (i.e., that we can observe), but we also have make inferences about what's going on in their minds (i.e., that is hidden from plain sight): their intentions, thoughts, and feelings. Will my parents understand if I am not going home for Thanksgiving this year? Does my boss realize how much work I put into the annual report? Will the audience appreciate my jokes during my presentation? Does my friend really believe that he is going to win the lottery? Will readers understand the dozens of cultural references that I have carefully hidden in this dissertation?

These are just a few examples of the type questions that all of us face on a daily basis, each of which require us making inferences about others' beliefs and mental states. Fortunately, nature has equipped humankind with an astonishingly advanced cognitive toolbox, that allows us to represent what's happening in others' minds—even if these mental states are different from our own. This remarkable ability goes by many names, most prominently: theory of mind (Dennett, 1978; Leslie, 1987; Premack & Woodruff, 1978), mind reading (Heyes & Frith, 2014; Sperber & Wilson, 2002), and mentalizing (Bateman & Fonagy, 2012; Frith & Frith, 2003). Research documents that most adults, and even most children between the ages of 6–7, understand that others can have beliefs different from one's own (Perner & Wimmer, 1985). Moreover, people can distinguish between levels of mental representations (e.g., between "A believes X" and "B believes that A believes X"), and by adulthood, they can represent about *fourth-order* levels of shared knowledge (e.g., Ali knows that Stephanie knows that Nik knows that Andras knows X; Kinderman, Dunbar, & Bentall, 1998). Finally, people seem to be able to track others' beliefs relatively effortlessly and automatically (van der Wel, Sebanz, & Knoblich, 2014), and engage in implicit mentalizing even without being consciously aware of doing so (Schneider, Slaughter, & Dux, 2015), which further highlight how well-fitted humans are when it comes to reading others' minds.

I.1 Belief-based utility

While social and cognitive psychology often investigates belief-forming processes and the ability to infer others' mental states on its own right, mainstream economists have long considered beliefs to be secondary: intermediate constructs, or, decision aids, that allow individuals to make better decisions. However, this reductionist approach has never been universally accepted among economists, and an alternative paradigm—that it is essential to understand how people form beliefs and what preferences they have over beliefs—is gaining foothold even in traditional fields of economics, such as finance (Sicherman, Loewenstein, Seppi, & Utkus, 2016), labor economics (Ariely, Kamenica, & Prelec, 2008), health economics (Oster, Shoulson, & Dorsey, 2013), and public policy (Hauser, Gino, & Norton, 2018). Nevertheless, since economics relies on a narrower and more parsimonious definition of what beliefs are—and what beliefs are not—than other disciplines, in which the delineation between beliefs and non-beliefs is less clear, in my work I focus on beliefs as defined in economics:

Beliefs are subjective probability assessments or expectations over outcomes or states of the world.

I.1.1 What belief-based utility is (and what it is not)

By *belief-based utility*, I refer to the utility derived from holding these subjective probability assessments, whether or not they correspond to objective probabilities, and independent of the utility derived from actual outcomes or states of the world. This definition has an important property that makes it particularly appealing to empiricists: According to this definition, beliefs have a normative benchmark, that is, an objective probability of outcomes or states, which the subjective belief can be compared to, and evaluated against. This implies that beliefs are—in principle—always *verifiable* (or falsifiable), unlike, for instance, taste or preferences, which are idiosyncratic and don't have a normative benchmark.

It is important to highlight that belief-based utility is different from *conceptual consumption* (e.g., Ariely & Norton, 2009) which describes the phenomenon when people derive utility from consuming non-physical goods, i.e., concepts (e.g., most of modern entertainment), in addition to physical consumption. Although the term "belief" is often conflated with mental states in general, or more vaguely refer to anything that "happens in the mind," it is important to highlight that belief-based utility, as defined above, is different from the utility experienced when consuming non-physical goods or when facing non-physical outcomes. Outcomes in standard economic (expected utility) models are not restricted to physical

ones: outcomes can include feelings such as anger or pride, and abstract concepts such as friendship, reputation, or identity. These can be added to any expected utility function as additional elements to the vector of outcomes that the decision-maker cares about, without affecting the subjective probability assessments in any way.

To demonstrate the difference between belief-based utility and conceptual consumption, consider the following example: Miriam, a talented young woman from New York, is contemplating whether she should pursue a career in stand up comedy, or whether she should climb the corporate ladder, just like the rest of her friends. Now, assume that Miriam is perfectly aware that her chance of ever becoming successful as a comedian is objectively low—and she has accurate beliefs about her expected lifetime earnings. In this situation, Miriam might still prefer the fulfilling but risky career with a lower expected lifetime income (i.e., becoming a stand up comedian) to the safe but dull career with a higher salary (i.e., getting a corporate job), even if her beliefs about the lifetime incomes are perfectly accurate. There is no need for belief-based utility in this case: Miriam simply cares more about feeling fulfilled (a non-physical outcome) than about her material wealth (a physical outcome).

Now, imagine a different scenario, in which Miriam has unrealistic beliefs about becoming a wildly successful stand up comedian. One mediocre performance after another, but she refuses to update her expectations (which would allow her to make objectively better decisions, and most likely, give up comedy), because she simply cherishes the idea of becoming a superstar. In other words, she derives utility *directly* from maintaining her dream, regardless of the expected outcomes (non-physical included). Doing so will make her feel good in the present, even if in reality her chance of becoming a celebrated stand up comedian is extremely low. In this second scenario, Miriam prefers the career in comedy over the corporate job *not* because of the extra feeling of fulfillment (a non-physical outcome), but because she will be able to maintain her cherished dream (i.e., a subjective probability assessment of becoming a successful comic), that is, derive utility directly from her beliefs.

I.1.2 The brief history of belief-based utility

The idea that people gain utility *directly* from the beliefs they hold, regardless of objective outcomes and states of the world, was already recognized by pioneering neoclassical economists, who built a new approach to economics on Bentham's notion of utility maximization. Bentham himself provided a list of the sources of cardinal utility (pleasure) and cardinal disutility (pain) which featured belief-based elements, most prominently expectation and imagination: "The pleasures of expectation are those that result from contemplating any sort of pleasure thought of as future, accompanied with the sentiment of belief" (Bentham, 1789, p. 36–37).

Given Bentham's embrace of belief-based utility, it would have been natural that beliefs would end up as part of neoclassical economics, which embraced Bentham's concept of utility as its foundation. Yet, though early neoclassical economists discussed belief-based elements of utility extensively (see Loewenstein, 1992), they struggled to incorporate beliefs into the new framework they were developing for economics. Sidestepping these challenges, the emerging preference-based ordinal utility and the revealed preference approach (Samuelson, 1948) remained agnostic towards beliefs as sources of utility and inferred utility from revealed preferences. This approach was bolstered by the ascendancy of behaviorism in psychology with its similar positivist focus on observable behavior.

Around the same time, Neumann and Morgenstern axiomatized Bernoulli's (1738 / 1954) theory of expected utility (Neumann & Morgenstern, 1947). In Expected Utility Theory (EUT) "beliefs" are objective probabilities which function as decision weights applied to the utilities associated with outcomes or states of the world. Therefore, probabilities in EUT are merely *decision aids*, towards which the decision-maker has a perfectly neutral attitude—the decision-maker's choices reveal her *preferences over outcomes*, weighted by objective probabilities. In other words, in EUT people derive cardinal utility only from outcomes.

However, individuals rarely have access to objective probabilities (except for, perhaps, in the domain of gambling), and must rely on their imperfect, subjective assessments instead. Savage was the first who incorporated this insight into EUT. His Subjective Expected Utility (SEU) model (Savage, 1954) relaxes the assumption that decision-makers rely on the objective probabilities of outcomes and allows them to make subjective probability judgments instead, which implies that beliefs can substantially deviate from objective probabilities (Equation I.1):

$$E[U(x)] = \sum_i p(x_i)u(x_i) \tag{I.1}$$

where $u(x_i)$ is the utility of outcome x_i, and $p(x_i)$ corresponds to the individual's subjective belief about the probability of outcome x_i. Importantly, however, subjective probabilities in SEU are still nothing more than *decision weights*, just like the objective probabilities in EUT. A SEU decision-maker is still perfectly neutral towards the content of her beliefs: She only cares about the accuracy of beliefs, since holding more accurate probability judgments (i.e., ones that resemble objective probabilities more) allows her to make better decisions and increase her utility. This implies that a SEU decision-maker never misses an opportunity to obtain free information, and always updates her beliefs according to the Bayes

rule. Eventually, by gathering more and more information, her beliefs will approximate the objective probabilities, eliminating prior biases that she might have had. Furthermore, there is no reason to expect systematic biases in prior, and especially in posterior, beliefs: if beliefs are based on sparse, noisy and incomplete information, it is just as likely that someone will underestimate the probability of some outcome as overestimating it. Yet, a multitude of phenomena suggest that people do not behave according to the above principles, and empirical evidence overwhelmingly supports the idea that people also care *directly* about their beliefs, that is, they derive utility from holding a particular set of beliefs, regardless of the outcomes in expectation (see e.g., Bénabou & Tirole, 2016; Golman, Hagmann, & Loewenstein, 2017; Loewenstein & Molnar, 2018). Without making assumptions about complex, non-linear relationships between outcomes and beliefs, belief-based utility can be incorporated into economic models according to the following simple equation (Equation I.2):

$$U = E[U(x)] + \theta v(P) = \sum_i p(x_i)u(x_i) + \theta v(P) \qquad \text{(I.2)}$$

Total utility U is the sum of two parts: 1) the expected utility $E[U(X)]$ which expresses the decision-maker's expected utility over outcomes; and 2) the belief-based utility $v(P)$ which indicates the utility derived from the set of beliefs P. The weight parameter θ captures the relative importance of utility from outcomes and utility from beliefs, that is, to what extent does the decision-maker care about obtaining better outcomes (i.e., based on accurate beliefs) versus holding more desirable beliefs.[1]

I.1.3 The next step: deriving utility from *others'* beliefs

The two approaches that I briefly introduced so far—the psychology of mind reading and belief-based utility—offer a logical next step: What if people also derive utility from what others believe, in addition to their own beliefs?[2] There is a long history of incorporating others' beliefs and intentions into game theory (Battigalli, Corrao, & Dufwenberg, 2019; Carpenter & Matthews, 2003; Charness & Rabin, 2002; Geanakoplos, Pearce, & Stacchetti, 1989), and research also shows that people are well-adapted to inferring others' intentions in economic interactions (Cushman, 2015; Heintz, Karabegovic, & Molnar, 2016; Rand, Fudenberg, & Dreber, 2015; Sutter, 2007). However, this work has exclusively focused on

[1] Note that if $\theta = 0$ (i.e., the individual is neutral towards her beliefs), the model reduces to SEU.

[2] Throughout my dissertation, I simply refer to "utility from others' beliefs," which is more parsimonious than the fully accurate description: *utility from one's second-order beliefs about someone else's beliefs*. However, for the sake of simplicity, I keep using the shorter expression of "utility from others' beliefs," but this always refers to the individual's second-order beliefs about another person's beliefs.

"strategic" interactions, in which knowing about someone else's mental state can actually help one to make better decisions. In such strategic contexts—which can be as innocuous as playing rock-paper-scissors with friends, or as consequential as launching a military strike—outcomes depend on how accurately one can infer others' mental states and intentions. Therefore, the preference for knowing what others believe always serves an instrumental goal in these situations, and thus, requires relatively little explanation.

Other work has hypothesized that preferences over what others believe can facilitate coordination within groups or help strengthen group identity (e.g., Abelson, 1986; Bénabou, 2013). Golman, Loewenstein, Moene, and Zarri (2016) review studies and theories that assert that people have a preference for belief consonance—that is, to maintain similar beliefs to that of their peers with whom they associate with. As the authors discuss, "people want to achieve belief consonance both because it cements their connection to groups, because it protects core values and beliefs about the self, and likely because they don't want to write off investments that they made on the basis of their beliefs" (p. 172). Therefore, the dominant beliefs that are shared within a group the individual belongs to—or aspires to belong to—will be prescriptive for the individual. Expressing dissent with the group's dominant belief system could lead to detrimental psychological and social consequences: weaken one's identity and social ties within the group, or in the worst case, culminate in ostracism. A Republican, for example, might lose friends by openly expressing a belief that climate change is caused by human activity, or raising concerns about domestic gun violence, and this social cost looms much larger than the benefit of holding and articulating an opposing belief. Such motives at the individual level can lead to enormous societal consequences, for example, to "pluralistic ignorance"—in which the majority openly supports a norm or regulation that contradicts with the majority's (private) preferences (Prentice & Miller, 1993). However, such motives are still fundamentally instrumental—since caring about what others believe serves the ultimate goal of tightening one's connection to their peers.

By contrast, in my dissertation I argue that people care about what others believe in a far broader set of situations, even if there is apparently no instrumental value in doing so. Furthermore, I demonstrate that people have preferences over what others *should* believe, and that if there is a discrepancy between others' current beliefs and their desired beliefs (according to one's preferences), then one will experience negative emotions and discomfort, which, in turn, might trigger actions: People will attempt to change others' beliefs to correct potential misunderstandings. Understanding the roots and consequences of why people care about what others believe, and whether they take actions to influence those beliefs is particularly important in the age of information, in which we spend an increasingly larger share of our lives creating, spreading, and consuming information.

I.2 Overview of dissertation

In Chapter II,[3] I challenge the idea that people have an intrinsic distaste for encountering differences in beliefs. Instead, I argue that people do not find others' beliefs disturbing because these are different from their own beliefs, but because they are convinced that others hold false beliefs. In a set of studies featuring self-recalled personal experiences and vignette scenarios, I demonstrate that participants are more disturbed and express stronger negative feelings in general, when they think that others hold false beliefs, compared to cases in which others' beliefs are different, even when participants' objective knowledge about others' beliefs is held constant. These findings highlight the possibility that most of the effects and behaviors that have been previously attributed to belief homophily, i.e., the desire to reduce differences one's own and others' beliefs, have been misattributed, and should be attributed to the intrinsic desire to avoid others holding false beliefs instead. The difference between these two interpretations, although subtle, is consequential: for example, it can inform how we should design policies aimed to improve the quality of public discourse. If it is the perceived incorrectness, not the difference, of beliefs that fuels the conflicts between opposing sides, then it is a better approach to highlight the possibility that others might be right, rather than trying to bridge differences between beliefs.

In Chapter III,[4] I investigate the desire for revealing contextual information to a partner when such information clarifies or corrects the partner's beliefs about the interaction but has no further consequences to the decision-maker (i.e., does not affect outcomes). I focus on situations in which making the right choice involves selecting the "lesser of two evils," however, only seeing the chosen option can lead others to misunderstand the decision maker's intentions. I show that in these difficult situations, people are willing to sacrifice money in order to explain their choices (i.e., revealing the context of their choice), even after a one-shot, anonymous interaction, when, without such costly communication of information, their partner would be likely to misunderstand their intent and thus form incorrect beliefs about the interaction. These findings highlight the role of intrinsic preferences for what others believe, even when there are no obvious instrumental reasons to care about these beliefs.

In Chapter IV,[5] I argue that people care directly about what others believe in the context of moral behavior and punishment—that people care about what transgressors think about punishment they receive, and specifically that they understand they have been punished, and why. I show that people who would otherwise enact harsh punishments, are willing to compromise and punish less severely, if by doing so they can tell the transgressor why

[3] Co-authored with George Loewenstein (Molnar & Loewenstein, 2020).
[4] Co-authored with Shereen Chaudhry (Molnar & Chaudhry, 2020).
[5] Co-authored with Shereen Chaudhry & George Loewenstein (Molnar, Chaudhry, & Loewenstein, 2020).

they are punishing them. And this is not because they want to deter bad behavior in the future—punishers glean value directly from affecting the beliefs of the transgressors in this way. Thus, people punish others not only because they want to restore justice (reduce the transgressor's welfare), but also because they want to affect what transgressors believe. This chapter also demonstrates that people care about what transgressors think about the punishment they receive—that they understand they have been punished, and why, though, according to our studies, not necessarily by whom. This insight has practical implications, for example, in the legal system: if victims are provided alternative ways of communicating messages to transgressors, via, for instance, mediation or victim impacts statements, they may be willing to settle legal suits for lower amounts or ask for lower punitive damages.

Finally, I conclude my dissertation by discussing the limitations of the present work in Chapter V. In addition, I highlight some future avenues for research, as well as review three areas of potential applications, in which gaining a deeper understanding of the causes and consequences of caring about others' beliefs might prove to be essential: 1) political behavior; 2) mental health; and 3) the future of work. Taking into account an intrinsic preference over others'mental states can help us to better understand a plethora of contemporary societal issues: the polarization of political beliefs and belief-based geographic sorting; the dramatic deterioration of public trust in democratic institutions and the media (and the emergence of "fake news"); the psychological effects of automatization and human-computer interactions; and the worsening mental health conditions due to people feeling misunderstood, isolated, and "left behind" by their peers and society, which might contribute to the recent surge in anti-establishment and extremist sentiments.

My work also expands our understanding of belief-based preferences to include preferences over second-order beliefs, i.e., the beliefs of others, that go beyond image motivation and strategic considerations (e.g., Battigalli et al., 2019). While previous theories of belief-based motives have examined individuals' preferences over *their own* non-instrumental cognitive states (e.g., Golman et al., 2017; Golman, Loewenstein, Molnar, & Saccardo, 2019; Hertwig & Engel, 2016), such as knowing whether one has an untreatable illness (Ganguly & Tasoff, 2017; Oster et al., 2013), or whether one's stock portfolio has decreased in value (Karlsson, Loewenstein, & Seppi, 2009; Sicherman et al., 2016), my dissertation demonstrates that such belief-based motives can be extended beyond the individual. That is, I argue that people also have an instrinsic preference for what others (should) believe, and this desire has important implications to their well-being and behavior in a multitude of domains.

This page intentionally left blank

"That was excellently observed, say I, when I read a passage in an author, where his opinion agrees with mine. When we differ, there I pronounce him to be mistaken."

— *Jonathan Swift* (1801)

Chapter II

The False and the Furious

With the dramatic increase in political polarization in the United States (Frimer, Skitka, & Motyl, 2017; Iyengar & Westwood, 2015; Mitchell, Gottfried, Kiley, & Matsa, 2014) and worldwide (Bessi et al., 2016; McCoy, Rahman, & Somer, 2018) there has been growing academic interest in the phenomena of belief homophily and belief dissonance: the tendency to associate with others who hold similar beliefs, and the distaste for encountering differences in beliefs.[1] Research documents the consequences of belief homophily (and belief dissonance), including geographic segregation by political views (Bishop, 2009; Motyl, Iyer, Oishi, Trawalter, & Nosek, 2014), political polarization and the spread of misinformation (Bessi et al., 2016), and selective self-exposure to media presenting concordant perspectives (Bakshy, Messing, & Adamic, 2015; Del Vicario et al., 2016; Flaxman, Goel, & Rao, 2016), creating what Cass Sunstein (2001) referred to as an "echo chamber."

A range of possible mechanisms have been proposed for why people don't like encountering opposing beliefs, including challenges to identity (Akerlof & Kranton, 2000; Bénabou & Tirole, 2011; Wood, Pool, Leck, & Purvis, 1996); aversion to cognitive dissonance and negative emotions (Frimer et al., 2017; Dorison, Minson, & Rogers, 2019); the desire for harmonious interactions (Kahan, 2015; Van Boven, Ehret, & Sherman, 2018); and the fear of being forced to "write off" investments based on existing beliefs (Abelson, 1986; Golman et al., 2016). In this chapter, however, we question whether social scientists have been correct in identifying belief-dissonance as the primary phenomenon that produces these effects, and whether an alternative theory could do a better job explaining these attitudes and behaviors. Our findings highlight the possibility that the behaviors that have been previously attributed to belief homophily, have been misattributed, and should be attributed to the desire to avoid others holding false beliefs instead. The difference between these interpretations is consequential and can inform policies aimed to improve the quality of public discourse.

[1] Following the definition of beliefs introduced in Chapter I, we limit our discussion to *beliefs* that are subjective assessments about verifiable and factual states of the world. We do not discuss disagreements over non-verifiable, non-factual statements, i.e., subjective *preferences* or *taste*.

II.1 The distaste for others holding false beliefs

Most adults can effortlessly infer others' beliefs and mental states, even if these drastically differ from their own—this astounding ability is usually referred to as theory of mind, mind reading, or mentalizing (see Frith & Frith, 2003; Leslie, 1987; Premack & Woodruff, 1978). Even most children by the age of 7 understand that others can have false beliefs (Perner & Wimmer, 1985). Here, we argue that this remarkable ability can backfire, making people feel miserable when confronting others who hold incorrect beliefs. Furthermore, we propose that many—if not most—of the behaviors that have been previously attributed to belief homophily have been misattributed, and in fact arise from the powerful distaste for encountering others who hold false beliefs. Testing between these alternatives of when and why people find others' beliefs disturbing—the *different belief* and *false belief* account—might seem trivial, but it is in fact challenging. The problem is that, if someone else has a different belief than oneself, given that a belief is by definition something that one believes to be true (see naïve realism: Pronin, Gilovich, & Ross, 2004; Reeder, Pryor, Wohl, & Griswell, 2005; Ross & Ward, 1996), it naturally also follows that the other person has a false belief. If this were not the case, that would be a sign either that one is not confident in one's own belief (so it is not truly a belief) or that one does not know for sure what the other person believes (in which case one cannot be certain that their belief is different). Yet, despite these empirical challenges, we believe that the distinction between these two accounts is not only meaningful from a psychological perspective but is also testable. It is consequential as well, because the two theories point to different prescriptions for reducing polarization and improving the quality of public discourse. For instance, if it is the perceived incorrectness, not differences, that fuels conflicts, then it is a better approach to highlight that others might be right (e.g., by shaking their confidence in their own beliefs, Babcock, Loewenstein, & Issacharoff, 1997), rather than trying to bridge differences (e.g., alerting people that differences in beliefs aren't as large as they think, Van Boven, Judd, & Sherman, 2012).

What makes false beliefs disturbing?

Subjective confidence. Truth is typically seen as desired, and people pursue accurate knowledge (e.g., theories of self-assessment, Sedikides, 1993; W. B. Swann, 1984). Based on the inherent desirability of truth, we should expect that *being convinced* that another party has incorrect beliefs should be disturbing; in contrast, *merely suspecting* that someone is acting upon false beliefs should be much less unnerving. By contrast, we would not expect a similar effect of confidence that beliefs differ: There is no comparable reason why being confident that another person's beliefs are different from one's own would be more disturbing than simply suspecting this.

Negative consequences. Another source of unsettling feelings from another person's false beliefs may arise from the prospect of negative consequences of the actions the other person takes based on those beliefs. We should expect to observe much milder frustration when another person is holding an incorrect, but relatively innocuous belief (e.g., believing that New York is the capital of the U.S., and as a result, answering incorrectly in a bar trivia contest), but much greater frustration if the setting was a televised game show with huge stakes (e.g., incorrectly believing that New York is the capital of the U.S., and losing $\$100,000$ in a game show).

Who is affected. Naturally, not only what those consequences are, but who they happen to, matters: Believing that other people are holding, and acting on, false beliefs is not always painful, even if the potential consequences are as severe as they could get. For example, if one detested the game show contestant, the fact that they lost as a result of their false belief might actually be pleasurable. However, it is painful when these false beliefs affect someone who the individual cares about: It would be heartbreaking to witness one's best friend or one's child giving the wrong answer.

II.2 Overview of studies

We gathered empirical evidence for the main hypothesis—that people are more disturbed by others' false beliefs than by different beliefs—in five pre-registered studies ($N = 2,835$). In Studies 1 and 2 we asked participants to recall real life experiences—situations in which they encountered others with different or incorrect beliefs, and measured how disturbed they were in these situations. We hypothesized that people who were asked to recall a situation in which the other person had incorrect beliefs would associate the situation with stronger negative emotions (and weaker positive emotions) than people who were asked to recall a situation in which the other person merely had different beliefs from the participant. While Study 1 provides a proof-of-concept, in Study 2 we also measured the effect of potential negative consequences, as well as participants' subjective confidence that others have false or different beliefs, in order to investigate the mechanisms which make these encounters particularly frustrating. We found consistent support for the main hypothesis in Studies 1–2, however, these studies had the limitation of relying on self-recalled stories, which could have introduced potential confounds due to the unobserved differences in recalled events (e.g., variation in the consequences or levels of confidence that others hold false beliefs). To address this limitation, in Study 3 we moved away from self-recalled real-world stories, and presented participants with standardized hypothetical vignettes. This allowed us to hold most aspects

of the situations constant (e.g., content of beliefs, social proximity, negative consequences), while still being able to experimentally manipulate whether participants construe the scenario in terms of incorrect or different beliefs. We achieved this by presenting participants with scenarios in which their confidently held beliefs differed from those of another person. Then, in different experimental conditions, we highlighted the element of the scenario that each of the theories sees as key: *differences in beliefs* or the *incorrectness of the other person's beliefs*. Finally, participants reported how disturbed they would have felt in these situations. Study 3 not only conceptually replicated the findings of Studies 1–2, but also provided evidence that being more confident about others holding incorrect beliefs—but not different beliefs—is associated with stronger negative reactions, providing further support for our main hypothesis.

In Study 4 we tested the hypothesis that false beliefs are primarily disturbing if they directly or indirectly affect someone who the participant cares about. We put participants in the same hypothetical scenarios as in Study 3, but we always highlighted that the other person had incorrect beliefs. We then manipulated, between participants, whether these false beliefs could affect someone who the participant would care about (e.g., their sibling) or someone who they wouldn't care about much (e.g., a stranger). Importantly, we held the potential consequences across conditions constant, so this study did not test how the severity of potential consequences would affect emotional reactions. Consistent with our predictions, we found that people would be more disturbed if the same false belief affected someone who they care about. Finally, in Study 5 we investigated the consequences of the distaste for others holding false beliefs: We studied whether believing that others hold incorrect beliefs would jeopardize social interactions and trigger various types of avoidance behaviors. We also increased external validity by eliciting participants' beliefs about real issues and presented them with Tweets that conflicted with their views. Consistent with Studies 1–4, we found that higher confidence that others hold false beliefs—but not different beliefs—was associated with stronger negative reactions. Moreover, being convinced that someone holds incorrect beliefs also triggered avoidance behaviors and made people less interested in establishing relationships with others.

II.3 Study 1: Self-construed recalled stories

Methods

A priori power analysis and pre-registration. We conducted an a priori power calculation in G*Power 3.1.9.2 (Faul, Erdfelder, Buchner, & Lang, 2009) to determine the sample size. To be able to detect a moderate difference (Cohen's $d \geq 0.50$) between people who classify the situation as "different beliefs" and people who classify the situation as "incorrect beliefs" in an independent samples t-test (assuming equal sample sizes), at the conventional significance level ($\alpha = .05$) and at a reasonably high power level ($1-\beta = .95$), we needed 210 observations. Based on this power calculation, we aimed to collect 200 responses, and we stopped data collection after reaching 200 responses (after exclusions). The sampling procedure and exclusion criteria, as well as the hypotheses, methods, measures, and analyses were pre-registered at AsPredicted.org: http://aspredicted.org/blind.php?x=nu9d6m.

Participants and ethics statement. We recruited 206 participants on Amazon Mechanical Turk.[2] Six (2.9%) were excluded from the data analysis, four who quit before finishing the study and two who failed the attention check. The final sample contained 200 responses (50.0% female; $M_{age} = 40.7$ years). The study was reviewed and approved by the Institutional Review Board at Carnegie Mellon University and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020). They were instructed to "recall a recent instance in which you [i.e., the participant] were aware that a person or group of people believed one thing and you [i.e., the participant] believed something else." We chose this wording that encouraged participants to think in terms of differences between their own and the other's beliefs rather than the other's false beliefs, as a conservative test of our hypotheses. Then, participants described who this other person (or group of people) was, and described this situation in as much detail as possible (open-ended response). Next, participants indicated the extent to which they felt the following three positive and three negative emotions: calm, happy, relaxed, disturbed, frustrated, and upset. The six emotions were presented in a random order. Participants provided their responses on continuous scales from 0 (*not at all*) to 100 (*extremely*). After rating each of the six emotions, participants responded to the following question: "Which of the following statements better describes the way YOU personally think about this situation?" Participants had to select one of the following (presented in a random order): (1) "The other

[2]The eligibility criteria were the following: located in the U.S., 99% or higher approval rating, and completed at least 5,000 HITs. Participants could complete the study only once.

person (or group of people) had different beliefs than me." or (2) "The other person (or group of people) had incorrect beliefs." We used this question to assess whether participants spontaneously classified the recalled situations in terms of others holding *different beliefs*, or *incorrect beliefs*, which allowed us to compare the emotional responses between these groups. Importantly, participants had to indicate their emotions *before* they were asked to construe the situation in terms of different or incorrect beliefs, which ensures that they could not draw inferences about their emotions from their own self-reported construal. On the next screen, participants indicated how confident they were that: 1) the other person had incorrect beliefs in the recalled situation; 2) the other person had different beliefs. The order of these two questions was randomized. Participants provided their responses to these questions on continuous scales from 0 (*not confident at all*) to 100 (*very confident*). Then, participants reported if there were any negative consequences of the other person holding those particular beliefs (open-ended response), and how often they experienced situations like the recalled one (5-point Likert scale: (*never, rarely, sometimes, often, very often*). Finally, participants answered an attention check question and indicated their age and sex.

Results

Main results. All 200 participants recalled a personal situation in which another person or a group of people held different and/or incorrect beliefs. The majority of participants ($n = 133$, 67%) reported that they are involved in similar situations at least sometimes or more frequently. Next, we looked at the proportion of people who construed the recalled situation in terms of different beliefs rather than in terms of incorrect beliefs, and vice versa. Although we hypothesized that the latter characterization would be more prevalent, we did not observe this pattern: Participants were about equally likely to classify the recalled situations in terms of different beliefs ($n = 111$, 56%) or as incorrect beliefs ($n = 89$, 44%). A Chi-square goodness of fit test revealed that these proportions are not significantly different from each other (i.e., not different from 50%), $\chi^2(1, N = 200) = 2.42, p = .120$.

More importantly, however, people who construed the recalled situation in terms of incorrect beliefs were significantly more disturbed, $M = 60.73$, than people who construed it in terms of different beliefs, $M = 41.27$, $t(191) = 4.062$, $p < .001$, Cohen's $d = 0.58$, 95% CI $[10.01, 28.91]$. Furthermore, we observed a significant difference between these groups in each of the six emotions: People who described the situation in terms of incorrect beliefs were also significantly more frustrated ($p < .001$), more upset ($p = .019$), less calm ($p = .026$), less happy ($p < .001$), and less relaxed ($p = .005$) than people who described the situation in terms of different beliefs (Figure 1 and Table 1). This suggests that the results obtained are not specific to using one particular label (i.e., "disturbed").

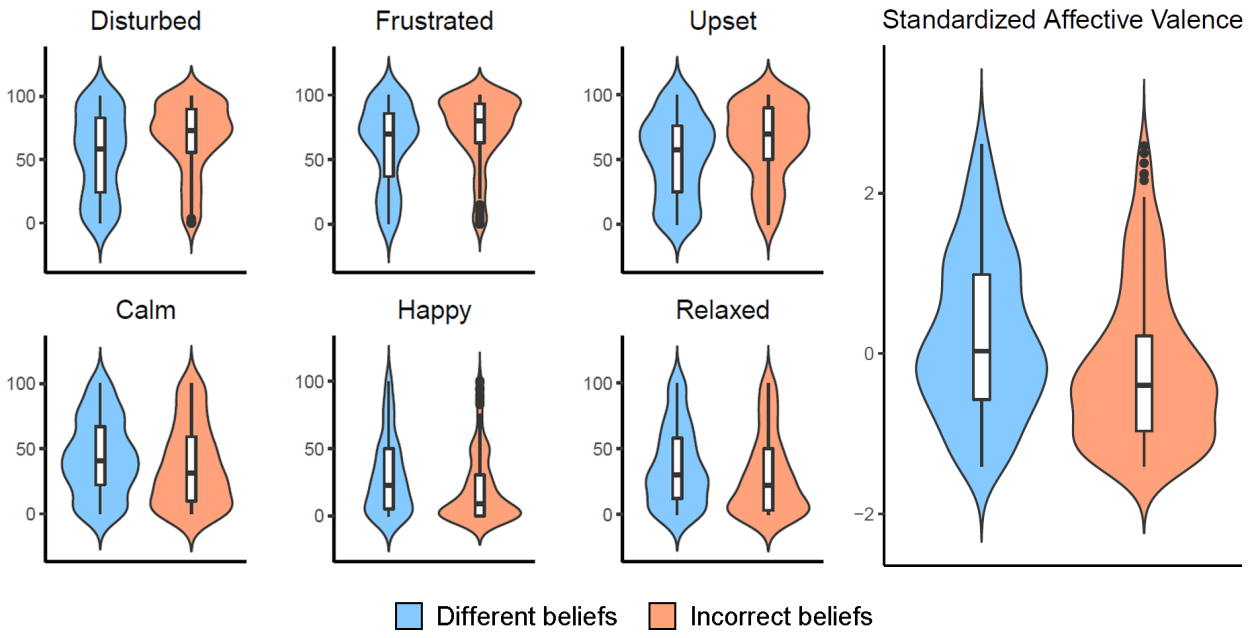


Figure 1: Violin plots of the six individual emotions (disturbed, frustrated, upset, calm, happy, and relaxed) and the compound measure of standardized affective valence in self-recalled real life situations (Study 1). Participants reported systematically stronger negative (and weaker positive) emotions when they construed the recalled situation in terms of incorrect beliefs than when they construed the situation in terms of different beliefs. The white boxes indicate interquartile ranges (IQR). The horizontal black line within each IQR indicates the median. We report the detailed *t*-test statistics in Table 1.

Table 1: Mean emotion ratings, Study 1

	Described situation in terms of...		
Emotion	different beliefs M [95% CI]	incorrect beliefs M [95% CI]	Test statistics
Disturbed	41.27 $[34.92, 47.62]$	60.73 $[53.81, 67.65]$	$t(191) = 4.062$, $p < .001$ Cohen's $d = 0.58$
Upset	43.20 $[37.05, 49.35]$	54.40 $[47.46, 61.35]$	$t(188) = 2.367$, $p = .019$ Cohen's $d = 0.34$
Frustrated	54.98 $[48.47, 61.49]$	73.82 $[67.93, 79.71]$	$t(198) = 4.205$, $p < .001$ Cohen's $d = 0.58$
Calm	48.68 $[42.7, 54.67]$	38.47 $[31.89, 45.05]$	$t(190) = 2.251$, $p = .026$ Cohen's $d = 0.32$
Happy	35.98 $[30.34, 41.62]$	18.82 $[14.15, 23.49]$	$t(197) = 4.593$, $p < .001$ Cohen's $d = 0.63$
Relaxed	42.35 $[36.72, 47.98]$	30.13 $[23.89, 36.38]$	$t(189) = 2.848$, $p = .005$ Cohen's $d = 0.40$
Affective valence	0.252 $[0.06, 0.44]$	-0.314 $[-0.50, -0.13]$	$t(196) = 4.188$, $p < .001$ Cohen's $d = 0.59$

Since we observed strong and significant correlations between the six emotions, all $|r| > .455$, all $p < .001$, which suggested reasonable factorability, we conducted a principal components analysis. The initial eigenvalues indicated that the first component alone explained 67% of the variance. The second, third, fourth, fifth, and sixth components had eigenvalues below one, and explained only 12%, 8%, 6%, 4%, and 3% of the variance, respectively. The component matrix indicated that the first component corresponds to affective valence, in which the three positive emotions had positive loadings, while the three negative emotions had negative loadings (see Tables 13–15 in Appendix VII.1.1). Therefore, we created a new compound variable—*affective valence*—which is the standardized, arithmetic mean of the standardized ratings of the six emotions (the ratings of the negative emotions were multiplied by -1). We also observed a significant difference in this compound measure of affective valence between the group of people who described the situation in terms of incorrect beliefs and the group who described it in terms of different beliefs: The incorrect beliefs group reported significantly lower affective valence, $M = -0.31$, than the different beliefs group, $M = 0.25$, $t(196) = 4.188$, $p < .001$, Cohen's $d = 0.59$, 95% CI $[0.30, 0.83]$, see Figure 2a.

Additional analyses: Confidence ratings and demographic factors. Next, we analyzed participants' subjective confidence ratings that others had different, or incorrect, beliefs. Participants who described the recalled situation in terms of different beliefs were significantly more confident that the other person had different beliefs, $M = 94.01$, than people who described the situation in terms of incorrect beliefs, $M = 80.93$, $t(103) = 3.856$, $p < .001$, Cohen's $d = .60$, 95% CI $[6.35, 19.80]$. On the other hand, the latter group was significantly more confident that the other person had incorrect beliefs, $M = 94.00$, than the former group, $M = 52.49$, $t(127) = 12.860$, $p < .001$, Cohen's $d = 1.66$, 95% CI $[35.13, 47.90]$.

Finally, we conducted an OLS linear regression analysis to test whether these confidence ratings affected how disturbed participants felt in the recalled situations (regardless of whether they construed the events in terms of incorrect or different beliefs). This regression analysis revealed that people had significantly lower affective valence when they were more confident that the other person had incorrect beliefs, $\beta = -0.011$, $SE = 0.002$, $t(195) = 5.359$, $p < .001$ (Table 2).[3] However, higher confidence that the other person had different beliefs did not affect discomfort ratings in any way, $p = .364$. While both age and sex were significant predictors of affective valence, more importantly, including these do not change the main results (Model 2). These results hold even if we exclude those participants who were fully confident that the other person had different beliefs (Models 3–4).

[3]Coefficient estimates and corresponding test statistics reported in the main text are extracted from Model 2 (Column 2), which has the highest adjusted R^2.

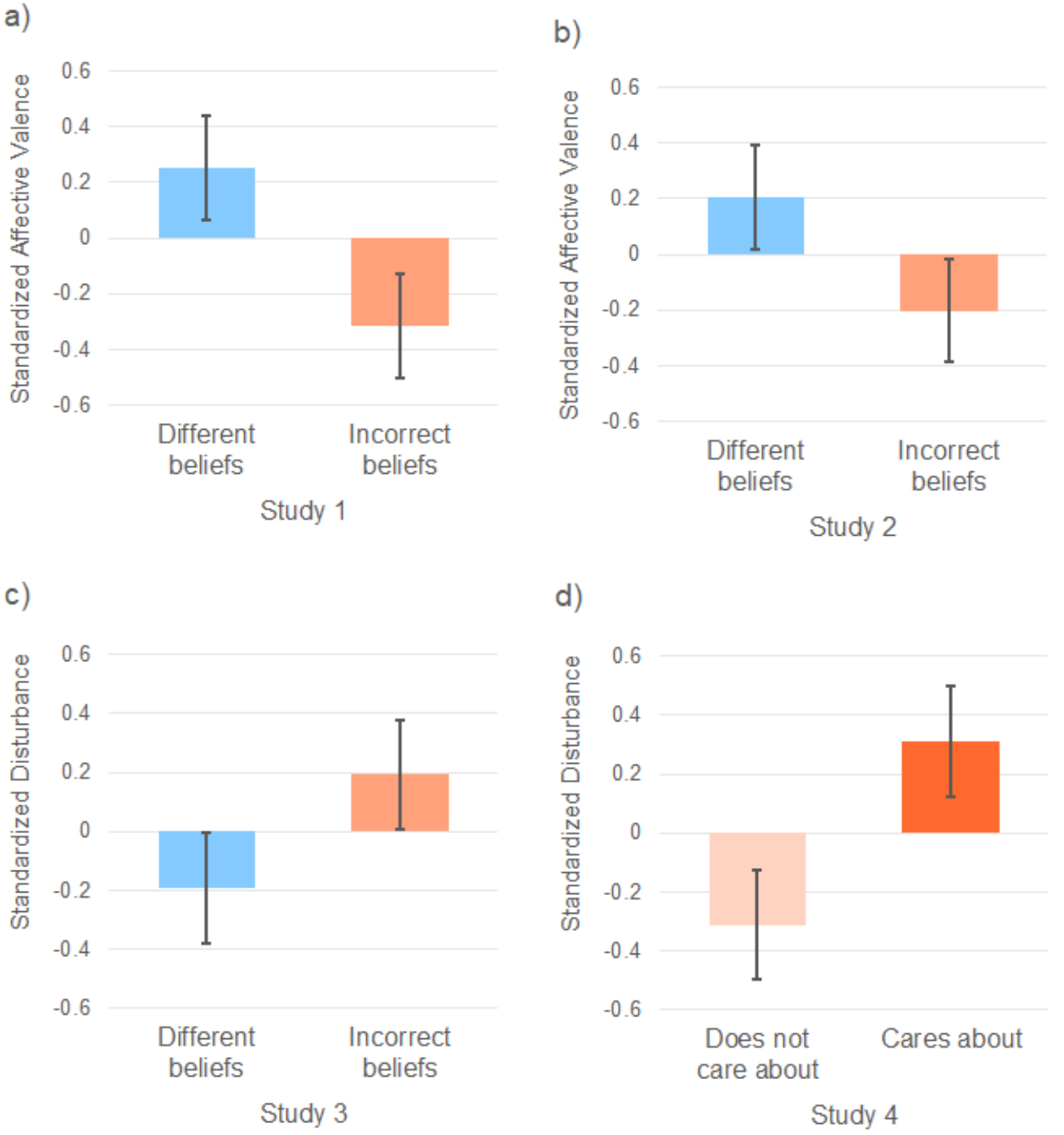


Figure 2: Main results in Studies 1–4. Note that in Studies 1 and 2 the dependent measure is the standardized affective valence (i.e., lower values mean stronger negative affect), whereas in Studies 3 and 4 the dependent measure is the standardized disturbance (i.e., higher values mean stronger negative affect). All of the differences depicted above (i.e., between conditions within each study) are significant at $p < .001$. Error bars represent 95% Confidence Intervals.

Table 2: OLS regression results, Study 1

	Standardized Affective Valence			
	Full sample		Limited sample[1]	
	(1)	(2)	(3)	(4)
Confidence: other incorrect (0–100)	-0.010^{***}	-0.011^{***}	-0.007^{*}	-0.007^{*}
	(0.002)	(0.002)	(0.003)	(0.003)
Confidence: different beliefs (0–100)	0.003	0.003	0.001	0.001
	(0.003)	(0.003)	(0.003)	(0.003)
Sex (female = 1)		-0.485^{***}		$-0.353^{\dagger}$
		(0.132)		(0.179)
Age (years)		0.013^{*}		0.007
		(0.006)		(0.008)
Constant	0.496	0.281	0.280	0.225
	(0.327)	(0.365)	(0.396)	(0.444)
Observations	200	200	94	94
R^2	0.118	0.185	0.050	0.093
Adjusted R^2	0.109	0.168	0.029	0.053

Note: $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

[1] *Only those participants who were not fully confident (< 100) that beliefs were different.*

Discussion

Study 1 provides preliminary evidence that people experience stronger negative affect when they encounter someone who holds beliefs that they deem incorrect, as opposed to others whose beliefs are different from their own. However, this study has a major shortcoming: Participants were not assigned randomly to recall a situation in which they encountered false or different beliefs; rather, they construed the recalled stories themselves, which might have introduced potential confounds. For instance, if people who tend to have stronger emotional responses in general were more likely to construe situations in terms of incorrect beliefs than people who tend to have milder reactions, then the differences in reported affective ratings could be explained by differences in participants' inherent characteristics (strength of emotional reactions), instead of the nature of beliefs involved in the recalled stories. We address this limitation in Study 2, by randomly assigning participants to either recall a situation that involved incorrect or different beliefs. Furthermore, we measure participants' subjective confidence in the correctness of their *own beliefs*, and also record some other features of the interaction that might explain what makes others' beliefs particularly disturbing.

II.4 Study 2: Guided self-recalled stories

Methods

A priori power analysis and pre-registration. We conducted an a priori power calculation in G*Power 3.1.9.2 (Faul et al., 2009) to determine the sample size. To be able to detect a moderate difference (Cohen's $d \geq 0.34$)[4] between people who recall a situation in which the other person held "different beliefs" and people who recall a situation in which the other person held "incorrect beliefs" in an independent samples t-test (assuming equal sample sizes), at the conventional significance level ($\alpha = .05$) and at a reasonably high power level ($1 - \beta = .95$), we needed 226 observations per condition. Based on this power calculation, we aimed to collect 200 responses per condition (400 total), and we stopped data collection after reaching 200 responses per condition (after exclusions). The sampling procedure and exclusion criteria, as well as the hypotheses, methods, measures, and analyses were pre-registered at AsPredicted.org: http://aspredicted.org/blind.php?x=wh56w8.

Participants and ethics statement. We recruited 498 participants on Amazon Mechanical Turk.[5] One hundred participants (20%) were excluded from the data analysis: 65 (13%) who quit before finishing the study,[6] 24 (5%) who failed the attention check, and 11 (2%) who submitted duplicate responses. The final sample contained 398 responses (49.3% female; $M_{age} = 38.8$ years). The study was reviewed and approved by the Institutional Review Board at Carnegie Mellon University and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020), in which they were randomly assigned to either the *incorrect beliefs* or the *different beliefs* condition: Participants were instructed to recall a recent instance in which they were aware that a person or group of people had either incorrect beliefs, or beliefs different from their own. Then, they described this situation in as much detail as possible (open-ended response). Following the recall, participants indicated the extent to which they experienced the same six feelings as in Study 1—calm, happy, relaxed, disturbed, frustrated, and upset, presented in a random order. Then, they indicated how confident they were that: 1) the other person had incorrect beliefs; 2) the other person had different beliefs; and 3) they (i.e. the participant) had correct beliefs in the recalled situation.

[4] This was the weakest effect we observed among the three negative emotions in Study 1.

[5] The eligibility criteria were the following: located in the U.S., 99% or higher approval rating, and completed at least 5,000 HITs. Participants could complete the study only once (however, a few participants attempted to submit duplicate responses).

[6] Among these participants, 54 quit before answering any questions, including the prompt.

Next, we asked participants how close this other person was to them at the time when the recalled events happened (from 0 = *not at all* to 100 = *very*), and whether there were any immediate or long-term negative consequences of the other person holding their beliefs (*yes, no, n/a*). If they answered "yes," we asked them to describe the negative consequences and rate how serious they were (from 0 = *very mild* to 100 = *very serious*). Then, we asked participants to indicate for each of the following categories whether they describe the beliefs involved in the recalled situation: political beliefs, religious beliefs, scientific beliefs, financial beliefs, beliefs about other people, beliefs about health and wellness, beliefs about sports, beliefs about work, and beliefs about products (from 0 = *not at all* to 100 = *very well*). Finally, we recorded the age and sex of participants, and as a manipulation check, participants answered the same self-construal question that we used in Study 1: "Which of the following statements better describes the way YOU personally think about this situation?" Participants had to select one of the following: "The other person (or group of people) had different beliefs than me." or "The other person (or group of people) had incorrect beliefs." We included this manipulation check at the end of the survey, so it could not interfere with any of the other dependent measures.

Results

Manipulation check. The majority of participants, 74%, construed the recalled situation in terms consistent with the experimental manipulation: Out of 200 participants who were asked to recall a situation in which someone held incorrect beliefs, 170 (85%) construed this interaction in terms of incorrect beliefs (rather than different beliefs), and out of the 198 participants who were asked to recall a situation in which someone held different beliefs, 125 (63%) construed the interaction in terms of different beliefs (rather than incorrect beliefs). The analyses reported in the following sections are all *intention-to-treat*, that is, we compared participants' responses between experimental conditions, regardless of how they construed the recalled situation. Since there was a substantial minority of participants (26%) who did not comply with the instructions, intention-to-treat analyses yield a more conservative estimate of the true effect of the experimental manipulation.

Main results. Participants who were instructed to recall a situation in which the other person held incorrect beliefs were significantly more disturbed, $M = 67.70$, than people who recalled a situation in which the other person held different beliefs, $M = 54.00$, $t(385) = 4.425$, $p < .001$, Cohen's $d = 0.44$, 95% CI $[7.61, 19.79]$. Furthermore, we observed a significant difference between these groups in each of the six emotions that we measured: People who recalled situations involving incorrect beliefs were significantly more frustrated

($p < .001$), more upset ($p < .001$), less calm ($p = .029$), less happy ($p < .001$), and less relaxed ($p = .025$) than people who recalled situations involving different beliefs (Figure 3 and Table 16).

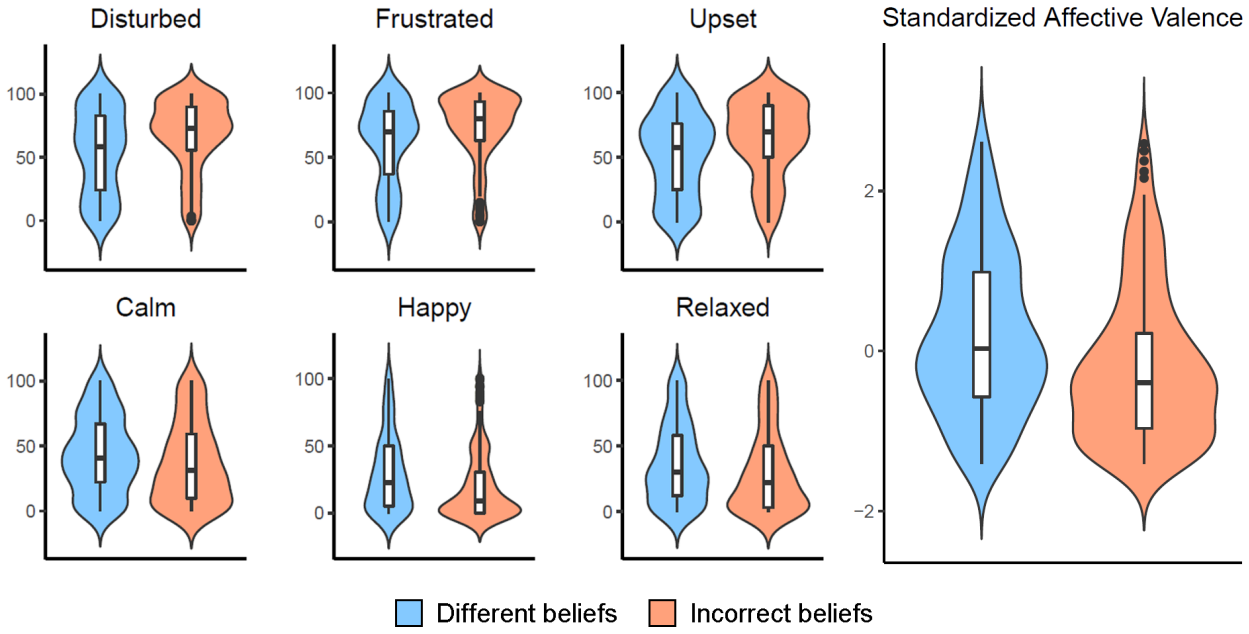


Figure 3: Violin plots of the six individual emotions (disturbed, frustrated, upset, calm, happy, and relaxed) and the compound measure of standardized affective valence in self-recalled real life situations (Study 2). Participants reported systematically stronger negative (and weaker positive) emotions when they recalled situations in which the other person held incorrect beliefs, as opposed to situations in which other held different beliefs. The white boxes indicate interquartile ranges (IQR). The horizontal black line within each IQR indicates the median. We report the detailed *t*-test statistics in Table 16 in Appendix VII.1.2.

As in Study 1, we observed strong positive correlations between all six emotions, all $p < .001$, so we conducted a principal component analysis. This analysis indicated that all emotions load into a single component which can be interpreted as affective valence (see Tables 17–19 in Appendix VII.1.3). We created a compound measure of affective valence, which is the standardized, arithmetic mean of the standardized ratings of the six emotions (the ratings of the three negative emotions were multiplied by -1). People who recalled a situation involving incorrect beliefs had a significantly lower affective valence, $M = -0.20$, than people who recalled a situation involving different beliefs, $M = 0.20$, $t(395) = 4.127$, $p < .001$, Cohen's $d = 0.41$, 95% CI $[0.21, 0.60]$, see Figure 2b. Next, we analyzed participants' subjective confidence that they (i.e., the participant) had correct beliefs, and that others had different, or incorrect, beliefs. Participants who recalled a situation in which the other person held incorrect beliefs were more confident that they (i.e., the participant) had correct

beliefs, $M = 91.06$, than participants who recalled a situation in which the other person held different beliefs, $M = 84.20$, $t(369) = 3.679$, $p < .001$, Cohen's $d = 0.37$, 95% CI $[3.19, 10.53]$. Participants recalling incorrect beliefs were also more confident that the other person had incorrect beliefs, $M = 89.20$, than people who recalled different beliefs, $M = 78.96$, $t(390) = 4.196$, $p < .001$, Cohen's $d = 0.42$, 95% CI $[5.44, 15.04]$. However, we did not find a significant difference in the confidence rating that the other person had different beliefs: People who recalled incorrect beliefs were just as confident that the other person had different beliefs, $M = 91.01$, as people who recalled different beliefs, $M = 93.05$, $t(363) = 1.193$, $p = .234$, Cohen's $d = 0.12$, 95% CI $[-1.32, 5.39]$. Finally, we compared the presence or absence of negative consequences between the two conditions. Participants who recalled situations involving incorrect beliefs were significantly more likely to indicate that there were immediate or long-term negative consequences of the other person holding those particular beliefs, $n = 74$ (37%), than people who recalled stories involving different beliefs, $n = 35$ (17%), $\chi^2(1, N = 299) = 28.294, p < .001$.

Robustness checks. To investigate the combined effects of these confidence ratings, as well other potential predictors, such as demographic variables, closeness between the participant and the other person, and whether there were negative consequences of the belief held by the other person, we conducted a regression analysis (Table 3). This analysis revealed that the confidence that the other person held incorrect beliefs had a significant and negative effect on affective valence, $\beta = -0.005$, $SE = 0.002$, $t(393) = 2.308$, $p = .022$, even when controlling for the experimental condition. Neither of the other two confidence ratings predicted affective valence significantly in any of the models, all $p > .10$ (Table 3, Models 2–4). Closeness between the participant and the other person had a significant positive effect on affective valence, $\beta = 0.003$, $SE = 0.001$, $t(390) = 2.108$, $p = .036$.[7] Consistent with the results of Study 1, women reported significantly lower affective valence than men, $\beta = -0.296$, $SE = 0.094$, $t(390) = 3.140$, $p = .002$. More importantly, including closeness and demographic variables in the analyses did not change the main results (Table 3, Model 3). The presence of negative consequences not only had a significant negative effect on affective valence, $\beta = -0.647$, $SE = 0.181$, $t(289) = 3.580$, $p < .001$, but including this predictor in the regression eliminated the main effect of experimental condition, $p = .450$. There was no significant interaction between experimental condition and the presence of negative consequences, $p = .938$ (Table 3, Model 4).

[7] We report the coefficient estimates and the corresponding test statistics obtained in Model 4 (Table 3, Column 4), which had the highest adjusted R^2 among the four models.

Table 3: OLS regression results, Study 2

	Dependent variable:			
	Standardized Affective Valence			
	(1)	(2)	(3)	(4)
Condition: incorrect beliefs	-0.406^{***}	-0.328^{**}	-0.302^{**}	-0.110
	(0.098)	(0.100)	(0.100)	(0.146)
Confidence: self correct (0–100)		-0.004	-0.005	-0.005
		(0.003)	(0.003)	(0.003)
Confidence: other incorrect (0–100)		-0.005^{*}	-0.005^{*}	$-0.005^{\dagger}$
		(0.002)	(0.002)	(0.003)
Confidence: different beliefs (0–100)		-0.001	-0.001	-0.001
		(0.003)	(0.003)	(0.003)
Closeness (0–100)			0.003^{*}	0.003^{*}
			(0.001)	(0.002)
Sex (female = 1)			-0.296^{**}	-0.274^{**}
			(0.094)	(0.105)
Age (years)			0.005	$0.008^{\dagger}$
			(0.004)	(0.004)
Negative consequences? (yes = 1)				-0.647^{***}
				(0.181)
Interaction: inc. beliefs × neg. cons.				-0.019
				(0.241)
Constant	0.204^{**}	1.048^{***}	0.861^{*}	0.954^{**}
	(0.070)	(0.316)	(0.346)	(0.366)
Observations	398	398	398	299
R^2	0.041	0.074	0.107	0.225
Adjusted R^2	0.039	0.065	0.091	0.201

Note: $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

The number of observations in Models 4 is smaller than the full sample, since this model excludes people who answered "N/A" to negative consequences.

Discussion

In Study 2 we replicated the results of Study 1: People who were instructed to recall situations in which others held incorrect beliefs reported stronger negative emotions than people who recalled situations involving merely different beliefs, and that it is the confidence that others hold incorrect beliefs—not the confidence that there is a difference in beliefs—that predicts affective responses. In addition, we also showed that the confidence that one has correct beliefs matters much less than believing that others hold incorrect beliefs.

However, Study 2 still had the limitation of relying on self-recalled stories, which introduced heterogeneity in situations, and thus, potential confounds. For instance, we found that people who recalled events involving incorrect beliefs were also more likely to report that there were negative consequences of the other person holding their beliefs. Our analyses revealed that this difference alone explains most of the effect of experimental condition on reported affective valence, therefore it is not clear whether encountering others holding incorrect beliefs *per se* makes people upset, or whether it is the possibility that incorrect beliefs lead to negative consequences that evoke stronger negative reactions. In Study 3, we address this limitation by introducing standardized scenarios, which allows us to hold most aspects of the situations constant while still being able to manipulate whether participants construe the scenario in terms of incorrect or different beliefs. We achieve this by presenting participants with scenarios in which their confidently held beliefs differ from those of another person. Then, across conditions, we highlight the element of the scenario that each of the theories sees as key: *differences in beliefs* or the *incorrectness of the other person's beliefs.*

II.5 Study 3: Standardized vignette scenarios

Methods

A priori power analysis and pre-registration. We conducted an a priori power calculation in G*Power 3.1.9.2 (Faul et al., 2009) to determine the sample size. To be able to detect a moderate difference (Cohen's $d \geq 0.50$) between the two conditions in a scenario in an independent samples t-test, at the conventional significance level ($\alpha = .05$) and at a reasonably high power level ($1 - \beta = .95$), we needed 105 observations per condition in each scenario. Based on this power calculation, we aimed to collect 100 responses per condition in each scenario (800 total). We stopped data collection after reaching 100 responses per condition in each scenario (after exclusions). The sampling procedure and exclusion criteria, as well as the hypotheses, methods, measures, and analyses were pre-registered at AsPredicted.org: https://aspredicted.org/blind.php?x=cm6tt2.

Participants and ethics statement. We recruited 907 participants on Amazon Mechanical Turk.[8] We excluded 78 participants (8.6%) from the data analysis: 51 (5.6%) who quit before finishing the study, 26 (2.9%) who failed at least one attention check, and one (0.1%) who submitted a duplicate response. The final sample contained 829 responses (46.4% female; $M_{age} = 38.2$ years). The study was reviewed and approved by the Institutional Review Board at Carnegie Mellon University and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey, in which they were instructed to read a hypothetical scenario and imagine how they would feel if they were involved in that scenario. Each participant was presented one of the following four hypothetical scenarios: mayor, promotion, savings, or weight loss. We used four different scenarios to cover a range of potential topics and social interactions, and also to have natural variation across scenarios in the extent to which participants would feel disturbed by different or incorrect beliefs.

Each scenario involved two people: the participant and another person (e.g., "Jill, your neighbor"). The vignette started with a brief description of the context, then proceeded with describing the other person's belief about a particular topic (e.g., "she is convinced that the mayor is deeply corrupt"). Then, we told participants what they would believe in this scenario (e.g., "you are virtually certain that the mayor is innocent"). Thus, the vignette established that the participant's confidently held beliefs are different from the confidently held beliefs of another person, whose beliefs—from the participant's point of view—are also

[8]The eligibility criteria were the following: located in the U.S., 99% or higher approval rating, and completed at least 5,000 HITs. Participants could complete the study only once.

incorrect (we report the full vignettes in Appendix VII.1.4). On the next screen, we included an attention check question then the study proceeded to the main experimental manipulation.

Participants were randomly assigned to either the *different beliefs* condition or the *incorrect beliefs* condition. In the different beliefs condition participants were presented with a reminder that highlighted that the other person had different beliefs than them (e.g., "you know that you and Jill have different beliefs about the mayor"). In the incorrect beliefs condition participants were presented with a reminder that highlighted that the other person had incorrect beliefs (e.g., "you know that Jill has incorrect beliefs about the mayor"). Then, in both conditions, participants reported how disturbed they would be in the scenario (see Figure 4), by indicating their response on a continuous scale from 0 (*not at all*) to 100 (*very*). To keep the experiment simple, and since we found a very strong correlation between all six emotions in Studies 1–2, we decided to use a single measure in Study 3. We picked "disturbed" to maximize the statistical power of our analyses, since this measure produced the largest effect sizes in Studies 1–2.

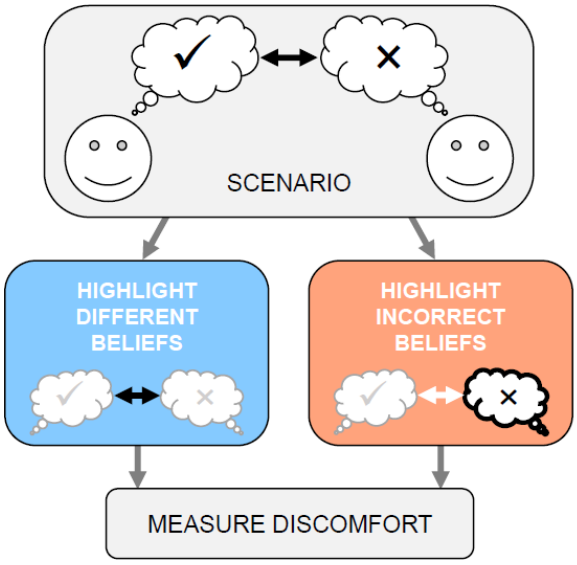


Figure 4: Flowchart of Study 3. Participants first read a scenario in which the participant's (left) beliefs were different from the beliefs of another person (right), whose beliefs—from the participant's point of view—were also incorrect. Participants were randomly assigned to either the different beliefs or the incorrect beliefs condition, in which we highlighted either the difference between beliefs (blue box) or that the other person had incorrect beliefs (orange box). Finally, we measured how disturbed participants would have felt.

Following the main dependent measure, we included the same three questions as in Study 2 as manipulation checks. We asked participants how confident they would be: 1) that they (i.e. the participant) have correct beliefs; 2) that the other person has incorrect beliefs; and 3) that the other person and them have different beliefs (presented in a random order; measured on continuous scales from 0 = *not at all* to 100 = *very*). Then, as a second attention check, participants had to indicate whether a statement about the other person's beliefs was true or false (e.g., "Jill believes that the mayor is innocent"). Finally, we recorded participants' age and sex. Participants were paid $0.30 for their participation.

Results

Manipulation checks. Although the experimental manipulation was quite subtle and each vignette made it very clear in *both* conditions that—from the participant's point of view—the other person not only had different beliefs but also had incorrect beliefs, highlighting the fact that the other person had incorrect beliefs (or different beliefs) had significant effects on confidence ratings. People reported significantly higher confidence that the other person had incorrect beliefs in the incorrect condition, $M = 77.94$, than in the different condition, $M = 73.47$, $t(822) = 2.896$, $p = .004$, Cohen's $d = 0.20$, 95% CI $[1.44, 7.50]$. Similarly, they reported significantly higher confidence that they (i.e., the participant) had correct beliefs in the incorrect condition, $M = 79.10$, than in the different condition, $M = 74.50$, $t(810) = 3.252$, $p = .001$, Cohen's $d = 0.23$, 95% CI $[1.82, 7.37]$. By contrast, we did not find any difference in participants' confidence that there was a difference in beliefs: Participants were just as confident about this in the incorrect condition, $M = 84.12$, as in the different condition, $M = 85.53$, $t(822) = 1.069$, $p = .286$, Cohen's $d = 0.07$, 95% CI $[-1.18, 4.00]$.

Main results. People in the incorrect belief condition reported being significantly more disturbed in three scenarios, $p = .010$, $p = .004$, and $p < .001$ in the mayor, promotion, weight loss scenarios, respectively, and marginally more disturbed in the savings scenario, $p = .054$, than people in the different belief condition (Figure 5 and Table 4). The effect sizes range from moderate (Cohen's $d = 0.270$) to strong ($d = 0.538$). To compare the two conditions across all scenarios, we standardized disturbance ratings within each scenario. Pooled across scenarios, participants reported that they would be significantly more disturbed in the incorrect belief condition, $M = 0.19$, than in the different belief condition, $M = -0.19$, $t(824) = 5.661$, $p < .001$, Cohen's $d = 0.39$, 95% CI $[0.25, 0.52]$, see Figure 2c.

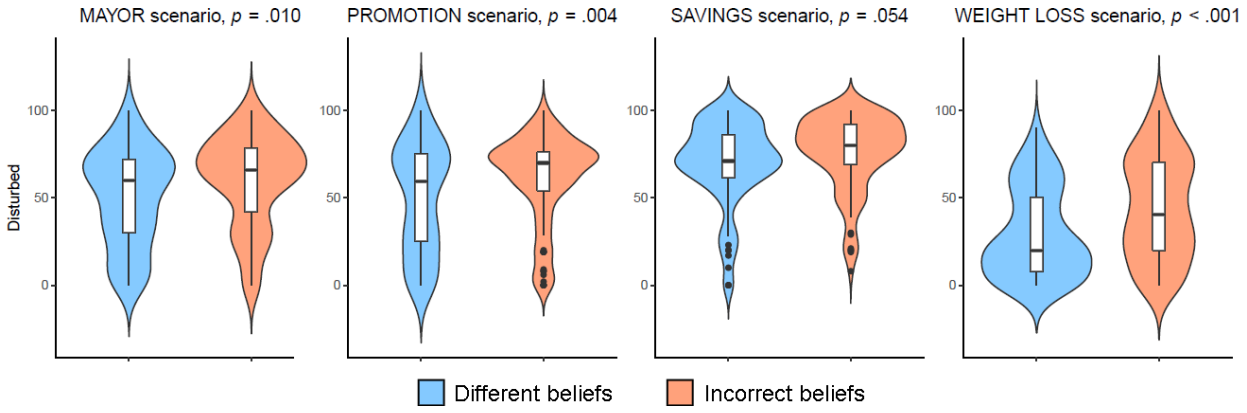


Figure 5: Violin plots of reported discomfort in Study 3. The white boxes indicate interquartile ranges (IQR). The horizontal black line within each IQR indicates the median.

Table 4: Mean disturbance ratings and test statistics in Study 3, by scenario and condition

Scenario	Condition		Test statistics[1]
	Different beliefs	Incorrect beliefs	
	M [95% CI]	M [95% CI]	
Mayor	50.38 [45.09, 55.66]	60.04 [55.03, 65.05]	$t(205) = 2.599$, $p = .010$ Cohen's $d = 0.361$
Promotion	49.81 [43.87, 55.75]	61.18 [56.33, 66.03]	$t(198) = 2.908$, $p = .004$ Cohen's $d = 0.403$
Savings	69.81 [65.21, 74.4]	75.93 [71.79, 80.08]	$t(203) = 1.940$, $p = .054$ Cohen's $d = 0.270$
Weight loss	29.81 [24.88, 34.73]	44.61 [38.97, 50.25]	$t(202) = 3.875$, $p < .001$ Cohen's $d = 0.538$
ALL (std'd)	-0.19 $[-0.29, -0.10]$	0.19 $[0.10, 0.28]$	$t(824) = 5.661$, $p < .001$ Cohen's $d = 0.390$

[1] *Pairwise comparisons between the corresponding rows (two-tailed t-tests).*

Robustness checks. To further test whether the main results are robust to the type of scenario and individual differences (i.e., demographic factors), we conducted a hierarchical OLS regression analysis (see Table 20 in Appendix VII.1.5). We also included the three self-reported confidence ratings (about the self being correct, the other person being incorrect, and that there is a difference in beliefs) as potential covariates. This regression analysis revealed that people were significantly more disturbed when incorrect beliefs were highlighted, than when different beliefs were highlighted, even after controlling for their confidences and demographic factors, $\beta = 9.27$, $SE = 1.787$, $t(819) = 5.185$, $p < .001$, (Table 20, Model 3). Importantly, higher subjective confidence that the other person held incorrect beliefs led to increased disturbance, $\beta = 0.23$, $SE = 0.050$, $t(819) = 4.508$, $p < .001$. By contrast, neither of the other two confidence ratings predicted discomfort ratings significantly, both $p > .10$. Consistent with the results of Studies 1–2, women were significantly more disturbed than men, $\beta = 3.75$, $SE = 1.783$, $t(819) = 2.102$, $p = .036$. In addition, we found a significant effect of age; older participants were more disturbed, $\beta = 0.34$, $SE = 0.071$, $t(819) = 4.737$, $p < .001$. We did not find similarly strong age effects in Studies 1–2, but this might be explained by the fact that in previous studies participants were asked to recall situations from their own life, whereas in Study 3 we provided standardized scenarios to everyone, and it is possible that these particular scenarios were more disturbing to older participants.

Mediation analysis. Finally, to investigate whether the main effect of experimental manipulation was driven by the differences in confidence ratings, we conducted a mediation analysis. We standardized the discomfort rating within each scenario, and also standardized the three measures of subjective confidences. We also controlled for sex and age in these models. A bootstrapped mediation with $5,000$ replications revealed that the confidence that the other person had incorrect beliefs significantly mediated the effect of experimental condition on reported disturbance, $\beta = 0.018$, 95% CI [0.006, 0.041], $p < .001$ (Figure 6). By contrast, neither the confidence that the participant had correct beliefs, nor the confidence that there was a difference in beliefs, mediated the effect of experimental conditions, both $p > .425$.

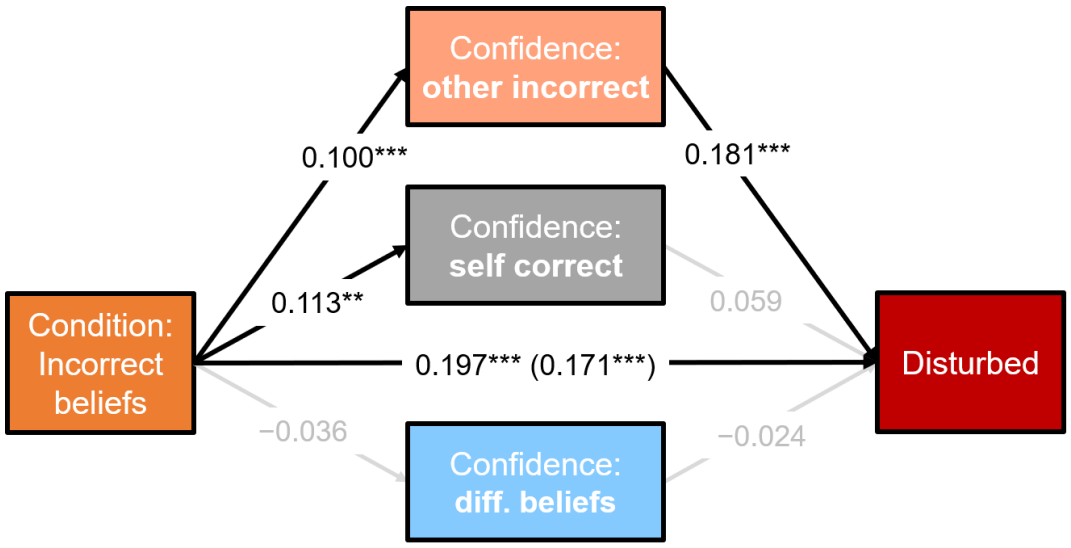


Figure 6: Mediation analysis, Study 3. Coefficients are standardized Beta coefficients.

Discussion

In Study 3 people reported that they would be more disturbed when we highlighted that another person had incorrect beliefs, as opposed to when we highlighted that they had different beliefs. Importantly, participants in both conditions read the exact same vignettes, thus, this effect cannot be explained by differences in social proximity, the presence and severity of negative consequences, or any other idiosyncratic feature of the scenarios.

II.6 Study 4: Caring about consequences

In Study 4 we test the hypothesis that false beliefs are primarily disturbing if they affect someone who the individual cares about. We put participants in the same hypothetical scenarios as in Study 3, but we always highlight that the other person has incorrect beliefs. We then manipulate, between participants, whether these false beliefs could affect someone who the participant would care about or someone who they wouldn't care about much.

Methods

A priori power analysis and pre-registration. We conducted an a priori power calculation in G*Power 3.1.9.2 (Faul et al., 2009) to determine the sample size. To be able to detect a moderate difference (Cohen's $d \geq 0.50$) between the two conditions in a scenario in an independent samples t-test, at the conventional significance level ($\alpha = .05$) and at a reasonably high power level ($1 - \beta = .95$), we needed 105 observations per condition in each scenario. Based on this power calculation, we aimed to collect 100 responses per condition in each scenario (800 total). We stopped data collection after reaching 100 responses per condition in each scenario (after exclusions). The study was pre-registered at AsPredicted.org: http://aspredicted.org/blind.php?x=ge6499.

Participants and ethics statement. We recruited 944 participants on Amazon Mechanical Turk.[9] We excluded 136 participants (14.4%) from the data analysis: 45 (4.8%) who quit before finishing the study, 90 (9.5%) who failed at least one attention check, and one duplicate response (0.1%). The final sample contained 808 responses (53.7% female; $M_{age} = 40.9$ years). The study was reviewed and approved by the Institutional Review Board at CMU and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020), in which they were instructed to read a hypothetical scenario and imagine how they would feel if they were in that scenario. Each participant was presented one of the following four hypothetical scenarios: mayor, promotion, savings, or weight loss. These were slightly modified versions of the scenarios used in Study 3, to allow for the main experimental manipulation.[10] Each scenario involved two people: the participant and another person (e.g., "James, who is a software developer"). The vignette started with a brief description of the context, then proceeded with describing the other person's belief about a particular topic (e.g., "He is

[9] The eligibility criteria were the following: located in the U.S., 99% or higher approval rating, and completed at least 5,000 HITs. Participants could complete the study only once.

[10] We report the full vignettes, as well as the attention check questions and the experimental manipulation used in these scenarios in Appendix VII.1.6.

convinced that a low-fat diet is ideal for people who want to lose weight"). Then, we told participants what they would believe in this scenario (e.g., "you are virtually certain that a low-carb, and not a low-fat diet, is ideal for people who wish to lose weight"). Thus, as in Study 3, the vignette established that the participant's confidently held beliefs are different from the confidently held beliefs of another person, and whose beliefs—from the participant's point of view—are also incorrect.

We manipulated between participants who was ultimately affected by these incorrect beliefs (thus, how much participants cared about the consequences of false beliefs). For example, James was either described as a passenger on a flight sitting next to the participant, who they have never met before (*does not care about* condition) or as the participant's only sibling, who they are really close with (*cares about* condition). Importantly, we did not manipulate the scenarios in any way that would have suggested a difference in potential consequences—only the identity of the person affected by the consequences was different.

After participants read the scenario, they answered an attention check question ("What is the name of the person in the scenario?"). On the next screen, participants were presented with a reminder that highlighted that the other person had incorrect beliefs (e.g., "you know that James has incorrect beliefs about diets."). Participants reported how disturbed they would be in the scenario, by indicating their response on a continuous scale from 0 (*not at all*) to 100 (*very*). Following the main dependent measure, participants indicated how confident they would be that the other person has incorrect beliefs ("How certain are you that [name] has incorrect beliefs?"). Participants indicated their certainty from 1 (*not at all*) to 5 (*very*). Since in Study 4 we did not manipulate whether participants were exposed to false or different beliefs, we did not ask them to indicate their confidence that beliefs were different. Then, participants answered two attention check questions: they selected who the other person was and indicated whether a statement about the other person's beliefs was true or false. Finally, we recorded participants' age and sex.

Results

Main results. People in the *cares about* condition reported being significantly more disturbed in all four scenarios than in the *does not care about* condition, all $p \leq .012$ (Figure 7). To compare the two conditions across scenarios, we standardized disturbance ratings within each scenario. Pooled across scenarios, participants reported that they would be significantly more disturbed in the *cares about* condition, $M = 0.31$, than in the *does not care about* condition, $M = -0.31$, $t(800) = 9.327$, $p < .001$, Cohen's $d = 0.66$, 95% CI $[0.49, 0.75]$ (Figure 2d). By contrast, we did not find any difference in the standardized confidence ratings between conditions: People reported that they would be about equally certain that the

other person holds incorrect beliefs in the *does not care about* condition, $M = -0.01$, as in the *cares about* condition, $M = 0.01$, $t(799) = 0.385$, $p = .701$, Cohen's $d = 0.03$, 95% CI $[-0.17, 0.11]$.

Robustness checks. Finally, we conducted an OLS regression analysis, in which we controlled for self-reported confidence that the other person had incorrect beliefs, as well as participants' age and sex. This analysis revealed that people were significantly more disturbed when they cared about the consequences of the false belief, than when they didn't, even after controlling for their certainty about the other person's incorrect beliefs and demographic factors, $\beta = 0.604$, $SE = 0.065$, $t(803) = 9.312$, $p < .001$. In addition, and consistent with previous studies, both self-reported confidence that the other person held incorrect beliefs and demographic factors are significant predictors of reported disturbance, see Table 22 in Appendix VII.1.8.

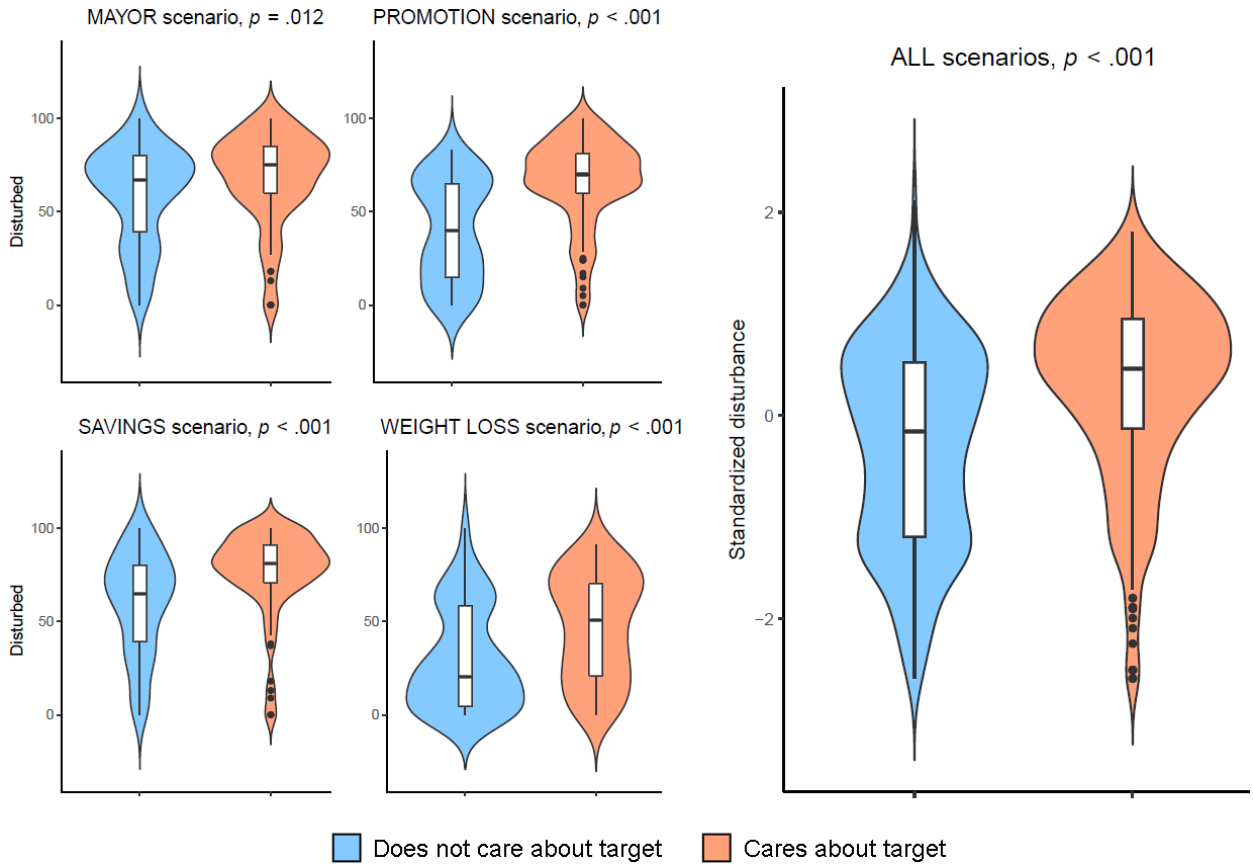


Figure 7: Violin plots of reported discomfort in Study 4. The white boxes indicate interquartile ranges (IQR). The horizontal black line within each IQR indicates the median. We report the detailed t-test statistics in Table 21 in Appendix VII.1.7.

Discussion

Study 4 demonstrates that even if we hold the content of a false belief, as well as its potential negative consequences, constant, it could evoke different levels of discomfort, depending on *who is affected* by those consequences. We also replicated the findings of Studies 1–3 that higher subjective confidence that the other person holds incorrect beliefs makes people more disturbed. While the preceding four studies established that encountering false beliefs affects emotional states negatively, Study 5 investigates whether these effects are strong enough to trigger meaningful changes in participants' behavior.

II.7 Study 5: Downstream behavioral consequences

In this final study we test whether people who are convinced that others hold false beliefs—as opposed to merely different beliefs—would be more likely to avoid interacting with those others, and would be less likely to form any kind of relationship with them. Besides investigating these downstream consequences of being disturbed by others' false beliefs, we also add more external validity by eliciting participants' actual views on real sociopolitical and scientific issues, and then present them with realistic Tweets that conflict with their beliefs.

Methods

A priori power analysis and pre-registration. We conducted an a priori power calculation in G*Power 3.1.9.2 (Faul et al., 2009) to determine the required minimum sample size. To be able to detect a significant effect with a weak/moderate effect size (standardized coefficient slope ≥ 0.15) of predictors in an OLS linear regression, at the conventional significance level ($\alpha = .05$) and at a reasonably high power level ($1 - \beta = .95$), we needed 567 observations. Based on this power calculation, we aimed to collect 600 responses. We stopped data collection after reaching 600 responses (after exclusions). The sampling procedure and exclusion criteria, as well as the hypotheses, methods, measures, and analyses were pre-registered at AsPredicted.org: https://aspredicted.org/blind.php?x=td9wh8.

Participants and ethics statement. We recruited 630 participants on Amazon Mechanical Turk.[11] We excluded 30 participants (4.8%) from the data analysis: 21 (3.3%) who failed at least one attention check, eight (1.3%) who submitted duplicate responses, and one (0.2%) who indicated "no opinion" on all six topics. The final sample contained 600 responses (48% female; $M_{age} = 39.5$ years). The study was reviewed and approved by the Institutional Review Board at Carnegie Mellon University and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020), in which they indicated their views on six divisive issues: climate change, vaccination, diets and weight loss, mask use and COVID-19, capital punishment, and police officers. For each issue, participants were presented with two opposing statements (e.g., "There IS convincing evidence that human activity contributes to global climate change" and "There is NO convincing evidence that human activity contributes to global climate change") and selected the statement that they agreed with more. We also allowed participants to indicate if they had no opinion on these issues (Table 5).

[11] The eligibility criteria were the following: located in the U.S., 99% or higher approval rating, and completed at least 1,000 HITs. Participants could complete the study only once.

Table 5: Topics and corresponding statements used in Study 5.

Topic / issue	Statement 1	Statement 2
Climate change	There IS convincing evidence that human activity contributes to global climate change.	There is NO convincing evidence that human activity contributes to global climate change.
Vaccination	The health benefits outweigh the potential risks and side effects of vaccination.	The risks and side effects outweigh the potential health benefits of vaccination.
Diets and weight loss	People who want to lose weight should eat less carbs.	People who want to lose weight should eat less fat.
Mask use and COVID-19	Masks are quite effective at reducing the spread of coronavirus.	Masks are not really effective at reducing the spread of coronavirus.
Capital punishment	Death penalty deters criminals from committing violent crimes.	Death penalty does NOT deter criminals from committing violent crimes.
Police officers	Most police officers act in the best interests of the public.	Most police officers do NOT act in the best interests of the public.

Note: For each of the above issues, participants could also select "no opinion / NA".

After participants indicated their views on the six topics, we randomly selected one issue for which they agreed with one of the two statements (thus, we excluded those issues about which they had no opinion). Then, we generated a hypothetical Tweet which featured the opposite of the participant's preferred statement. For example, if the "climate change" topic was randomly selected, and if the participant preferred the statement "There IS convincing evidence that human activity contributes to global climate change", we generated a Tweet that denied that human activity contributes to climate change (Figure 8).

Figure 8: An example of a generated (hypothetical) Tweet used in Study 5.

The identity of the author of the Tweet and the engagement statistics (likes, retweets) were hidden to avoid potential biases. To ensure that participants read the Tweet and stayed engaged with the content, we asked them to write a few sentences about it (minimum 100 characters). After participants finished writing down their thoughts, we asked them how disturbed the Tweet made them feel. Similarly to previous studies, participants also indicated how confident they were that: 1) they had correct beliefs; 2) the author of the Tweet had incorrect beliefs; and 3) there was a difference in beliefs (all measured on slider scales from 0: *not at all* to 100: *very*).

Next, we included the main dependent measures. The first five questions asked participants whether they would avoid interacting with the author of the Tweet in various contexts. In particular, we asked them how likely they would: avoid working with, avoid talking to, avoid hanging out with, avoid trusting, and block the other person (all measured on continuous slider scales from 0: *not at all* to 100: *very*). The remaining four questions asked whether participants preferred being in various relationships with the author of the Tweet. In particular, participants indicated whether they would prefer the other person to be their: neighbor, colleague, family member, and partner (all measured on continuous slider scales from −100: *prefer not* to 100: *prefer*). The actual questions used for the above nine dependent measures are reported in Appendix VII.1.9. Finally, we recorded participants' age, sex, political affiliation, and political ideology (ranging from very liberal to very conservative).

Results

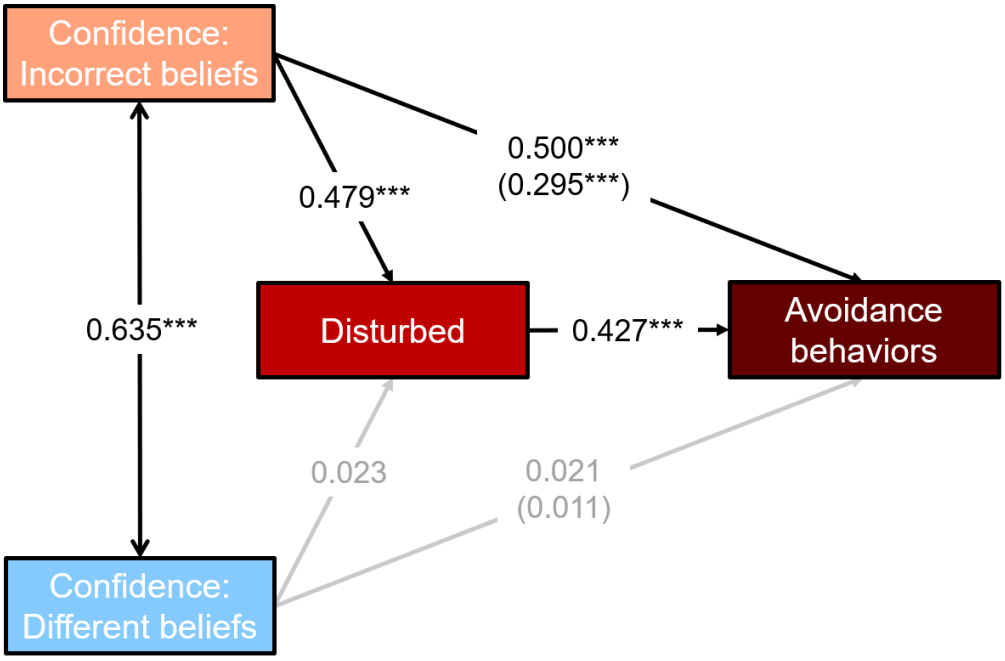
Main results. First, we conducted a linear OLS regression analysis to investigate if participants' subjective confidence ratings predicted how disturbed they were by the Tweet (see Table 23 in Appendix VII.1.10). This analysis revealed that participants who were more confident that the author of the Tweet had incorrect beliefs were significantly more disturbed by the Tweet, $\beta = 0.456$, $SE = 0.058$, $t(596) = 7.800$, $p < .001$. People who were more confident in the correctness of their own beliefs were also more disturbed by the Tweet, $\beta = 0.326$, $SE = 0.065$, $t(596) = 5.011$, $p < .001$. By contrast, participants who were more confident that there was a difference in beliefs were not more disturbed by the Tweet, $\beta = 0.028$, $SE = 0.065$, $t(596) = 0.433$, $p = .665$. These results are robust to demographic factors and political ideology (Table 23, Model 2).

Next, we looked at whether participants' subjective confidence ratings predicted their avoidance behaviors towards the author of the Tweet (for detailed results, see Table 24 in Appendix VII.1.10). Participants who were more confident that the author had incorrect beliefs reported that they would significantly more likely avoid working with them, $\beta = 0.536$, $SE = 0.063$, $t(596) = 8.545$, $p < .001$. They also indicated that they would significantly

more likely avoid talking to, avoid hanging out with, avoid trusting, and block the author of the Tweet, all $p < .001$. By contrast, neither participants' confidence that they had correct beliefs, nor their confidence that there was a difference in beliefs predicted any of the above avoidance behaviors, all $p > .334$. These results are also robust to demographic factors.

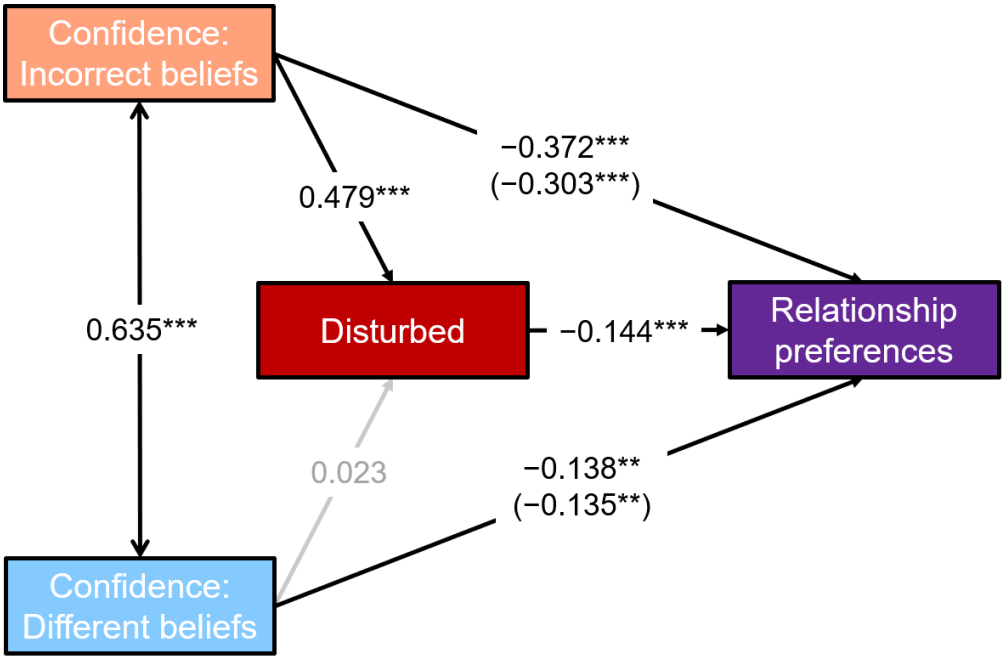
Finally, we investigated whether participants' confidence ratings predicted their relationship preferences (see Table 25 in Appendix VII.1.10). Participants who were more confident that the author had incorrect beliefs reported that they would prefer having all relationships with the author of the Tweet significantly less, all $p < .001$. By contrast, participants' confidence about the correctness of their own beliefs did not predict relationship preferences, all $p \geq .068$. Similarly, participants' confidence about differences in beliefs did not predict their preferences for less intimate social relationships, $p = .167$ and $p = .394$, for having the author of the Tweet as a neighbor and a colleague, respectively. However, for more intimate relationships (family and romantic), we did observe significant effects of the confidence about belief differences, in addition to the significant effects of the confidence about incorrect beliefs. People who were more confident that there was a difference in beliefs reported that they would prefer having intimate relationships with the author of the Tweet significantly less, $\beta = -0.337$, $SE = 0.106$, $t(596) = 3.178$, $p = .002$, and $\beta = -0.420$, $SE = 0.104$, $t(596) = 4.053$, $p < .001$, for familial and romantic relationships, respectively. These results are robust to demographic factors and political ideology.

Mediation analyses. To investigate whether reported disturbance mediated the effect of confidence ratings on avoidance behaviors and relationship preferences, we conducted mediation analyses. For these mediation analysis, we omitted participants' confidence that they had correct beliefs (since we focus on the distinction between others' incorrect beliefs and differences in beliefs). We also standardized all variables, and controlled for age, sex, and political ideology in these models. For ease of exposure, here we report the results of mediation analyses conducted on two *composite measures*: a composite measure of avoidance behaviors (Figure 9) and a composite measure of relationship preferences (Figure 10), which we obtained by taking the mean of the standardized values of the corresponding items. We report detailed mediation analyses (i.e., for each measure separately) in Appendix VII.1.11. A bootstrapped mediation with $5,000$ replications revealed that disturbance significantly mediates the effect of confidence that the author had false beliefs on avoidance behaviors, $\beta = 0.205$, 95% CI [0.155, 0.262], $p < .001$. By contrast, we found no such mediating effect of disturbance between participants' confidence about differences in beliefs and avoidance behaviors, $\beta = 0.010$, 95% CI [−0.033, 0.055], $p = .670$. Thus, participants' confidence about differences in beliefs did not predict either directly or indirectly avoidance behaviors.



Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 9: Mediation analysis: avoidance behaviors, Study 5. Coefficients: standardized Beta.



Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 10: Mediation analysis: relationship preferences, Study 5. Coefficients: standardized Beta.

Finally, a bootstrapped mediation with 5,000 replications revealed that disturbance significantly mediates the effect of confidence that the author had incorrect beliefs on relationship preferences, $\beta = -0.069$, 95% CI $[-0.123, -0.024]$, $p = .002$. By contrast, we found no such mediating effect of disturbance between participants' confidence about differences in beliefs and avoidance behaviors, $\beta = -0.003$, 95% CI $[-0.021, 0.013]$, $p = .670$. Thus, even though confidence about differences in beliefs had some direct effect on preferences for intimate relationships, this effect was independent of how disturbed participants felt.

II.8 General discussion

Across five studies we find that people are more disturbed in situations when the incorrectness of others' beliefs are highlighted, or when they recall a situation in which others held false beliefs, as opposed to cases when the differences between beliefs are highlighted, or when a recalled situation involves merely different beliefs. Furthermore, we also document the moderating role of potential negative consequences and that people are primarily disturbed when false beliefs could negatively affect someone they care about. Finally, we demonstrate that higher confidence that others hold false beliefs—but not different beliefs—triggers avoidance behaviors and reduces people's desire to form relationships with others. These results are inconsistent with, and support an alternative to, the theory of belief homophily: People are disturbed by others' false beliefs, rather than by differences in beliefs.

Although we show that people are primarily upset about others' false beliefs, we are not arguing that people cannot independently dislike situations characterized by belief dissonance. For example, when belief consonance makes social interactions more pleasant, one might even prefer to adopt shared beliefs that one does not necessarily deem to be true (e.g., Bénabou, 2013; Golman et al., 2016). Indeed, we did find some evidence for this motive in Study 5, as the lack of belief homophily *per se* had negative consequences for forming *intimate relationships*. However, outside of these close relationships, negative consequences occurred only when different beliefs were *also perceived as false*.

Which of these two interpretations is right (or, if both are right in different situations, understanding when one or the other is operative) is consequential because it can inform policies aimed to improve the quality of public discourse. If or when it is perceived incorrectness, and not differences, that fuels conflicts, then a better approach to reducing the tension caused by differences in perspectives could be to highlight the possibility that others might be right or that oneself might be wrong, rather than trying to bridge differences. Understanding what makes people disturbed when facing opposing beliefs also has practical implications for persuasion, since negative affect is an important predictor of how open and receptive they are to opposing views (Minson, Chen, & Tinsley, 2019).

This page intentionally left blank

"You must work very hard communicating [...] and take pains to communicate through the press. Staying both honest and credible with them is important. [...] He [a leader] must let the public know what he is doing and convince them he is doing the best he can to act in *their* interest. This is *very, very* important."

— *William D. Ruckelshaus* (1993)

Chapter III

The Lesser of Two Evils

Imagine being the resident of a small town whose economy is heavily dependent on a local copper smelter. Now, imagine that the head of the U.S. Environmental Protection Agency (EPA) decided to implement restrictions that caused the plant to close, causing the loss of over 500 jobs and over $20 million a year from your community. How do you think most people in your town would feel about the head of the EPA? However, what if the head of the EPA then revealed that the only other option was to leave the community at significant risk of arsenic-caused lung cancer due to the arsenic emitted by the copper smelter? Knowing that he faced two bad options would probably make his decision seem less callous and less disconnected, especially if his decision was in support of the health of the community. This scenario is based on a real dilemma that the EPA administrator, William D. Ruckelshaus, faced in 1983, regarding a copper smelter in Tacoma, Washington.[1] This case was famous because, in an unprecedented move, Ruckelshaus asked the community for their informed opinion about both options—that is, he revealed his choice set. Letting the Tacoma residents really see his choices ensured that they would better understand the context that he was facing, that both would have negative consequences, a move that likely garnered some sympathy for the difficulty of his decision, perhaps even mitigating potential damage to his public image.

While public figures have a vested interest in making sure that they have a positive image in the eyes of the public, it seems that people often display an intrinsic desire to have their decisions understood, wanting to explain themselves even in lower stakes situations and around complete strangers. Imagine that you just started riding a bike to work and are too afraid to ride on the road with cars, so you decide to ride on the sidewalk. Eventually, you pass a woman who yells at you to get off the sidewalk. Wouldn't you like to explain to her your situation? Or consider a situation when you cut someone off in traffic because you were running late. Don't you wish you could tell them about your circumstances?

[1] Parker, 1983

In this chapter, we investigate this phenomenon in the laboratory. We ask whether people care to have the context of their decisions understood even in an anonymous, one-shot interaction, when reputation should not matter. Specifically, we ask whether people are willing to incur a cost to reveal the choice set they faced to an anonymous partner who has no way of punishing them. Furthermore, we investigate what individual differences account for why some people feel the need to explain their choices, but others do not.

III.1 The threat of being misunderstood

People often find themselves in situations where they must choose from a constrained set of options, each of which, when evaluated separately, seems like a terrible choice. In such cases, despite their best intentions, decision-makers are perceived to have done the wrong thing, especially when observers have incorrect beliefs about the set of available options. In addition to famous social and moral dilemmas (e.g., the trolley problem, Sophie's choice), examples for such situations exist in almost all domains of everyday life: prioritizing which patients to save first, choosing between risky medical treatments, allocating budget cuts, making decisions about layoffs, assigning boring and unrewarding tasks that someone nevertheless has to do, deciding which internship to accept when they are all terrible, choosing a meal late at night when most places are closed, and choosing between two massively unpopular candidates running for office. With regard to the last example, during the presidential election of 2016, pundits and voters often and explicitly described it as choosing "the lesser of two evils," since the unprecedented low approval rate of both candidates forced voters to make a very difficult choice (Fabrizio, 2016).

(Mis)attribution and the correspondence bias

These choices can be difficult and frustrating to make, even when the decision-maker, in absence of an audience, has a clear preference over options—why? One reason is that people do not want their actions to be misinterpreted by others; they do not want to be misunderstood or misjudged. Observers are eager and quick to make inferences about underlying dispositions and preferences from observed behavior and choice (Heider & Simmel, 1944; Jones & Davis, 1965; Nisbett & Ross, 1991), often overattributing to dispositional factors and underestimating situational influences (Quattrone, 1982). Context is often an important, if not the primary, determinant of behavior (Mischel, 1977), and as a result, incomplete information about the context of a choice can cause onlookers to make unwarranted judgments about a decision-maker's intention and character. Indeed, judging a person based on her behavior can lead to both mistaken beliefs about her preferences (Jones & Harris, 1967) and also mistaken predictions of her future behavior (Bierbrauer, 1979).

An extreme version of this mistake is the correspondence bias when observers, completely neglecting the causal role of the situational constraints, incorrectly draw inferences about the decision-maker's underlying characteristics and dispositions from behavior that can be *fully* explained by the particular context in which they occur (Gilbert & Malone, 1995). Many of the examples provided earlier fall into this category: In most cases, a manager does not lay employees off because she has a perverse preference for firing people, a doctor does not postpone a patient's care out of personal discrimination, and many people who voted for Trump or Clinton did so because they strongly wanted to avoid electing the other candidate, not because they actively supported the person they voted for. As Miller and Nelson (2002) pointed out, observers tend to interpret others' actions as approach-motivated (i.e., choosing the option they like the most), even when they know that their own choices are motivated by avoidance (choosing the option they dislike the least).

Avoiding misattribution by revealing the choice set

In such "lesser of two evils" situations, decision-makers can avoid misjudgment by others by disclosing their full choice set, conveying that they were constrained in some way (e.g., the company board required them to lay off 40% of employees; they had to prioritize among patients who were all in critical condition; their department could only afford to hire one new faculty member). In correcting the misjudgment, revealing could help maintain a desirable social image, assuage the negative feelings of others, or simply convey the truth of their intentions (good or bad). In this sense, revealing the choice set can be considered a special case of signaling preferences and intentions: Instead of sending signals directly through the choices they select, decision-makers in this case send a message that is not directly payoff relevant (i.e., does not affect the welfare of others) (Golman, 2016). Rather, such a message draws the observers' attention to the context or choice set, which then allows them to adjust their dispositional inferences about the decision-maker in light of the newly discovered situational factors (Trope & Liberman, 1993).

People may want to avoid misjudgment for strategic (extrinsic) or non-strategic (intrinsic) reasons. When people are expecting to interact with the observer in the future (providing a possibility of reward or punishment), or when their reputation is at stake, they have a *strategic* motive to boost their image by revealing their choice set. This also means that they would choose not to reveal their choice set, if it could make them look worse. On the other hand, if a person gets some intrinsic or psychological benefit from revealing the choice set, they may do this even in anonymous, one-shot interactions, when reputation should not matter. Furthermore, depending on the underlying intrinsic drive, they may choose to reveal even if it conveys selfishness or some other negative attribute.

Previous work that has examined choice-set revealing (or hiding) behavior has done so with the presumption of its being a tool primarily to obtain a positive (or prevent a negative) image in the eyes of others, especially with respect to perceptions of generosity and fairness (Cappelen, Halvorsen, Soerensen, & Tungodden, 2017; Dana, Cain, & Dawes, 2006; Dana, Weber, & Kuang, 2007; Grossman & van der Weele, 2017). In contrast, we leave open the possibility for other motivations, and we investigate what those motivations are. Furthermore, in most of those previous studies, revealing behavior was often tied to the original decision itself: Either it was known before the decision was made that there would be a chance to reveal the choice set, or after the option to reveal was unveiled, the decision maker could return to alter their initial decision. Effectively, most studies examined the *ex ante* choice to reveal. This is important because, when the decision-maker knows in advance (before making a choice) that she will be able to provide excuses, apologies, or reveal intentions *after* implementing her choice, she can (and will) change her behavior accordingly (Andreoni & Rao, 2011; B. Ho, 2012).

What is common in these studies is that the decision-maker's behavior and the observer's initial beliefs upon learning about the decision-maker's behavior are interdependent, since at the time of decision the decision-maker has control not only over her choice but also over what the observer will know about that choice. Thus, fully disentangling the choice and decision to reveal the choice set is essential if we want to study the intrinsic motive to have choices understood, ruling out strategic, payoff relevant motives. By contrast, we have people make a decision without knowing about the opportunity to reveal. Later, or *ex post*, we give participants the opportunity to reveal (without the ability to change their original decision). In this way, the choice to reveal is only made afterwards and cannot influence the original decision. To our knowledge, no previous research has examined such *non-strategic ex post* revealing behavior, when the choice has been already made, observed, and is beyond the decision-maker's control. Are people willing to incur a cost to reveal information in such cases? And if so, what is their main motivation to do so?

III.2 Reasons for revealing the choice set

Similarly to previous work, we believe that the motivation to reveal requires caring about the beliefs of others.[2] However, this includes a variety of beliefs and a variety of reasons for caring about those beliefs. In this section, we review several of these potential intrinsic motivations that might be behind the drive to reveal the choice set.

[2]In our design, the decision maker is fully aware of their choice and its material consequences, so we rule out the types of beliefs about the self and self-signaling that do not depend on the beliefs of others.

Impression management and social image

Perhaps the account most frequently cited as an explanation for why people prefer to manage what is known about their choice set is to maintain a good impression in the eyes of others. People tend to judge, and be judged by, others along two universal dimensions (warmth and competence, see Fiske, Cuddy, & Glick, 2007), so they engage in a multitude of behaviors trying to maintain a positive impression: help others, be polite, act generously, behave competently, follow rules and social norms. Often times, they even incur costs to do so. Such concerns for a good impression and social image are not limited to strategic situations where reputation matters (i.e., involve reciprocity), so it seems that people have an intrinsic desire to uphold a favorable social image of themselves, which has been documented in laboratory studies featuring one-shot, anonymous interactions lacking the threat of punishment or the promise of reward (Andreoni, Rao, & Trachtman, 2017; Cappelen et al., 2017; Dana et al., 2006; Dana et al., 2007; Grossman & van der Weele, 2017).

Concern for the feelings or expectations of others

A second account that could explain the desire to reveal the choice set focuses on decision-maker's concern for the feelings of others, or beliefs-based altruism (te Velde, 2018). A person's beliefs could lead them to feel a variety of emotions, such as disappointment, surprise, anger, curiosity, satisfaction, or relief. Instead of being motivated by to instill a positive image of themselves in the eyes of others, the decision maker might be motivated to inspire beliefs that lead to positive, rather than negative, emotions in the other person. If a choice between "two evils" results in a bad outcome for another person, revealing the choice set might move the affected party from thinking that the world is a harsh place, resulting in sadness or disappointment, to thinking that others do their best to be generous, resulting in gratitude or pleasant surprise. A more self-serving explanation for preventing negative feelings in others, especially feelings of disappointment, could stem from wanting to alleviate one's own negative feelings (e.g., guilt). This is the central element in the theory of guilt aversion, which assumes that people incorporate others' beliefs (and even second-order beliefs) into their utility functions, and they suffer disutility when their actions fail to meet others' expectations (Battigalli & Dufwenberg, 2007). This could lead to behavior that expresses a preference for not letting others feel frustrated and disappointed (Charness & Dufwenberg, 2006; Dufwenberg & Gneezy, 2000). Regardless of the source of concern for others' feelings, this remains an additional, alternative explanation for revealing a choice set in a "lesser of two evils" scenario.

Preference for honesty

Finally, the desire to reveal the choice set can be based on a third motivation, that is neither related to others' perception of the decision-maker's niceness or the emotions of others: the desire for being honest to others. In the absence of revealing the choice set, observers will be left with an incorrect view of the decision maker's choice, and if people simply care that the view is incorrect (without necessarily desiring that the view be positive), they might be driven to reveal the choice set. It is reasonable to consider this drive because honesty is a highly valued character trait (Saxe, 1991), and strongly associated with the concept of morality and integrity in many cultures and religions (Carter, 1996). As such, it represents a construct that is fundamentally distinct from the two universal dimensions (warmth and competence) (Goodwin, 2015), and it has been shown that people even prioritize obtaining information about others' honesty and morality, rather than learning about their competence or warmth (Brambilla, Rusconi, Sacchi, & Cherubini, 2011). The desire to appear honest can be interpreted as a special case of impression management, however, it is worth noting that being perceived as honest does not necessarily imply that the person will be liked, since the person can still be perceived as mean ("brutally honest"), thus we should treat the desire for honesty as separate from the desire to be perceived as nice.

Most previous research on revealing behavior has, as mentioned before, used revealing primarily as a tool to detect concern for signaling generosity or hiding selfishness. In contrast, in the "lesser of two evils" situations, revealing could also be used simply to be or appear honest: While people who are primarily concerned about being perceived as fair or generous would not be willing to reveal their choice after selecting the *worse* option—they should only be willing to do so after selecting the *better* option—people who desire to be honest would reveal their choice set unconditionally, i.e., regardless of what their choice was. In the following sections, we turn to an empirical investigation of these motives and study revealing and hiding behavior in two behavioral experiments.

III.3 Study 6: The Good, the Bad, and the Reasonable

To study the desire to reveal the choice set, we used an abstract laboratory study. We put decision-makers into a dictator game context that only allowed them to choose between two *bad* allocation choices, i.e., options that *both* favored the decision-maker over the recipient. Furthermore, they were told that recipients would not know the options the decision-makers faced, only the one option they chose. This asymmetry of information could lead to a fundamental attribution error: Allocators would initially appear selfish, regardless of their allocation choice, even if they chose the *more* generous option by sacrificing some of their money. After the chosen option was shown to the recipient, the allocators were offered the opportunity to reveal the choice set to the recipient for a price, and they were told that the recipient would know about this decision. Under this context, we investigated whether allocators were willing to sacrifice some of their money in order to have their intentions be understood—that is, whether they were willing to pay to disclose information that had not initially been available to the recipient.

Across three conditions we manipulated whether revealing the choice set signaled that the decision-maker's intention was positive (i.e., generous), neutral (reasonable), or negative (selfish). If people care primarily about either looking generous or about the feelings of others, they should be more likely to reveal the choice set in the positive and neutral conditions than in the negative condition. Furthermore, people could be motivated to *improve* their image or the feelings of others and separately to *avoid hurting* their image or the feelings of others. When people can reveal clear generosity (i.e., that they had sacrificed some of their own money to give their partner more), they get to achieve both goals. But when their choice was simply reasonable (i.e., it increased both their own and the other person's payoffs), this would not make them look especially good; it would only avoid hurting their image. This implies that people would be less willing to pay for disclosing information when doing so merely signals that they had been reasonable than if they could reveal their self-sacrificing generosity. However, if people also care about honesty or being understood by others, we should see a significant amount of people revealing in the negative condition. Furthermore, if this latter drive represents the primary, or only, motivation to reveal, then revealing behavior should not differ across conditions.

Methods

Participants. We recruited 240 Amazon Mechanical Turk workers (108 female, mean age = 35.5) to participate in a short study on decision-making. The study was reviewed and approved by the IRB at CMU. Informed consent was obtained from all participants.

Procedure. Participants were randomly assigned to the role of the allocator or the recipient. In the first stage, every pair played a one-shot, anonymous, discrete choice dictator game with asymmetric information. The allocator had to choose between two possible allocations: "Option A" or "Option B" (see Table 6). The allocator was told that the recipient would know that she (i.e., the allocator) faced two options, but that the recipient would only see the option the allocator selected—the other option would remain unknown, though the recipient would be aware there was another option in the choice set. Thus, in the first stage, there was an information asymmetry between the allocator, who knew the full choice set, and the recipient, who knew only the chosen option. The recipient was told the same instructions, making her also aware of the information asymmetry. At this point, the allocator was not aware that she would later be given a chance to reveal the other option in the second stage.

Table 6: Experimental conditions and payoffs, Study 6

Condition[1]	Option A		Option B		Option A vs
	Allocator	Recipient	Allocator	Recipient	Option B[2]
Generous (G)	\$1.80	\$0.60	\$2.00	\$0.20	−20¢ / +40¢
Reasonable (R)	\$1.80	\$0.60	\$1.40	\$0.20	+40¢ / +40¢
Selfish (S)	\$1.80	\$0.60	\$1.40	\$0.80	+40¢ / −20¢

[1]The names of conditions reflect the allocator's intention, had she chosen Option A.
[2]Change in payoffs (allocator / recipient) if Option A is selected over Option B.

Before making the allocation decision, participants were randomly assigned to one of three conditions: "Generous", "Reasonable", or "Selfish." The conditions are named for how the allocator's choice of Option A would most likely be perceived, given the other option. A novel feature of our design eliminates wealth effects across conditions: We designed the allocations such that the most frequently chosen option had the same payoffs across conditions (i.e., Option A was kept constant across conditions).

Option B was designed such that most participants would select Option A (\$1.80 / \$0.60), the social-welfare maximizing option, in all conditions. In the Generous condition, Option B provided \$2.00 to the allocator and \$0.20 to the recipient. We expected that most participants would select Option A because it offered them a cheap way to do something generous. In the Reasonable condition, Option A offered *both* people a gain of 40 cents relative to Option B (\$1.40 / \$0.20), and thus, it was the reasonable choice over B. In the Selfish condition, we made it relatively expensive to be generous, thus we expected that most participants would be attracted to the selfish Option A.

Manipulating Option B across conditions allowed us to investigate whether allocators were more likely to reveal Option B when revealing signaled their generosity compared to when revealing highlighted something neutral, like reasonableness, or something negative, like selfishness. One important assumption of this design is that the allocator's choice, Option A, looks selfish to the recipient who does not know the choice set. To ensure this was the case, we assessed whether the recipient was unhappy about the allocation, and whether she believed that the alternative, Option B, would have granted her a larger amount of money. After receiving the allocator's transfer, we asked the recipient to rate this transfer ("How do you feel about your partner's transfer?" -100 = *very negative* ... 0 = *neutral* ... $+100$ = *very positive*) and elicited her expectations about the hidden alternative ("How much would Option B have given to your partner / to you?"). in an incentive compatible way (for details, see Appendix VII.2.1).

After the allocator reached a decision, we offered her an opportunity to reveal the unchosen alternative to the recipient. This opportunity was a surprise: The allocator was not told in advance that she would have the opportunity to reveal the other option. To reveal the unchosen alternative, she had to pay \$0.05 out of the money she previously allocated to herself (see Appendix VII.2.2). Keeping the alternative option hidden did not cost anything.[3] After the allocator reached a decision, the recipient was informed about it, and rated the allocator's transfer again. We hypothesized that these second ratings would be more favorable than the initial ratings, if and only if the revealed choice signaled the intention to be generous or reasonable, but not selfish. In the absence of revealing, the recipient would not obtain any new information, so we did not expect ratings to change in those cases.

Results

Recipients' initial ratings and guesses. To confirm that recipients evaluated Option A negatively, we looked at the guesses and ratings made by recipients *who received Option A* across conditions ($n = 84$). The distribution of guesses about the alternative option (B) suggests that the experimental manipulation worked: Most recipients guessed that the allocator could have given them more money with Option B ($n = 67$, 80%). The majority also rated the initial transfer (Option A) negatively or neutrally ($n = 56$, 67%). Furthermore, an OLS linear regression analysis revealed a significant negative correlation between recipients' initial ratings of Option A and their guesses about Option B: Higher guesses about Option B are associated with lower ratings of Option A, $\beta = -24.4$, $t(83) = 2.441$, $p = .017$ (see Appendix VII.2.3).

[3] We included this cost to reduce spurious choices to reveal and to determine whether this behavior was valued enough to give up money to perform it.

Allocators' choice. The majority of allocators chose Option A in each condition (see Table 7). Consistent with both a concern about maintaining a positive image and for the feelings of others, among allocators who chose Option A, more people decided to pay for revealing their alternative option in the Generous (52%) and Reasonable (33%) conditions than in the Selfish condition (5%). Chi-square tests of goodness-of-fit yielded a significant difference in the willingness to reveal between the Generous and Selfish conditions, $\chi^2(1, N = 45) = 12.417$, $p < .001$, $\phi = 0.525$, and between the Selfish and Reasonable conditions, $\chi^2(1, N = 61) = 6.592$, $p = .010$, $\phi = 0.329$, but not between the Generous and Reasonable conditions, $\chi^2(1, N = 62) = 2.134$, $p = .144$, $\phi = 0.186$. It is also worth mentioning that, among those who selected Option B in the Selfish condition—and thus, acted generously at a large cost—6 out of 18 allocators (33%) revealed Option A, providing further support for the hypothesis that allocators revealed their choice set more likely when it signaled generosity. By contrast, only three people, across all conditions, paid to reveal the choice set when it made them look bad or unreasonable, suggesting little to no preference for being honest.

Table 7: Proportion of allocators choosing and revealing options across experimental conditions

Study	Condition[1]	*n*	Chose A		Revealed B[2]		Chose B		Revealed A[3]	
			n	%	*n*	%	*n*	%	*n*	%
Study 6	Reveal-G	40	23	58%	12	52%	17	43%	1	6%
	Reveal-R	40	39	98%	13	33%	1	3%	1	100%
	Reveal-S	40	22	55%	1	5%	18	45%	6	33%
Study 7	Reveal-G	81	60	74%	30	50%	21	26%	6	29%
	Reveal-S	81	60	74%	19	32%	21	26%	8	38%
	Hide-G	80	60	75%	55	92%[4]	20	25%	14	70%[4]
	Hide-S	86	62	72%	54	87%[4]	24	28%	22	92%[4]

[1]R: Reasonable; G: Generous; S: Selfish; [2]Only among who chose A; [3]Only among who chose B

[4]For the Hide conditions in Study 7, the table displays the proportion of who kept the alternative visible.

Recipients' final ratings. Finally, we look at recipients' ratings *after* the alternative option was revealed (or kept hidden) to investigate whether these final ratings differ from the initial evaluations (see Table 26 in Appendix VII.2.3). Consistent with expectations, recipients' ratings of the transfer improved dramatically in those cases—and only in those cases—when revealing signaled the generosity or reasonableness of the allocator. Within-subject paired *t*-tests revealed that the increase in ratings was significant in all three conditions.[4] By contrast, ratings did not change significantly (all $p > .10$) and stayed predominantly negative or neutral after allocators decided to keep hidden the alternative option.

[4]In the Selfish condition, we compared ratings of recipients whose partner initially selected Option B, and then, by revealing Option A, signaled their generosity.

III.4 Study 7: Revealing and hiding intentions

Study 6 provides an important proof of concept—that people are willing to pay to explain their choice, even *ex post*—and also suggests this behavior is *not* motivated by the desire to be or appear honest. Rather, the results suggest the behavior is driven either by allocators' desire to look generous (or to avoid looking selfish) or by their concern for others' feelings. However, Study 6 also has a potential confound: Although the design encourages most participants to select Option A regardless of condition, there is still a sizeable portion of people who select Option B. This means that there may be some self-selection such that the allocators who choose Option A in the Generous condition systematically differ from the allocators who choose Option A in the Selfish condition (e.g., the latter group may have stronger self-regarding preferences, and more reluctant to give up any money), which might serve as an alternative explanation for the observed difference in the allocator's behavior across conditions. The primary goal of Study 7 was to replicate the main results of Study 6—that allocators are willing to pay to reveal the choice set to explain their choice, especially when doing so reflects positively on them—but in a more controlled setting, ruling out potential self-selection.

Secondly, not all generously-behaving allocators decided to reveal their choice set, and we wanted to understand whether this was a fully-informed choice, implying that they did not care much about others' beliefs, or whether it stemmed from misunderstanding the beliefs and feelings of the recipient. Did some allocators incorrectly believe that their partner was satisfied with the transfer, and for that reason find it unnecessary to reveal the choice set? Or did they know that their partner was dissatisfied, yet did not care enough about how their partner felt? To assess this, we included measures of allocators' second-order beliefs (i.e., beliefs about the recipients' beliefs).

Third, even if people were more likely to reveal if they chose generously than if they chose selfishly, we will not be able to say that those revealers do not care about honesty at all—revealing signals both honesty *and* generosity in those cases. Thus, in Study 7 we also introduced new conditions in which the choice was to hide rather than to reveal the choice set. If decision-makers care primarily about looking generous or managing others' feelings and do not care about being honest, then we should find that those whose choice appears selfish are willing to pay to hide the choice set. This would also mean that people would be willing to pay to create ambiguity in order to prevent a selfish choice from being known with certainty. The fact that many people in Study 6 left the choice set hidden suggests that many decision-makers did not have a problem with leaving another person in the dark about their actions. However, leaving the misunderstanding in place was a passive

choice; it is not clear to what extend it was actively preferred. With the new conditions, we will be able to see whether people really do not care if another person misunderstands their decision, as long as they can protect either their image or the other person's feelings. Lastly, we included new measures to probe the underlying motivations for revealing (and hiding) as well as the effects of these actions on the recipients. These measures assessed individual differences in allocators' concern for others' feelings, their desire to maintain a positive impression, their beliefs about the recipient's curiosity, and recipients' perceptions of allocator's trustworthiness.

Methods

A priori power analysis and pre-registration. We determined the required sample size per condition by an a priori power analysis using G*Power 3.1.9.2. (Faul et al., 2009). To be able to detect a difference of 15% in the proportion of revealers between conditions in a Chi-square test (Cohen's $w \geq 0.3$), with a significance level of $p = .05$ and power of $1 - \beta = 0.90$, we needed a sample size of 58 observations (pairs) per condition, therefore we decided to stop data collection after we reached 60 pairs of participants per condition. Since we planned to compare only those allocators who chose Option A, the stopping rule referred to the number of allocators choosing Option A, not the total number of allocators recruited. Therefore, we stopped data collection, after we reached at least 60 allocators choosing Option A in all four conditions. The study design, including all conditions, all dependent variables, the desired sample size, and the planned analyses were pre-registered at https://aspredicted.org/xs99v.pdf.

Participants. We recruited 674 AMT workers. Because one of the main goals was to control for self-selection, we also wanted to rule out potential confounds due to high attrition rates. As Zhou and Fishbach pointed out, high attrition rates are common in online studies, especially on AMT, and this can lead to major confounds and biases in the collected data (Zhou & Fishbach, 2016). However, we observed a very low attrition rate in Study 7, ruling out any potential confound due to attrition effects: Only 18 participants (2.7%) quit before completion. Six hundred fifty-six participants (293 female, mean age = 36.3) completed the experiment, and we included all of them in the final sample. The study was reviewed and approved by the IRB at CMU. Informed consent was obtained from all participants.

Procedure. Study 7 had a 2x2 between-subjects design: We manipulated whether the allocator's choice could be seen as generous or selfish after revealing the choice set, and whether the alternative option would be kept hidden by default (and could be revealed at a cost, as in Study 6), or would be revealed by default (and could be hidden at a cost).

The experimental procedure in the Reveal conditions was similar to the conditions in Study 6, with the following differences: We removed the Reasonable payoff set, and thus, used only the Generous and Selfish payoff sets. Then, we sorted participants randomly into these two conditions *after* they chose either Option A or B. To avoid any deception, we did this in the following way. First, the allocators were told that they had to choose between two options, but they had only probabilistic information about one option. Option A was always the same as in Study 6 ($1.80 for the allocator and $0.60 for the recipient), while Option B could be either $2.00 / $0.20 or $1.40 / $0.80 (the same as in the Generous and Selfish conditions in Study 6). Allocators were told that Option B was one of these allocations, but at the time of their choice, they were not able to tell which one (they were told that each allocation for Option B had a 50% chance to be realized, and the actual allocation would be revealed immediately after their choice). After allocators made their choice, we randomly sorted them into two conditions: In the Generous condition Option B was $2.00 / $0.20, while in the Selfish condition Option B was $1.40 / $0.80.

The second major difference from Study 6 was that we also elicited allocators' second-order beliefs about recipients' guesses and ratings: Each allocator was incentivized to predict their partner's guess about the alternative option in the first stage and how their partner would rate their transfer in the second stage (after the alternative option had been or had not been revealed). By comparing recipients' first-order beliefs and guesses to these new second-order measures, we can assess whether non-revealing behavior is driven by inaccurate second-order beliefs (e.g., non-revealers believe that the recipients rate the transfer positively). Furthermore, as one might argue, the wording of the rating used in Study 6 ("How do you feel about your partner's transfer?") might have been slightly misinterpreted by some recipients, who, instead of rating the allocator's *act* of choosing a certain allocation and transferring some amount, might have simply rated the transfer (the amount of money received) itself. To rule out any such misinterpretation, we asked recipients a second question both before and after revealing: "How do you feel about your partner as a person?" -100*: Very untrustworthy ...* 0*: Neither untrustworthy, nor trustworthy ...* $+100$*: Very trustworthy.* This question ought to measure recipients' attitude towards the allocators. We also asked allocators to indicate their second-order beliefs about their partner's trustworthiness ratings. Note that if participants interpret the transfer rating question correctly (measuring attitude towards the allocator and her actions), then both the transfer ratings and trustworthiness ratings measure the same construct, thus we should expect a very strong and positive correlation between. However, if there is any misinterpretation of the transfer ratings, then we should expect a divergence between transfer and trustworthiness ratings, as they do not necessarily measure the same constructs.

Third, we asked recipients in the first stage how curious they were about the alternative allocation ("To what extent do you want to know the unchosen option?" 0*: I do not want to know at all ...* 100*: I strongly want to know*), and we elicited allocators' beliefs about their partner's curiosity about the unselected option. Strong feelings of curiosity are often associated with negative emotions, feelings of deprivation, and people find such situations aversive (Litman & Jimerson, 2004; Loewenstein, 1994). Thus, if allocators believed that their partner was very curious about the (hidden) unchosen option, they could have decided to reveal the choice set to alleviate such negative feelings. Therefore, this measure about their partner's curiosity ought to capture such concerns for recipients' emotions.

To assess differences in social preferences as a potential explanation for the decision to reveal, we included six questions after the second stage to investigate whether allocators care about the feelings of strangers and close others, and whether they want to be perceived as fair and honest by strangers and close others (see Appendix VII.2.4). Since the experiment featured an anonymous one-shot interaction, we included questions about close others to serve as control measures, which should not have any association with the allocators' decision to reveal. By contrast, we expected a significant correlation between allocators' concerns for strangers (and the desire to be perceived fair and honest by them) and the decision to reveal, moderated by the experimental condition: Allocators who have stronger concerns for strangers would be more likely to reveal the choice set in the Generous condition, but less likely to reveal the choice in the Selfish condition.

The procedure in the Hide conditions was the same as in the Reveal conditions described above, but the allocators were told that the unchosen alternative would be revealed to the recipients by default. Then, they had the costly opportunity to hide the alternative option for 5 cents. If they decided to hide the alternative, the recipient was informed about this, and was shown only the chosen option. Everything else, including the first- and second-order belief measures was the same as in the Reveal conditions. Thus, the Hide conditions mirrored the Reveal conditions: Whereas in the Reveal conditions the allocators could pay to reveal the unchosen alternative, in the Hide conditions they could pay to hide it.

Results

Allocators' choice. The majority of allocators selected Option A (74%). We limit all subsequent analyses to these participants. Consistent with our previous findings and basic impression management goals (i.e., want to look positive), allocators were more willing to reveal their alternative if they could signal their generosity and/or make the other person feel good, even after controlling for self-selection. A Chi-square tests of goodness-of-fit yielded a significant difference in the willingness to reveal between the Reveal-Generous (30 out of 60,

50%, see Table 7) and Reveal-Selfish (19 out of 60, 32%) conditions, $\chi^2(1, N = 120) = 4.174$, $p = .041$, $\phi = 0.186$. This finding rejects the hypothesis that the difference in revealing behavior between Generous and Selfish conditions can be explained *entirely* by self-selection. However, the much smaller effect size obtained in Study 7 ($\phi = 0.186$, cf. $\phi = 0.525$ in Study 6) suggests that in the first experiment the large differences were *partially* caused by self-selection. More interesting is, however, the finding that a substantial proportion of allocators (32%) were willing to pay to reveal the choice set that made the recipient think that they had been selfish. This is inconsistent with wanting to look generous or wanting to make the other person feel good, and it is more consistent with the idea that a substantial portion of people do care about honesty in and of itself, even when it signals selfishness and/or hurts the feelings of others.

By contrast, almost no one was willing to pay to hide their choice set (13 out of 122, 11%), and a Fisher's exact test revealed no significant difference in the willingness to hide between the Hide-Generous (5 out of 60, 8%) and Hide-Selfish (8 out of 62, 13%) conditions, $p = .560$. A direct comparison between the Reveal-Selfish and Hide-Selfish conditions further substantiates this claim: A Fisher's exact test revealed that allocators were significantly more likely to pay to reveal their selfishness (32%), than to hide it (13%), $p = .016$. This asymmetry between the Reveal and Hide conditions (i.e., lot of people pay to reveal their choice but very few people pay to hide their choice set) suggests that people in this scenario do not care enough about looking generous or enough about the other person's feelings that they are willing to obfuscate their choice set and create a state of misunderstanding in the other person in order to obtain that goal.

Accuracy of allocators' beliefs.[5] One simple explanation for why 48% of allocators failed to reveal the choice set even when it could have put them in a good light, is that these people made a genuine mistake, informed by their biased beliefs: They (incorrectly) believed that their partner would be satisfied with the transfer, even without knowing what the alternative was. By comparing allocators' second-order beliefs about the recipients' ratings to recipients' actual ratings we can investigate whether they were any systematic biases in allocators' beliefs. Figure 11 illustrates the first- and second-order beliefs of participants after the alternative option had been revealed (or kept hidden). First, when looking at recipients' ratings (light grey bars), we see that the ratings on average were predominantly negative or neutral, except for when generous allocators revealed their choice set or kept it revealed, in which cases the average rating was positive, replicating the results of Study 6.

[5]The results discussed in this section are based on allocators' second-order beliefs about transfer ratings only. We found a significant and strong positive correlation between transfer and trustworthiness ratings, suggesting that the two ratings are measuring a similar underlying construct.

Figure 11: Allocators' and recipients' mean rating of the transfer (Option A) after the alternative option (Option B) was revealed (or kept hidden), Study 7. Error bars indicate $\pm 1SE$

Independent samples t-tests revealed that recipients' ratings in the Reveal-Generous condition were significantly higher after the alternative had been revealed ($M = 63.7$) than after the alternative had been kept hidden ($M = -26.4$), $t(58) = 6.549$, $p < .001$, Cohen's $d = 1.691$. Even though recipients' final evaluation of the transfer in the Reveal-Selfish condition was negative regardless whether the alternative had been revealed ($M = -15.40$) or kept hidden ($M = -40.17$), revealing was associated with marginally significantly better ratings, $t(58) = 1.846$, $p = .070$, Cohen's $d = 0.504$. A similar pattern emerged in the Hide conditions: Recipients in the Hide-Generous rated the transfer significantly higher after the alternative was left visible ($M = 45.93$) than after the alternative had been hidden ($M = -15.40$), $t(58) = 2.378$, $p = .021$, Cohen's $d = 1.226$. Recipients in the Hide-Selfish condition also rated the transfer better after the choice set was left visible ($M = 1.98$) than after it had been hidden ($M = -25.38$), however, this difference was not significant, $t(58) = 1.377$, $p = .174$, Cohen's $d = 0.492$.

Next, we looked at allocators' second-order beliefs about recipients' ratings. If allocators who did not reveal the alternative in the Reveal-Generous condition had biased beliefs, then we should expect them to hold overly optimistic beliefs about the recipients' ratings. However, we find that allocators were surprisingly accurate in the Reveal-Generous condition. Those who revealed the alternative believed that their transfer would be rated very positively ($M = 61.3$), whereas those who did not reveal correctly inferred that their transfer would be seen negatively ($M = -14.4$). The difference between the beliefs of revealers and non-revealers is large and significant, $t(58) = 6.316$, $p < .001$, Cohen's $d = 1.632$. Furthermore, these second-order guesses about recipients' ratings were not significantly different from how recipients actually rated the transfer, $t(58) = 0.194$, $p = .847$, Cohen's $d = 0.050$, and $t(58) = 0.895$, $p = .375$, Cohen's $d = 0.231$, in cases when the choice set had been revealed and had been kept hidden, respectively. Thus, allocators in the Reveal-Generous condition anticipated correctly that their transfer would be rated negatively if they did not reveal the choice set. This finding rules out that the decision to not reveal was based on biased beliefs.

By contrast, allocators in the Reveal-Selfish condition tended to systematically and significantly overestimate recipients' ratings, regardless of whether they had revealed the choice set, $t(118) = 2.831$, $p = .005$, Cohen's $d = 0.517$. However, allocators' guesses were still directionally consistent with recipients' actual ratings: Allocators who revealed thought that their transfer would be rated significantly higher ($M = 16.3$) than allocators who did not reveal ($M = -16.9$), $t(58) = 2.446$, $p = .017$, Cohen's $d = 0.662$. This pattern is consistent with how recipients evaluated the transfer in the Reveal-Selfish condition, and the similar effect sizes ($d = 0.662$ and $d = 0.504$, for allocators and recipients, respectively) suggest that allocators had accurate beliefs about the magnitude of the difference in the ratings be-

tween cases when the choice set had been revealed and cases when it had been kept hidden. Again, these results indicate that allocator's behavior whether to reveal the alternative could not be primarily driven by biases in their beliefs about what their partners would think or feel—most non-revealers were aware that their transfer would be viewed negatively if they failed to reveal their choice set.

Allocators also held accurate beliefs in the Hide-Generous condition. Allocators' belief about the recipients' rating after the alternative had been kept revealed ($M = 38.60$) was not significantly different from recipients' actual ratings ($M = 45.93$), $t(108) = 0.708$, $p = .481$, Cohen's $d = 0.135$. Allocators were slightly pessimistic in the Hide-Selfish condition: Allocators' belief about the recipients' rating after the alternative had been kept revealed ($M = -17.13$) was marginally significantly lower than how recipients actually rated the transfer in these cases ($M = 1.98$), $t(106) = 1.955$, $p = .053$, Cohen's $d = 0.376$. We did not conduct any t-tests measuring the accuracy of allocators' beliefs in cases when the alternative had been hidden in the Hide conditions, because the small sample sizes in these cases ($n = 5$ and $n = 8$, in the Hide-Generous and Hide-Selfish, respectively) would not warrant that the basic assumptions for t-tests would hold.

Individual predictors of the decision to reveal. Since the majority of allocators (89%) in the Hide conditions kept the alternative option visible, the sample who hid the alternative option in these conditions was too small to be able to draw any reliable statistical inference about their individual characteristics, preferences, or beliefs. Thus, we limit the analyses in the following section to allocators who participated in the two Reveal conditions ($n = 120$). Note also, that because the number of people who hid the alternative in the Hide conditions was small (13/122), we do not run similar regressions on the Hide conditions.

First, we examine the measures assessing participants' concerns towards the beliefs and feelings of strangers and close others. We observed significant positive correlations between all six measures, all $r > .279$, all $p < .001$, suggesting reasonable factorability. Thus, we conducted a principal components analysis and identified two primary factors that could account for these six measures: "attitude towards close others" and "attitude towards strangers" (for detailed results, see Appendix A2). These compound measurements indicate to what extent allocators cared about the beliefs and feelings of close others or strangers, respectively. The compound measures were created by taking the mean of the three corresponding items, and had reasonable reliability, Cronbach's $\alpha = 0.77$ (close others) and $\alpha = 0.89$ (strangers). We then conducted an OLS linear regression analysis, with allocators' decision to reveal the alternative (or keep it hidden) as the dependent variable, and the following variables as potential predictors: condition (Generous or Selfish), attitude towards close others, attitude

towards strangers, second-order belief about recipient's guess about recipient's alternative payoff, second-order belief about recipient's curiosity, and demographic factors (sex, age). To study the effects of different predictors, we created six models, gradually adding new predictors, or omitting non-predictive variables. We also considered possible interactions between condition and the other predictors. Table 8 summarizes the results of this analysis.

Model I has a single predictor (condition) and yields the same result as the Chi-square test reported earlier in this section: Allocators are significantly less likely to reveal the alternative if they are in the Selfish condition. In Models II and III we add the two compound variables (attitudes towards close others and strangers) and the interaction terms between these variables. As expected, the attitude towards close others does not predict the decision to reveal, and adding this variable as a predictor to Model II does not improve the predictive power. Thus, we omit this variable from subsequent models.

By contrast, the attitude towards strangers is a significant predictor of the decision to reveal, and improves the predictive power of the model. Model III indicates that people who care more about strangers' feelings and how strangers perceived them are more likely to pay to reveal the alternative option. Also, a marginally significant interaction emerges between this variable and the condition, which suggests that in the Selfish condition, allocators who care more about strangers are *less* likely to reveal the alternative. Adding the effect of attitude and the interaction between attitude and condition eliminates the main effect of condition, which suggests that the difference between the Generous and Selfish conditions is driven by the differential behavior of people who care about the attitude of strangers.

So far, we have established that individuals' concern for strangers predicts whether they will reveal the choice set to their anonymous partners. What about allocators' second-order beliefs about recipients' expectations? Surely, it feels much worse to withhold the information that the alternative option would have been terrible for the recipient, when she was expecting the opposite, than when she was expecting that the alternative was indeed terrible. But are allocators more likely to reveal the alternative, when they believe that their partner thinks that the alternative option would have been more favorable to them? To answer this question, we add allocators' second-order guesses about the recipients' alternative payoff to the model.The results clearly support that allocators' beliefs about recipients' expectations matter: Allocators who thought that their partner expected a better alternative were more likely to reveal the choice set in the Generous condition (when the alternative was in fact worse for the recipient), but *less* likely to reveal it in the Selfish condition (when the alternative was in fact better). Adding these second-order beliefs increases the predictive power of the model, and more importantly, the attitudes towards strangers remain a significant predictor.

Table 8: OLS linear regression analysis: Allocators' decision to reveal, Study 7

	Model I	Model II	Model III	Model IV	Model V	Model VI
Constant	0.500* (0.063)	0.011 (0.363)	0.131 (0.188)	–0.132 (0.225)	–0.145 (0.247)	–0.024 (0.285)
Condition: Selfish	–0.183* (0.089)	0.536 (0.551)	0.291 (0.265)	0.607† (0.324)	0.386 (0.402)	0.471 (0.432)
Attitude towards close others	–	0.006 (0.004)	–	–	–	–
Attitude towards close others X Condition: Selfish	–	–0.008 (0.006)	–	–	–	–
Attitude towards strangers	–	–	0.006* (0.003)	0.006* (0.003)	0.006* (0.003)	0.006* (0.003)
Attitude towards strangers X Condition: Selfish	–	–	–0.007† (0.004)	–0.008* (0.004)	–0.008* (0.004)	–0.008* (0.004)
2nd order belief about recipient's alternative payoff	–	–	–	0.258* (0.125)	0.253† (0.132)	0.223† (0.129)
2nd order belief about recipient's alternative payoff X Condition: Selfish	–	–	–	–0.305† (0.176)	–0.297 (0.181)	–0.270 (0.180)
2nd order belief about recipient's curiosity	–	–	–	–	0.000 (0.003)	–
2nd order belief about recipient's curiosity X Condition: Selfish	–	–	–	–	0.003 (0.004)	–
Sex: female	–	–	–	–	–	0.187 (0.140)
Sex: female X Condition: Selfish	–	–	–	–	–	–0.218 (0.190)
Age	–	–	–	–	–	–0.003 (0.006)
Age X Condition: Selfish	–	–	–	–	–	0.004 (0.009)
Adjusted R^2	.027	.028	.048	.068	.061	.050
Model Significance	.041	.098	.033	.023	.048	.097

Note: the OLS linear regression analyses above are limited to allocators who chose Option A in Reveal conditions ($n = 120$). Coefficients are unstandardized coefficients with standard errors in parentheses. Dependent variable is the choice whether to reveal Option B (1) or keep it hidden (0).
*$p < .05$, †$p < .10$, two-tailed.

What about curiosity? Were allocators more likely to reveal if they believed that their partner was more curious, thereby satisfying their curiosity? In Model V we added allocators' second-order beliefs about their partner's curiosity. However, neither this variable, nor its interaction with the condition predicts the allocators' action. Thus, allocators do not reveal their choice set to merely satiate others' curiosity.

Finally, in Model VI we test whether adding demographic factors (sex and age) can improve the predictive power of the model. However, neither variables are significant predictors, thus there is no evidence that demographic factors per se have any predictive power in explaining the decision to reveal the alternative. In sum, Model IV emerges as the model with the highest predictive power and suggests that attitudes towards strangers and beliefs about recipients' expectations are the two factors that significantly predict the decision to reveal actions. The more an allocator cares about strangers, and the higher her guess about her partner's belief about the alternative is, the more likely she is to reveal her choice set in Generous condition, and the less likely she is to reveal her choice set in the Selfish condition.

III.5 General discussion

When facing only bad options, one's choice is bound to create a misunderstanding, if the recipient of the choice or some other audience does not know the full set of choices, or is unaware of choice-limiting constraints. In this paper we investigated for what intrinsic reasons and to what extent people might be willing to reveal their choice set in such cases. We found that, many people are indeed willing to pay for revealing the context of their choice ex post, even to anonymous individuals in one-shot interactions. In Study 6, we demonstrated that, when the choice set reflected positively on the allocator, over half of allocators were willing to pay to reveal the choice set (52%). Fewer participants were willing to pay to reveal the other choice (33%) when revealing neither signaled generosity nor selfishness—only reasonableness—while almost no participants chose to reveal the choice set when it signaled selfishness (5%). In Study 7, we replicated the results of Study 6, and demonstrated that the main effects observed in Study 6 cannot be fully explained by self-selection of participants: More people preferred to reveal when it signaled generosity ("generous revealing," 50%) than when it signaled selfishness ("selfish revealing," 32%). However, this also made it clear that a significant portion of people reveal when it is likely to have a negative impact on both the allocator's image and on the recipient's feelings. This behavior of revealing selfish intentions seems largely to be explained by a preference for either being or appearing transparent. In their text comments, many of these allocators expressed the desire to be honest: e.g., "I felt like I should be honest about the reality of the decision, even though it showed I kept more for myself than for them"; "so my partner can see how everything really happened"; "I have nothing to hide!!" Others anticipated that the recipient would see their choice as reasonable or understandable: e.g., "They will see that I was just trying to maximize my bonus"; "I just thought it would make more sense to my partner if they saw what the options were, and it would explain why they received a lesser amount than my portion."

Consistent with this, we found that revealing generous intentions, unlike selfish intentions, to anonymous recipients was correlated with a stronger concern for the thoughts and feelings of strangers (but not close others). The difference between Studies 6 and 7 due to self-selection is interesting in that it suggests that the people who tend to behave generously also tend to desire to be or appear honest. However, when such people were put in a situation in which they appeared selfish, and their actions could not be interpreted as generous no matter what they did, they still preferred to be honest by revealing their choice set.

We also found a remarkable asymmetry between the desire to reveal, and the desire to hide the choice set: While most people were willing to effectively hide their selfish intentions by *not revealing* the choice set, almost no one was willing to *hide* their selfish intentions

when they had the opportunity to actively hide the choice set. That is, allocators seemed to be quite likely to passively allow a misunderstanding when it could help their image or prevent negative feelings in recipients, but they seemed decisively unwilling to actively create a misunderstanding in order to accomplish the same goal.

Did revealing behavior achieve its goal? We verified that this act of disclosing information was indeed effective at improving recipients' ratings of the transfer and the perceived trustworthiness of the allocator in the generous conditions. In Study 7, we ruled out that non-revealing allocators underestimated these beneficial effects of revealing, and thus made a mistake by choosing not to reveal. On the contrary, non-revealers were aware that recipients expected a much better alternative option (before it was revealed) and that they were dissatisfied. These results provide strong support against the explanation that non-revealing behavior is driven by inaccurate beliefs, like those stemming from the curse of knowledge (Camerer, Loewenstein, & Weber, 1989). Instead, as we found in Study 7, non-revealers had weaker concern about the thoughts and beliefs of strangers than revealers did.

If people were willing to pay to reveal the context when it would project good intentions, why were they not willing to pay to hide the context when it could hide negative intentions? Most people in the Hide-Selfish condition left their choices known to the recipient. One explanation is that, in this experimental setup, hiding the unchosen allocation would still project bad—although more ambiguous—intentions since both allocations asymmetrically favored the allocator. As one allocator commented, "Either way, I look greedy. I'd rather they just know the details." However, the act of hiding may also be qualitatively different than the act of revealing, signaling a negative trait: dishonesty. Indeed, of the allocators in the Hide-Selfish condition who did not pay to hide their choice set, a large portion (44%) explained this choice as the desire to be honest and transparent with their partner. By contrast, those who chose not to pay to reveal in the Reveal-Selfish condition in Study 7 explained their decisions in the comments primarily by saying revealing was too expensive (46%) or that they did not want their partner to know because it would make them look bad (39%). The difference is that keeping the choice set hidden required *action* in the Hide-Selfish condition but *inaction* in the Reveal-Selfish condition. As some research suggests, acts of commission are seen as more blameworthy than acts of omission (Spranca, Minsk, & Baron, 1991), and this may serve as an additional reason for why allocators were unwilling to pay to hide their choices.

Overall, these experiments have demonstrated that people are indeed willing to incur a cost to correct a misunderstanding by revealing their choice set even in a one-shot, anonymous setting, when there is no chance of future interaction, and there can be no reputational consequences. And they do this for a variety of reasons, with the strongest motivation de-

riving from wanting to send a positive signal, whether that is for self-serving image concerns or out of concern for the feelings of others. However, there are certainly some limitations and open questions. For instance, we do not have conclusive evidence regarding whether transparency and the act of revealing per se—regardless of the content to be revealed—has intrinsic value and is appreciated by the audience. We do not find an unconditional significant increase in ratings after the choice set is revealed—only if doing so reveals the generosity of the decision-maker. Therefore, it is possible that recipients did not appreciate the decision-makers' honesty at all, rather, viewed these actions negatively ("brutal honesty"). By contrast, allocators seemed to think revealing their selfishness would be perceived positively on average, suggesting those allocators may have misguided beliefs. Thus, it is not clear whether these allocators revealed because they thought honesty would make them look good to others or whether they are simply the type of people who want to live by the principle of honesty, regardless of how it affects others.

Furthermore, we did not manipulate *how* and *when* the context would be revealed. In our experiments, it was always the decision-maker who made an active choice between revealing or withholding information, immediately following his or her decision. But in real life, often it is the receiver or a third-party (e.g., a journalist, a co-worker) who can reveal such information, or the information is discovered accidentally. Voluntary disclosure by the decision-maker can be in some cases associated with honesty and transparency, while in other cases it can raise suspicion about the decision-maker's strategic motives, or even be interpreted as bragging (Berman, Levine, Barasch, & Small, 2015). There can also be a considerable delay between the choice and revealing the context, which might attenuate or enhance the effect of disclosure. Finally, in our experiments the decision-makers could reveal the *full* context of the choice (i.e., all alternatives), and if they did so, the audience was aware that the full context had been revealed. By contrast, if the choice set is more complex and contains several alternatives, the decision-maker might be motivated to reveal only partial information. Similarly, when the audience can not verify whether the full context or only some part of it has been revealed, they might perceive information disclosure as less credible, and remain suspicious about the decision-maker's intentions. We believe that the above are all relevant questions, which have important implications for psychology, public relations, and organizational behavior, and that the current paper serves as a basis for future research addressing these issues.

This page intentionally left blank

"It's not about the money...
It's about sending a message!"

— *The Joker*

Chapter IV

Avenging Minds: Belief-based Revenge

IMAGINE that you work at a company where one of your colleagues, Bob, has repeatedly claimed credit for work that you did, and as a result, is likely to receive a large year-end bonus. Luckily, you are moving to a different state for a new job, so you won't ever have to work with him again. But before you leave, you are in a position to pay him back: You are about to submit the annual performance evaluations of your peers, and know that if you write a negative review of Bob, it will impact his bonus. However, imagine that Bob would never find out that his bonus was reduced because of his bad behavior, or because of your review. Would your desire for payback be satisfied by writing that review? Would it be more satisfying if you knew that Bob would learn the reason for his reduced bonus, even if you would never work with him again? As the philosopher Peter French put it:

> *"Revenge is a very personal matter, and when it is inflicted, it is important that the target grasp the reason why. If the target does not know that he or she is paying the penalty because of his or her specific prior harming or injuring of someone or of the avenger himself or herself, the act of revenge has misfired."* (French, 2001)

If French is right in the quote above from his book on revenge, then most people would likely prefer that Bob know the reason he was punished, even if they received no material benefit from doing so. This intuition is ubiquitous in Western culture: Countless "revenge stories"—e.g., *The Odyssey*, *Hamlet*, *The Count of Monte Cristo*, *Once Upon a Time in the West*, *Gladiator*, and *Django Unchained*—depict protagonists going to great lengths, even risking their lives, just to make sure that the antagonists learn *why*, and *by whom*, they are punished. Doing so does not confer any instrumental value to the protagonist: Typically, the punishment itself is fatal; the antagonist meets their maker shortly after learning by whom and for what reason they are being punished, which leaves no room for deterrence (changing their future behavior for the better). Yet, as the audience witnesses these events

unfolding, most would feel dissatisfied if the protagonist enacted their revenge by sending the transgressor to his grave with no awareness of the reason for his death. Fortunately, most cases of revenge in real life (e.g., getting back at a selfish co-worker or teaching a lesson to an unfaithful partner) are rarely this dramatic, and more often than not involve non-physical (e.g., financial, psychological, social) forms of punishment. In the workplace, for example, revenge can take many forms, ranging from passive-aggressive behavior (e.g., withholding support) to public humiliation (Bies & Tripp, 1998; Bies & Tripp, 2006). However, affecting what transgressors believe—whether they understand the reason for being punished—can be just as important in these mundane cases as in those captivating fictional stories.

In this chapter we test whether people, when taking revenge, want to affect transgressors' *beliefs*—consistent with French's assertion, as well as with the plot-line of so many stories and anecdotes from everyday life. We present evidence from three studies which all show that people want transgressors to understand that they have been punished, and why, though not necessarily by whom. Furthermore, we find that, although there are multiple motives behind the drive to take revenge, belief-based motives often take priority over other concerns that previous researchers considered paramount in the decision to punish (i.e., distributive and retributive motives).

IV.1 Beyond retribution and deterrence

Most of the existing literature on punishment and negative reciprocity (e.g., Glaeser & Sacerdote, 2000; Carlsmith, Darley, & Robinson, 2002; R. Ho, ForsterLee, ForsterLee, & Crofts, 2002; Crombag, Rassin, & Horselenberg, 2003; Carlsmith, 2006; Schumann & Ross, 2010; Crockett, Özdemir, & Fehr, 2014) focuses on the central question of whether punishment is an end-in-itself (i.e., retribution or "just deserts," Kant, 1952, 1991) or whether it serves another, ultimate purpose (i.e., deterrence, Bentham, 1789).[1] Retributive punishment is triggered by unjust harms, and victims report a moral imperative to punish wrongdoers to restore justice, without further considerations for the beliefs or the future behavior of the transgressors. By contrast, deterrent punishment is driven by a more calculated, utilitarian reasoning: its goal is to reduce future harms by deterring wrongdoers from re-offending, and by deterring potential offenders among observers (for an extensive recent review on retribution and deterrence, see Osgood, 2017). People who have deterrence motives care about the

[1]Research suggests that there can be other ultimate goals of punishment as well, other than deterrence, for example: complying to, and enforcing, social norms (e.g., Carpenter, Matthews, & Ong'ong'a, 2004; Fehr & Fischbacher, 2004; Boyd, Gintis, & Bowles, 2010) or impressing observers (reputation management, e.g., Barclay, 2006; Santos, Rankin, & Wedekind, 2011; Raihani & Bshary, 2015). However, these goals are similar to deterrence in the sense that they are all based on some type of utilitarian calculation, as opposed to retribution. More importantly, these accounts are also agnostic towards the beliefs of the transgressors, as they only care about direct and indirect outcomes.

beliefs of the transgressor only to the extent to which those beliefs can influence future behavior (and reduce future harms), and oftentimes deterrence can be achieved without changing the mind of the transgressor (e.g., by incapacitating the offender, Darley & Pittman, 2003). Both retribution and deterrence share an important feature: an exclusive focus on *outcomes*, which may be defined either as the objective material welfare or as the subjective suffering of the transgressor and other individuals (or some combination of objective and subjective outcomes).

Recognizing that neither of these accounts do a satisfactory job at explaining a wide range of punishment behaviors—such as those of the punishers in the examples provided in the introduction—another line of research argues for the existence of a third motive for revenge, beyond retribution and deterrence: the desire to *affect the transgressor's beliefs* by communicating certain aspects of the punishment (e.g., the reason, the source) to them. The "understanding hypothesis" (French, 2001; Miller, 2001; Gollwitzer & Denzler, 2009) posits that punishers want transgressors to understand the reason why they are being punished. This is also consistent with work on the "expressive function" of punishment (Feinberg, 1965; Masclet, Noussair, Tucker, & Villeval, 2003; Xiao & Houser, 2005; Sarin, Ho, Martin, & Cushman, 2020), which argues that punishment, to a large extent, serves a symbolic, communicative purpose, in addition to retribution and deterrence. Gollwitzer and Denzler (2009) provided the first piece of empirical evidence for the understanding hypothesis, by showing that victims are more satisfied (measured indirectly as implicit goal fulfillment) when offenders signal that they understand why revenge was imposed upon them. Subsequent studies replicated this finding, demonstrating that people are more satisfied (measured directly) if they can explain to transgressors why they have been punished (Gollwitzer, Meder, & Schmitt, 2011), especially if the transgressor also acknowledges that he or she understands (Funk, McGeer, & Gollwitzer, 2014). Similarly, Sarin et al. (2020) demonstrated in a set of hypothetical vignette studies that most people think that punishments should be informative, that is "sufficiently semantically related to the offense that its communicative intent is apparent" (p. 16).

Although these studies provide preliminary evidence for the understanding hypothesis, they have limitations that the current research was designed to address. Most importantly, in prior studies participants did not base their decision to punish (or not punish) on how the punishment would affect transgressors' beliefs. In Gollwitzer and Denzler (2009), Gollwitzer et al. (2011), and Funk et al. (2014), participants enacted their revenge first, so the act of revenge itself could not have been influenced by considerations of how revenge would affect the offender's beliefs. The punisher, then, either could or could not communicate with the offender. The key dependent measure in these studies was punishers' post-communication

satisfaction. In Sarin et al. (2020) participants evaluated hypothetical vignettes, taking the perspective of uninvolved third parties, and were not asked to make any actual punishment decisions. None of this previous work utilized incentivized (non-deceptive) behavioral studies that featured interactions between real participants. Therefore, it remains unclear whether considerations of how punishment will affect what offenders believe, *per se*, affects potential punishers' decisions about whether to exact revenge.

Second, the above studies did not investigate *how strong* belief-based motives are, and whether people would trade off enacting retributive justice and affecting the offender's beliefs if both of these motives are present. Similarly, the research on the role of symbolic punishment in economic games (e.g., Masclet et al., 2003; Xiao & Houser, 2005) has only looked at whether people use symbolic punishment when that was their *only* option to punish. This line of research did not, however, investigate either the relative importance of different motives, or whether people would make trade-offs between different punishment goals.

Third, while these studies demonstrated that punishers value the ability to communicate with the offender, it remains unclear what dimension of communication makes revenge "sweet." In prior work, when punishers sent an explanatory message to the offender, that message always revealed not only the existence of the punishment but also its source—i.e., that the communicator was personally responsible for the punishment. Making the offender understand was thus always confounded with revealing the source of punishment. This poses challenges to the original interpretation of the results (i.e., the understanding hypothesis), since punishers might have an array of reasons for revealing the source, without necessarily making the offender understand the reason (we discuss these motives later in detail in the section "Preference for cognitive states: Belief-based motives").

IV.2 The three motives behind punishment

In this section, we lay out the three main motives behind punishment that we test in our experiments. Here we do not discuss deterrence motives since we create one-shot, anonymous interactions, so, in addition to belief-based motives, we focus only on the two outcome-based components of retribution: material outcomes and affective outcomes (i.e., suffering). Removing the chance of future interaction (thus, excluding deterrence motives) is essential: doing so allows us to study pure belief-based motives, which are not confounded by strategic considerations. The prospect of future interaction could introduce strategic elements and preferences over future outcomes (with potentially complex inter-temporal tradeoffs).[2]

[2]For example, if the punisher is worried about retaliation (i.e., decreased future welfare), she might prefer letting the punishment remain unnoticed. By contrast, if the punisher believes that punishment would change the transgressor's future behavior for the better (i.e., increased future welfare), the punisher might prefer to send a clear message.

Preference for material states: distributive justice

The first, and perhaps simplest motive for punishment is distributive justice: the desire to restore a fair allocation of wealth. This desire for distributive justice can be driven by either *negative reciprocity* (i.e., to harm those who have harmed others, e.g., Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Gächter, 2000), or inequality aversion (i.e., to reduce the difference in material welfare between the interactants, e.g. Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009; Raihani & McAuliffe, 2012). If people care *only* about distributive justice, they should be agnostic about how the transgressor feels and what the transgressor believes, since they only care about the (reduction of the) wealth of the transgressor. This also implies that the ability to inform the transgressor about *why* and *by whom* was the punishment enacted should *not* influence the punisher's choice. In an organizational setting, a manager could promote distributive justice in response to an employee's misbehavior or wrongdoing by withholding their bonus or some other benefit (e.g., extra paid leave, promotion). If the manager only cared about distributive justice—i.e., how equitably outcomes were distributed—then she would not care about whether the wrongdoer knew why he had suffered these negative outcomes, or even if he was aware that he had. Such a manager would be equally motivated to impose a "hidden" punishment (whereby the transgressor remains ignorant about the fact that he was punished) and an "open" punishment (whereby the transgressor is fully aware that he was punished).

Preference for affective states: comparative suffering

In addition to the objective welfare of the transgressor, the punisher might also care about how the transgressor *feels* in response to the punishment. The same material punishment might have very different affective consequences, depending on what the transgressor knows about the context (e.g., awareness of a loss, the reason for the punishment), and the punisher might want to make sure that the transgressor suffers for their misbehavior, to an extent roughly equivalent to the amount of suffering they caused (e.g., "comparative emotions", Frijda, 1994; "comparative suffering", Gollwitzer et al., 2011). Proportionality of suffering is also a key component of Just Deserts theory (Kant, 1952; Carlsmith et al., 2002), according to which transgressors deserve punishment proportional to the moral wrong they commit, and should suffer proportionally.[3] For example, if a manager is forced to choose between a harsh but hidden punishment (e.g., drastically reducing an employee's annual bonus, without ever

[3]Punishment can be also driven by how *the punisher* feels about the outcome, and acts of revenge can be used as a tool to regulate one's emotions (e.g., Gollwitzer & Bushman, 2012; Dickinson & Masclet, 2015; Jordan, McAuliffe, & Rand, 2015). However, as our paper focuses exclusively on the considerations for *the transgressor's* welfare, emotions, and beliefs, we do not make any hypotheses about the role of punisher's emotions in punishment decisions.

letting them know about what their bonus could have been) and a mild but open punishment (e.g., slightly reducing an employee's bonus, while also highlighting what their bonus could have been, so that they will feel bad about the situation), she might prefer the latter, even if ceteris paribus she would prefer to enact a harsher punishment. However, it is important to note that this individual would still not care about the transgressor's beliefs *directly*, but only to the extent that such beliefs would make the transgressor suffer. That is, this individual would be agnostic about the source of the transgressor's suffering—would not care whether the transgressor understands the reason for their misfortune.

Preference for cognitive states: belief-based motives

The punisher might also care directly about the transgressor's beliefs, independent of his or her objective material welfare and suffering. The idea that people have preferences over their own and others' beliefs, regardless of the instrumental value of such beliefs, has many applications in economics (see, e.g., Bénabou & Tirole, 2016; Loewenstein & Molnar, 2018; Battigalli et al., 2019). Here, we focus on two specific types of beliefs:

1. The transgressor's understanding of *why* they were punished (the reason)
2. The transgressor's understanding of *who* punished them (the source).

Understanding (reason for punishment). In accordance with the understanding hypothesis, we argue that punishers prefer offenders to know the reason for punishment. That is, it is not satisfying for a punisher to simply restore a fair allocation of resources between the victim and the offender *and* make the offender suffer for their misdeeds—it is also important for the transgressor to know that there is a causal link between their wrongdoing and the punishment. If this desire to make the offender understand the reason for punishment is strong enough, punishers might even let the transgressor get away with an objectively less harsh punishment (and suffer less), if doing so will ensure that the transgressor understands. For instance, if a manager cares about what a transgressing employee believes about the reason for receiving the punishment, in addition to the objective material welfare and/or the subjective feelings of the transgressor, she might prefer a milder but explanatory punishment—which makes it clear to the offender why they are being punished—over a harsher but unexplained punishment.

Personal vendetta or impersonal justice (source of punishment). Punishers might also care about what transgressors believe about the *source* of punishment. Making the punishment "personal" (i.e., if the transgressor knows the identity of the punisher) might confer further benefits, in addition to letting the transgressor know *why* they are being punished (for a review on the benefits of personal revenge, see Schumann & Ross, 2010). Most importantly, being identified as the source of the punishment might help to restore tarnished self-esteem (Crombag et al., 2003), boost self-efficacy and combat feelings of helplessness of the victim (Bies & Tripp, 2006); or signal to others (including the transgressor) that the victim does not tolerate unjust behavior and will respond to offenses (McCullough, Kurzban, & Tabak, 2013). The latter motive is particularly relevant in societies with a strong honor culture, where a personal vendetta is often the only way to "save face," i.e. to maintain the respect of the community (Jacoby, 1983; Sommers, 2009).

However, there are also major disadvantages and risks associated with personal revenge—disclosing that the victim was personally responsible for the punishment. In most industrialized societies, the civil and criminal justice systems have a de jure monopoly over enacting punishments for wrongdoings, so any personal revenge can be considered as a form of vigilantism, which itself may violate the law (Dumsday, 2009). Even if enacting revenge is not illegal in a particular situation, it might still violate social norms and might be perceived as illegitimate (Bies & Tripp, 2006), so people might prefer to remain anonymous to avoid stigmatization or legal charges. There are also a variety of other reasons why people might not want to disclose their identity to the transgressor: People might want to serve justice without escalating the conflict, especially if they are concerned about further retaliation (Eadeh, Peak, & Lambert, 2017). Finally, taking personal revenge (as opposed to relying on the impersonal justice served by institutions) might also make the victim look petty or jealous (Tripp, Bies, & Aquino, 2002). Therefore, while there are a number of good reasons for making the transgressor aware that the victim is responsible for their punishment, given these potential negative consequences, it is unclear whether, in a particular situation, people will have a preference for affecting the transgressor's belief by disclosing their identity.

IV.3 Overview of studies

We start our empirical investigation with an observational study (Study 8), in which we assess the prevalence of belief-based motives in real-life revenge decisions, in addition to the prevalence of distributive and retributive motives. Given that most people do not take revenge that often, instead of asking people directly in a survey to recall a situation in which they punished someone who had wronged them, we turned to an existing collection of revenge stories posted to an American question-and-answer website, Quora, where questions in various topics are asked, answered, and edited by Internet users. We then recruited independent coders to read and evaluate these stories, by answering a set of questions we created. This method not only allowed us to obtain less biased descriptions but also to investigate whether the presence or absence of certain elements of punishment made stories more satisfying and popular (as measured by the number of upvotes and views).

While Study 8 presents evidence for the main hypothesis—that people want offenders to understand the reason for punishment—it lacks proper experimental controls (i.e., each story is unique), and, since it is observational, it does not allow us to make any causal inference about the motives behind punishment. Thus, in Study 9 we put participants in a standardized hypothetical scenario that resembles the opening example involving the two co-workers: We tell participants to imagine that they had been treated unfairly by another person and then they can choose whether, and to what extent, to reduce this other person's bonus payment. Participants also have to choose between various messages to send to this hypothetical transgressor, which differ in how much information the transgressor knows about the punishment. By looking at the type of message people would prefer to send, we can determine whether people care about affecting the transgressor's beliefs—in addition to their welfare and suffering. On the other hand, Study 9 still does not allow us to tell: a) *how much* people are motivated by such belief-based concerns; b) whether they would actually implement the punishment if the interaction was real; and c) whether they would make any trade-offs between belief-based motives and distributive or retributive justice.

In Study 10, we re-create the scenario used in Study 9, but this time the transgressor is another (real) participant who allocates a disproportionate amount of work to the "victim" (i.e., the potential punisher). Participants actually have to complete the assigned work, and their decisions have real monetary consequences; thus, Study 10 mimics the dynamics of a real-life organizational setting. We also introduce a between-subjects experimental manipulation: We manipulate the type of message that people can deliver along with the punishment, which allows us to carefully disentangle different motives and identify mechanisms that guide punishment decisions.

IV.4 Study 8: Real-life revenge stories from Quora

Methods

Data. To create a database of stories that feature revenge, we sampled stories among the answers to the question "What was the best revenge you've ever gotten?" on Quora.[4] This question was posted on February 2, 2017, and has been extremely popular since then: It received over 900 answers and generated over 14 million page views as of February 2, 2020. In this thread people can share their experiences (either anonymously, or linked to their Quora account) with others, and describe how they had gotten revenge on someone else who had wronged them. The described situations are rich in detail (typically 1–1.5 pages long) and encompass a number of contexts and type of relationships (e.g., professional, romantic, casual), with varying forms and severity of punishments (e.g., from mild tantrums to severe financial damages or grave bodily harm). We include a screenshot of the Quora question and a sample response in Figure 12.

We recorded the first 100 answers (about 11% of total answers) in the order they appeared on the website on February 2, 2020. The order of answers to a question depends on various factors, and is determined by an unpublished machine learning algorithm using a combination of relevance, reputation of the answerer (including upvotes and downvotes), and date added.[5] We recorded the following for each answer (i.e., each "story"):

- Answer (i.e., the story)
- Name of the answerer (if available; only for determining the gender of the answerer)
- Title (or profession) of the answerer (if available)
- Engagement statistics: number of views and number of upvotes, as of February 2, 2020
- Date of answer
- Direct URL link to answer (for easy retrieval)

In addition, we also measured the word count and the length of each story (number of characters, including spaces). Most stories were about 1–1.5 pages long, with an average word count of $M = 620$ words ($SD = 284$) and an average length of $M = 3,326$ characters ($SD = 1,550$). Since there were a few very lengthy stories which would have made coding very tedious, we decided to remove three outliers that were over $10,000$ characters long (about four pages or $2,000$ words). Thus, we included 97 stories in the final data.

[4] www.quora.com/What-was-the-best-revenge-youve-ever-gotten
[5] www.quora.com/q/quoraengineering/A-Machine-Learning-Approach-to-Ranking-Answers-on-Quora

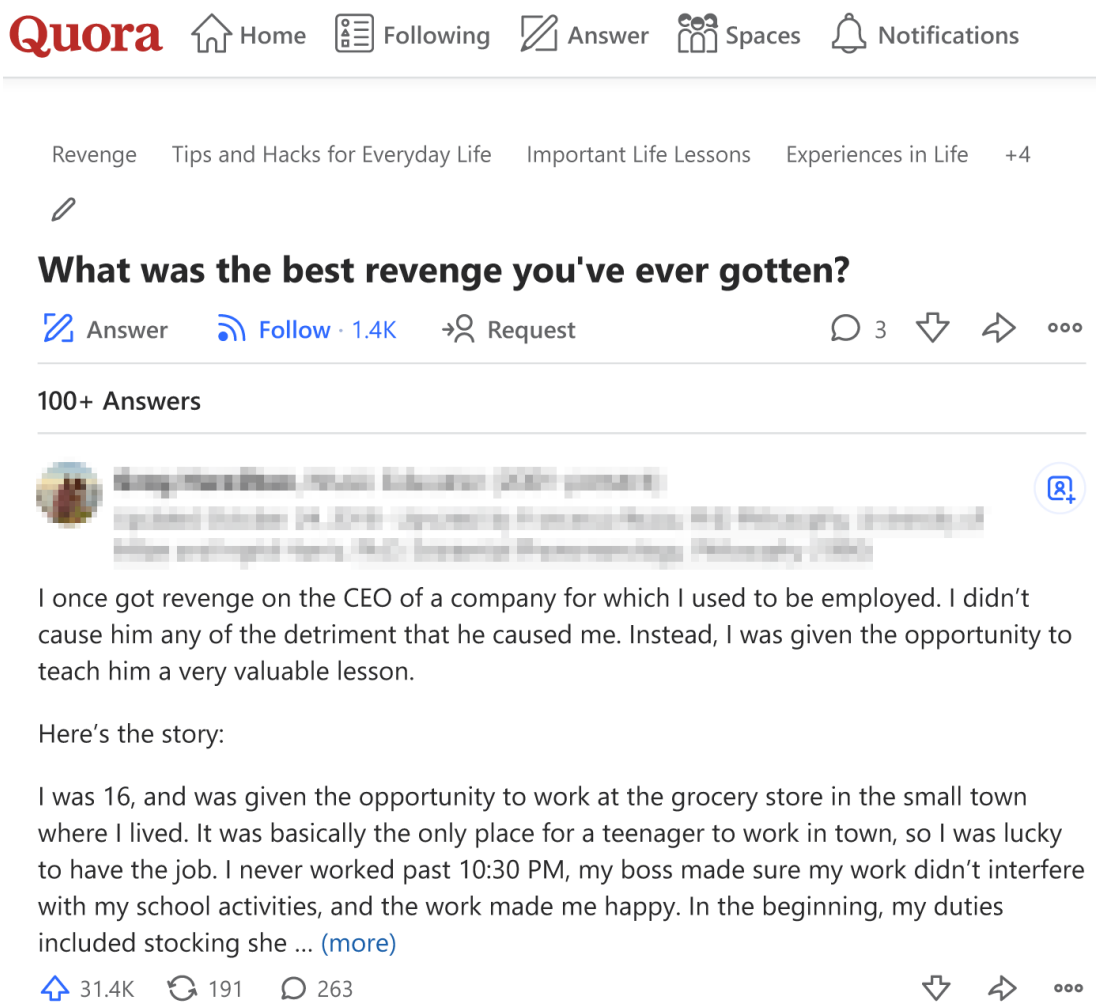


Figure 12: Screenshot of the Quora question used in Study 8. Each "revenge story" was a response to the question "What was the best revenge you've ever gotten?". We recorded the first 100 responses to this question, in the order as they appeared on the site on February 2, 2020. The screenshot above was taken of a fairly popular story, in which a former employee taught a lesson to their former CEO, following repeated abuse and bullying at the workplace. Note the author's emphasis on the lack of harm ("I didn't cause him any of the detriment that he caused me") and the role of belief-based considerations in the opening paragraph ("teach him a very valuable lesson").

Coding. We recruited 97 independent coders (58% female; age: $M = 37.0$ years) on Prolific (https://www.prolific.co) who were blind to the aim and hypotheses of the study.[6] We directed the coders to a Qualtrics survey (*Qualtrics*, 2020), in which each of them read five randomly selected stories, one at a time. Only the raw text of the stories and the names of the authors (if available) were displayed in this survey, everything else was removed, including formatting, images and videos, hyperlinks, and engagement statistics (number of views and upvotes). Coders answered the following questions about each story (responses in italics):

1. "Does the answer describe a transgression and a reaction to it?" *YES / NO*
2. "Was the transgressor punished in any of the following way(s)?" *(multiple choice)* *Physical punishment / Psychological punishment / Social punishment / Financial punishment*
3. "Which of the following BEST describes the type of interaction?" *(single choice)* *Professional relation / Intimate, close connection / Casual, loose social connection / Other*
4. "Did the transgressor suffer as a result of the punishment? If this was not described explicitly, would a reasonable person suffer if the described type of punishment was enacted upon them?" *Y/N*
5. "Did the transgressor learn the reason why he or she was punished? If this was not described explicitly, would a reasonable person learn the reason why he or she was punished?" *Y/N*
6. "Did the transgressor learn that the writer was the one who punished him or her? If this was not described explicitly, would a reasonable person learn that the writer was the one who punished them?" *Y/N*
7. "How much do the following describe the punisher's motive?" 0 *not at all* ... 4 *extremely*
   (a) "Get even / Payback (e.g. eye for an eye), restore fairness, serve justice."
   (b) "Inflict pain on the transgressor, make them suffer for what they did."
   (c) "Teach a lesson to the transgressor."
   (d) "Prevent the transgressor from harming others in the future."
8. "To what extent was the writer concerned that, if the transgressor found out why or who she/he was punished by, that the transgressor could retaliate and harm the writer again?" -3 *extremely unconcerned* ... $+3$ *extremely concerned*
9. "What is the perceived gender of the writer?" *Female / Male / Other / Cannot tell*

[6] Coders spent on average 26.6 minutes on the survey. Each coder was paid $4.50 upon completion.

Question 2 was displayed only if the coder answered "Yes" to Question 1. Questions 3–9 were displayed only if the coder selected at least one item in Question 2. If the coder either responded "No" to Question 1 or did not select any form of punishment for Question 2, the survey proceeded to the next story (or concluded if the current story was the last one). Since each of the 97 coders read a random subset of five stories, we obtained five independent evaluations for each story. Due to the large corpus of text to be evaluated (over 100 pages) and the nature of this coding strategy, it was not feasible to resolve potential disagreements via discussion and re-coding, instead, we calculated average agreements for each of the nine questions. For instance, if three out of five coders responded "YES" and two responded "NO" to a question, we assigned a value of 0.6 to that question, indicating that 60% of the coders agreed. We refer to this method as "majority agreement." We used this method of majority agreement when assessing the responses to Questions 1–6, and 9. For Questions 7–8, which had Likert scales, we calculated the mean of the ratings provided by the five coders.

Exclusions from analysis. We focus only on instances in which a transgressor is punished for their wrongdoing, so there are two conditions which must be met in each story: a) the story should describe a transgression and a reaction to it; and b) the reaction should involve some form of punishment. Therefore, we excluded all stories that failed to satisfy either of these two criteria. We excluded five stories which did not describe any transgression, and six stories which did not describe any form of punishment, according to the majority of coders (i.e., at least three out of five). Thus, there were 86 stories which described both a transgression and a punishment as a reaction to it, according to the majority of coders.[7]

Results

Context and form of punishment. There was substantial heterogeneity in both the types of situations and the form of punishment described in the stories. Most stories either took place in a professional setting (e.g., between co-workers or between an employee and a supervisor, 42%) or in intimate, close social, contexts (e.g., between family members or ex-partners, 29%). Only a minority of stories described a casual interaction (e.g., between people who barely knew each other, 19%), which indicates that in most cases (71%) the transgressor and the punisher had a shared history and knew each other well before the transgression and punishment occurred. Nine stories (10%) could not be classified, as there was no agreement between the majority of coders regarding the type of context.

[7] Out of these 86 stories, there were 54 stories that had full agreement among coders (i.e., all five coders indicated that the story described both a transgression and a punishment), 19 stories that had a strong agreement among coders (four out of five), and 13 had a majority agreement (three out of five).

The most common forms of punishment were *financial only* (23%) and *physical only* (21%), followed by *psychological only* (16%) and *social only* (12%). The rest of the stories either involved mixed punishments (12%), or could not be classified due to the lack of agreement between coders (16%). The form of punishment chosen by punishers strongly depended on the social context (Figure 13): While physical punishment was the predominant method of taking revenge in casual contexts (56% of those cases), such actions were relatively rare in other contexts (17% of professional settings and 12% of intimate situations). As we would expect, financial punishment was the most common in professional settings (31%), while psychological punishment was the most common in close social contexts (28%).

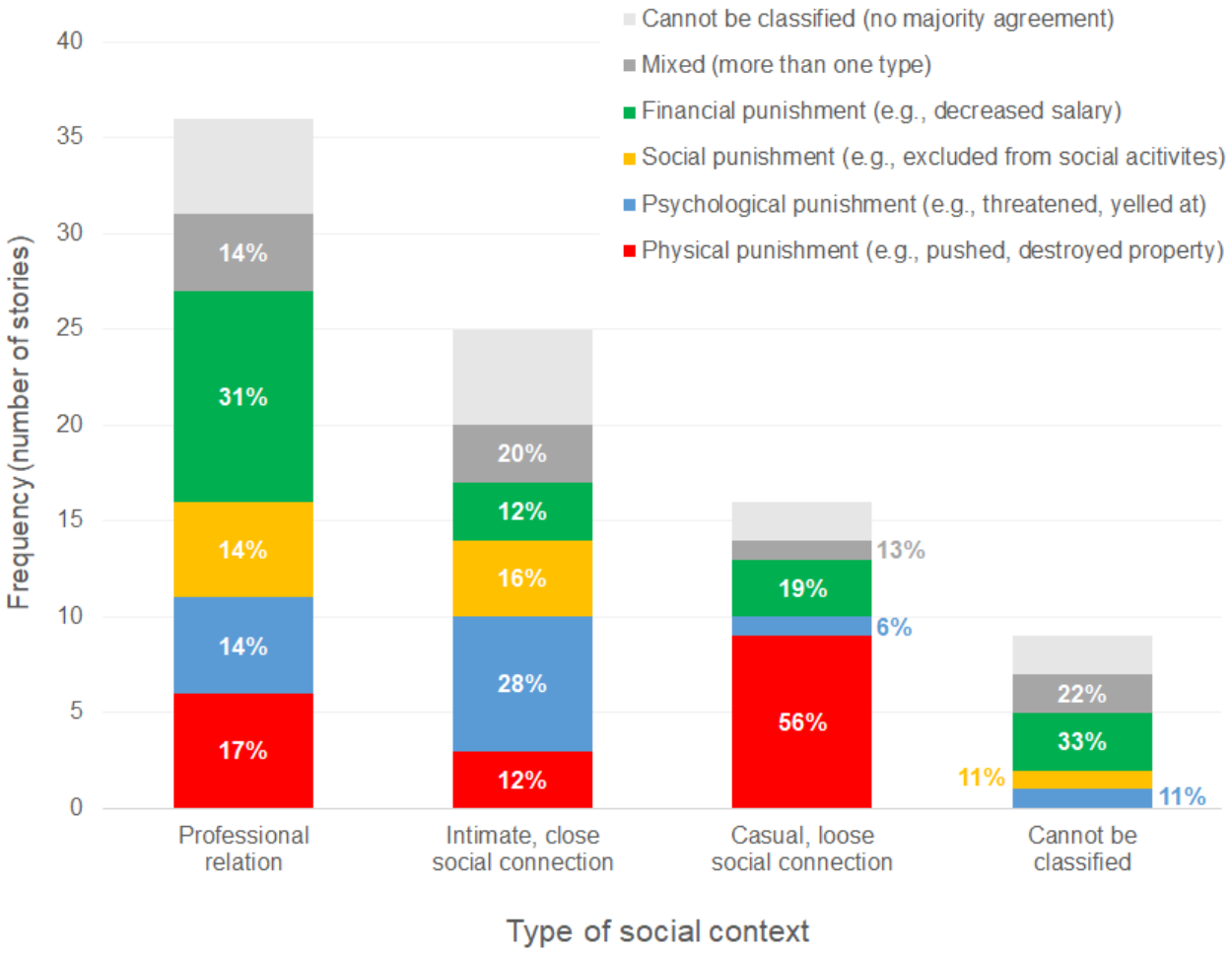


Figure 13: Frequency of different forms of punishment across social contexts, Study 8. The percentages indicate the relative frequencies of the different forms of punishment within each context.

Did the transgressor suffer as a result of the punishment?
Using the same method as for determining which stories described a transgression and a punishment (majority agreement, i.e., at least 60% of coders agreed), we find that:

- 79 stories (92%) describe the transgressor suffering,
- 6 stories (7%) describe the transgressor NOT suffering,
- 1 story (1%) is ambiguous (i.e., no majority agreement).[8]

Based on the above, it is clear that most stories described a situation in which the transgressor ended up suffering as a result of the punishment they received. However, the mere presence of suffering does not necessarily imply either that inflicting pain was the only goal or that is was the most important motive behind the act of punishment.

Did the transgressor learn the reason why they were punished, and by whom?
Using the majority agreement method, we find that:

- 59 stories (69%) describe the transgressor learning the reason,
- 22 stories (26%) describe the transgressor NOT learning the reason,
- 5 stories (6%) are ambiguous (i.e., no majority agreement).[9]

Thus, aligned with the hypothesis predicting that people care about what transgressors believe, we find that in the majority of the stories described the transgressor learning the reason why they had been punished. Similarly, consistent with the hypothesis that punishers prefer revealing their identity, we find that in the majority of stories the transgressors also learned the identity of the punisher:

- 54 stories (63%) describe the transgressor learning the identity,
- 28 stories (33%) describe the transgressor NOT learning the identity,
- 4 stories (5%) are ambiguous (i.e., no majority agreement).[10]

[8] If we classify stories only if there was a full agreement (i.e., 100%) among the five coders, the number of stories which describe, do not describe, or are ambiguous are 51, 0, and 35, respectively.
[9] If we classify stories only if there was a full agreement (i.e., 100%) among the five coders, the number of stories which describe, do not describe, or are ambiguous are 30, 6, and 50, respectively.
[10] If we classify stories only if there was a full agreement (i.e., 100%) among the five coders, the number of stories which describe, do not describe, or are ambiguous are 35, 13, and 38, respectively.

The relative importance of the motives behind punishment. While the previous results already establish that both revealing the reason for punishment and the identity of the punisher are quite widespread in situations involving revenge, these measures do not allow us to study the relative importance of different motives (i.e., distributive justice, retributive justice, etc.). To investigate this, we looked at coders' evaluation of the perceived motives behind each instance of revenge. Recall that coders rated the extent to which each of four motives—restoring justice; inflicting pain; teaching a lesson; deterring future harm—could have motivated the punisher. To test whether certain motives were stronger than others, we conducted paired samples t-tests (two-tailed) between each two of the four motives, using Bonferroni correction to adjust for multiple comparisons (factor of six since there were six related tests).

We find that restoring justice was rated the strongest among the four potential motives, $M = 2.75$ (out of 4), 95% CI [2.55, 2.95], although not significantly higher than the second most important motive: teaching a lesson, $M = 2.71$, 95% CI [2.52, 2.90], $t(171) = 0.429$, $p = 1$,[11] Cohen's $d = 0.045$. Both of the above motives were significantly stronger than the desire to inflict pain, $M = 2.09$, 95% CI [1.90, 2.29], $t(171) = 7.981$, $p < .001$, Cohen's $d = 0.701$, and $t(171) = 7.015$, $p < .001$, Cohen's $d = 0.677$, respectively. This highlights that even though the transgressor did suffer eventually as the result of punishment in the vast majority of stories (92%), making the transgressor suffer was not perceived as the main motive behind punishment (recall that these motives were not self-reported, therefore these results cannot be explained by a social desirability bias in reporting).

Finally, all previous motives (justice, teaching, suffering) were significantly stronger than the fourth motive: deterrence (preventing the transgressor from harming others in the future), $M = 1.39$, 95% CI [1.18, 1.60], $t(171) = 9.359$, $p < .001$, Cohen's $d = 1.385$, $t(171) = 12.069$, $p < .001$, Cohen's $d = 1.381$, and $t(171) = 5.265$, $p < .001$, Cohen's $d = 0.724$, respectively.

Are those stories more satisfying in which the transgressors learned why, and by whom, they had been punished?

While the previous sections provide evidence that belief-based motives behind punishment are widespread and perceived as important, we also tested whether fulfilling these motives makes observing the interaction *more satisfying* (as shown in lab studies, e.g. Gollwitzer et al., 2011). To investigate this, we looked at the engagement statistics—the number of upvotes and views each story received on the original website—as an indirect but externally

[11] Note that here we report the Bonferroni-corrected p values (unadjusted values x6).

valid measure of how satisfying stories are. Importantly, we did not ask the independent coders to rate how satisfying they found these stories, and they did not see the engagement statistics. Doing so allowed us to avoid any potential experimental demand effects, since the dependent measures of interest (number of upvotes and views) were obtained independently of the predictor variables (i.e., the coders' classifications).

It is worth to note that the number of views and upvotes are strongly and positively correlated, Pearson's $r = 0.69$, $p < .001$, and causality runs in both directions: Higher exposure (more views) may lead to more upvotes, and a higher number of upvotes makes a story more visible on the website, which generates more views. Because it would have been impossible to disentangle these two causal pathways in a non-experimental study, we decided to use both the total number of upvotes and the total number of views, independently, as proxy measures for *absolute* reader satisfaction.

In addition, we also calculated a *relative* measure of satisfaction: the number of upvotes *per views* (as % of total views), which allowed us to test if the presence of certain elements of punishment made stories *relatively* more satisfying (i.e., higher number of upvotes, given the same number of views), than stories that lacked these components. Since both the number of upvotes and the number of views (along with the story length) followed a strongly right-skewed distribution, we log-transformed these variables and included the log-transformed variables in subsequent analyses.

First, we compared whether stories in which the transgressor had learned the reason for being punished received more views and upvotes than stories that clearly lacked this belief-based element (the classification is based on coders' responses to Question 5). The stories in which the transgressor learned the reason for punishment received significantly more views, $M = 70,029$, 95% CI $[50,104,\ 89,954]$, than those in which the transgressor remained ignorant, $M = 30,305$, 95% CI $[20,318,\ 40,291]$, $t(80) = 2.746$, $p = .009$, Cohen's $d = 0.625$, The former group of stories also received significantly more upvotes, $M = 4,513$, 95% CI $[2,414,\ 6,612]$, than the latter group, $M = 1,364$, 95% CI $[558,\ 2,169]$, $t(80) = 2.987$, $p = .005$, Cohen's $d = 0.739$. Finally, stories in which transgressors learned the reason for punishment had a significantly higher upvotes/views ratio, $M = 6.12\%$, 95% CI = [3.48%, 8.75%], than those stories that lacked this element, $M = 3.51\%$, 95% CI = [2.49, 4.53], $t(80) = 2.64$, $p = .012$, Cohen's $d = 0.692$, see Table 9.[12]

[12]Note that while we report the means and 95% Confidence Intervals of the original variables, the test statistics reported in this section (t-test statistics, p values, and effect sizes) correspond to the results of the independent samples t-tests conducted on the *log-transformed versions* of these variables.

Table 9: The effect of the transgressor learning the reason for punishment.

Did the transgressor learn the reason for punishment?	Views	Upvotes	Upvotes /views	Word count	Story length
YES ($n = 59$)	70,029	4,513	6.12%	645	3,453
NO ($n = 22$)	30,305	1,364	3.51%	604	3,249
Standardized difference (Cohen's d)	0.625**	0.739**	0.692*	0.089	0.069

Note: five stories are omitted here because they could not be classified (i.e., no majority agreement between coders). $^{*}p < .05$; $^{**}p < .01$

These results already suggest that stories in which the transgressors learned why they had been punished received more attention and were appreciated more—both in absolute and relative terms—by the general public. However, it is possible that the increased popularity of this type of stories can be also explained by an alternative mechanism: If readers appreciate stories with more detail, and if those stories that described the transgressor learning the reason for punishment were also *more detailed in general*. While the level of detail is difficult to quantify directly, the length of the story can be considered as a coarse but objective proxy for the level of detail. Although stories that had the transgressors learn the reason for punishment were slightly longer on average, $M = 645$ words, 95% CI = [572, 719], than stories lacking this component, $M = 603$ words, 95% CI = [509, 697], this difference was not significant, $t(80) = 0.385$, $p = .702$, Cohen's $d = 0.089$. We obtained the same result when using the story length (number of characters) instead of word count, and found no significant difference in the average length between the two types of stories, $p = .769$.

Next, we looked at whether the audience also appreciated more when the punisher revealed their identity to the punisher, compared to cases when the punisher remained anoynmous (the classification is based on coders' responses to Question 6). While we also observe some differences in the number of views and upvotes between stories in which the transgressor had eventually learned the identity of the punisher and stories in which they did not, these differences are non-significant (both $p > .183$), and are much weaker than between stories that differed in whether the transgressor learned the reason for punishment (see Table 10).

Table 10: The effect of the transgressor learning the identity of the punisher.

Did the transgressor learn the identity of the punisher?	Views	Upvotes	Upvotes /views	Word count	Story length
YES ($n = 54$)	60,794	4,114	6.21%	655	3,500
NO ($n = 28$)	52,225	2,567	3.82%	558	3,015
Standardized difference (Cohen's d)	0.028	0.329	0.512*	0.314	0.277

Note: four stories are omitted here because they could not be classified (i.e., no majority agreement between coders). $^*p < .05$

However, stories revealing the identity of the punisher still received significantly more upvotes *relative* to the number of views, $M = 6.21\%$, 95% CI [3.34%, 9.07%], than those stories in which the transgressor never learned the identity of the punisher, $M = 3.82\%$, 95% CI [2.74%, 4.90%], $t(81) = 2.033$, $p = .048$, Cohen's $d = 0.512$. We did not find any significant difference in the word count or the length of the stories, both $p > .165$.

Finally, we conducted a linear regression analysis to investigate the combined effects of the two main predictors (i.e., the transgressor learning the reason for punishment and the transgressor learning the identity of the punisher) on the three dependent variables reported above (i.e., number of views, number of upvotes, upvotes/views ratio). In addition, we also included a potential third predictor: whether the transgressor suffered as the result of punishment. All three predictor variables were dummy variables (1 if majority "YES", 0 if majority "NO"), and the three dependent variables were the log-transformed measures of the number of views, number of upvotes, and the upvotes/views ratio, respectively. We built seven models for each dependent variable: three which had only one predictor, three which had two predictors, and one in which we included all three predictors. We also included the age of the stories (measured in days since the answer was posted) as a potential covariate in these models. We report the detailed results for all models in Appendix VII.3.1.

Regardless of which model we look at, neither learning the identity, nor suffering can predict either the absolute number of views or the absolute number of upvotes, all $p > .10$. Learning the reason for punishment, however, significantly predicts both the number of views, $\beta = 0.394$, $t(75) = 2.911$, $p = .005$, and the number of upvotes, $\beta = 0.469$, $t(78) = 3.199$, $p = .002$.[13] Consistent with the results reported in Tables 9 and 10, when *only* the reason

[13]Here we report the statistics extracted from the models with the highest adjusted R^2.

for punishment *or* the identity of the punisher is included in a model (but not both), each of these are significant predictors of the upvotes/views ratio, $\beta = 0.065$, $t(78) = 3.172$, $p = .002$, and $\beta = 0.047$, $t(79) = 2.356$, $p = .021$, for the reason and identity, respectively. When both variables are included, however, only disclosing the reason predicts the upvotes/views ratio significantly, $\beta = 0.053$, $t(75) = 2.046$, $p = .044$, whereas disclosing the identity has no longer a significant effect, $\beta = 0.019$, $t(75) = 0.783$, $p = .436$. This shows that learning the reason for punishment—but not the identity of the punisher—also led to higher relative satisfaction, in addition to higher absolute satisfaction among readers. The presence of suffering does not predict the upvotes/views ratio in any of these models, all $p > .10$.

To summarize the main findings of Study 8, we found that the desire to affect transgressors' beliefs about the reason for punishment and about the identity of the punisher are common in real situations involving punishment, and that such belief-based motives—as opposed to deterrence and making the transgressor suffer—are perceived just as strong as the desire to serve justice. Furthermore, we were able to show a link between the presence and absence of these components of punishment and the popularity of revenge stories. Situations in which the transgressor eventually learned the reason for punishment attracted significantly more views and upvotes, and had a higher upvotes/views ratio, than stories in which the transgressor remained ignorant, which provides further evidence for the understanding hypothesis and the importance of belief-based motives.

IV.5 Study 9: Unconstrained hypothetical revenge

While Study 8 provides evidence for the existence of belief-based motives driving punishment in real contexts, it does not allow us to make any casual inference about whether punishers deliberately decided to punish transgressors in this way, or whether they just had no other option to enact their revenge (e.g., they could not punish anonymously, or without letting the transgressor know the reason for punishment).

In Study 9 we deal with this issue by putting all participants in the same hypothetical scenario featuring the same offense (the scenario we opened the paper with, involving two co-workers), after which they could report the severity of punishment they would choose to impose, as well as the accompanying message for the transgressor. Thus, in Study 9 participants have the ability to influence what the transgressor believes. We chose a stylized workplace interaction and a financial punishment, since such events are, as found in Study 8, one of the most frequently reported revenge contexts, and financial punishment is the most frequent type of punishment in this context. This setup is especially relevant in organizational settings.

To create an an initial sense of unfairness—a transgression—which could then elicit a punitive reaction, we asked participants to imagine themselves in a situation in which an individual—the transgressor—behaved unfairly towards them: They allocated an unreasonable proportion of joint work (85%) to the participant, despite being paid equally for the task. We then measured whether, and to what extent, participants would reduce the payment of this hypothetical person.

Participants could also choose whether to send a message to the offender, and if they decided to do so, they could select from four preset messages, which differed in the amount of information conveyed about the punishment (i.e., whether they were punished, why they were punished, etc.). If punishers only care about the material welfare of the transgressor, then they should be indifferent about what the transgressor believes and therefore also be indifferent about the type of message to be sent to the transgressor. By contrast, if punishers care about the affective and cognitive states of the transgressor, we should see a tendency towards selecting more informative messages, i.e., messages that let the transgressor understand why, and by whom, their payoff has been reduced.

Methods

Participants and ethics statement. We recruited participants from two separate online participant pools: Amazon Mechanical Turk (MTurk) and Prolific. We aimed to collect 100 responses on each platform (i.e., 200 total), and ended up recruiting 201 participants: 100 participants on MTurk and 101 participants on Prolific.Among the 201 recruited participants, we excluded 7 (3.5%) who failed the attention check question. The final sample contained 194 responses: 96 from MTurk (44% female; age: $M = 38.6$ years) and 98 from Prolific (45% female; age: $M = 32.5$ years). The study was reviewed and approved by the IRB at CMU and conducted ethically. Informed consent was obtained from all participants.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020), in which they read a hypothetical scenario. In it, the participant was interacting with another person who was responsible for the allocation of work (a numerical addition task) between him or her and the participant. In the scenario, both the participant and the other person received a fixed compensation for the work, regardless of how the task was split between them. Then, the participant was told that the other person had allocated 85% of the work to them (i.e., to the participant), which, given the fixed compensation, would be considered very unfair. After learning about the unfair allocation of work, the participant indicated how he or she would have felt if this happened to them. First, the participant had to select from the following list of 12 feelings (presented in a randomized order): *amused, angry, betrayed, disappointed, envious, grateful, happy, pleased, proud, sad, satisfied, surprised*. Then, for each feeling that the participant selected, they had to indicate how strongly they would have experienced that feeling (continuous scale from *1: A little* to *7: Extremely*).

Next, the participant was informed that the other person was going to receive a surprise bonus of \$1. Before receiving this bonus, however, the participant had the opportunity to reduce this bonus and to send a message to the other person. Using a slider, the participant could adjust the other person's bonus to any amount X between \$0 and \$1 (including \$0 and \$1). Simultaneously (on the same screen), the participant could choose one of the following five options, presented in a randomized order:

- Do NOT send any message. [ignorance]
- Send a message: "You received an extra bonus of \$X." [bonus only]
- Send a message: "You received an extra bonus of \$X out of \$1." [suffering]
- Send a message: "You received an extra bonus of \$X out of \$1 because you were unfair to your partner." [justice]
- Send a message: "You received an extra bonus of \$X out of \$1. Your partner decided to reduce your extra bonus because you were unfair to them." [revenge]

As the participant adjusted the slider, the amounts X for all the options changed at the same time, and the individual could adjust the slider and choose between the options at will before clicking on "next" (see Figure 14). In subsequent sections we will refer to the above messages by the corresponding labels (in brackets; these labels were not displayed to participants). After participants reached a decision, they wrote an open-ended explanation for their choice and answered the attention check. Finally, we recorded participants' sex, age, and political affiliation.

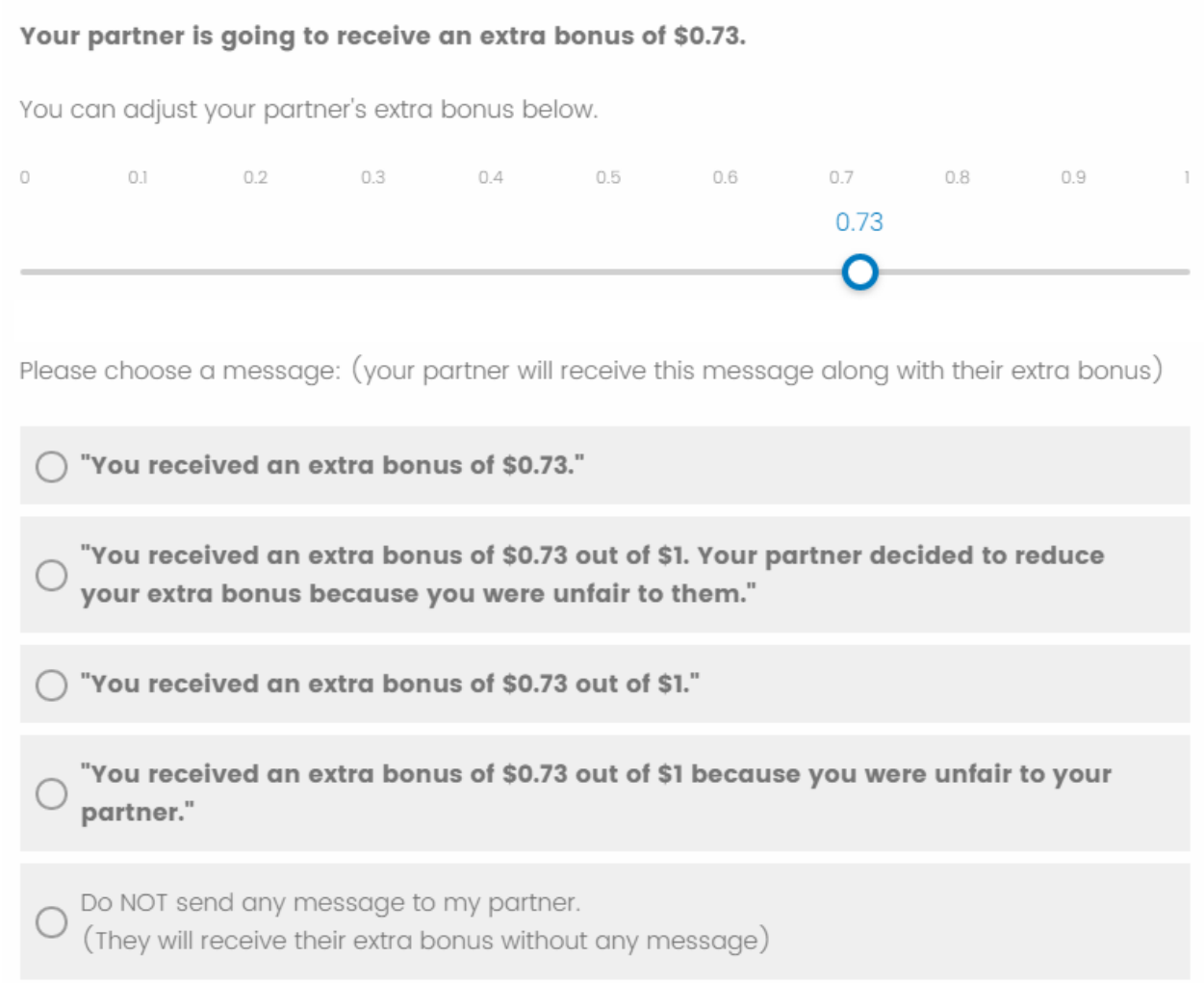


Figure 14: Decision screen in Study 9: Participants could freely adjust the other person's payoff, and at the same time, they could choose between not sending any message or sending one of the four preset messages.

Results

Punishment choices. Pooled across both samples, the overwhelming majority of participants, 169 people (87%), decided to reduce the other person's payoff (84% in the Prolific sample and 91% in the MTurk sample). The average reduction (including people who did not reduce) was \$0.76, leaving the unfair person only \$0.24 on average. The most frequently chosen amount (chosen by 40% of participants) was the full reduction (i.e., −\$1, leaving the partner \$0). These results indicate a strong overall preference for distributive justice, i.e., a desire to reduce the transgressor's material welfare. Figure 15 shows the full distribution of choices across the two subject pools.

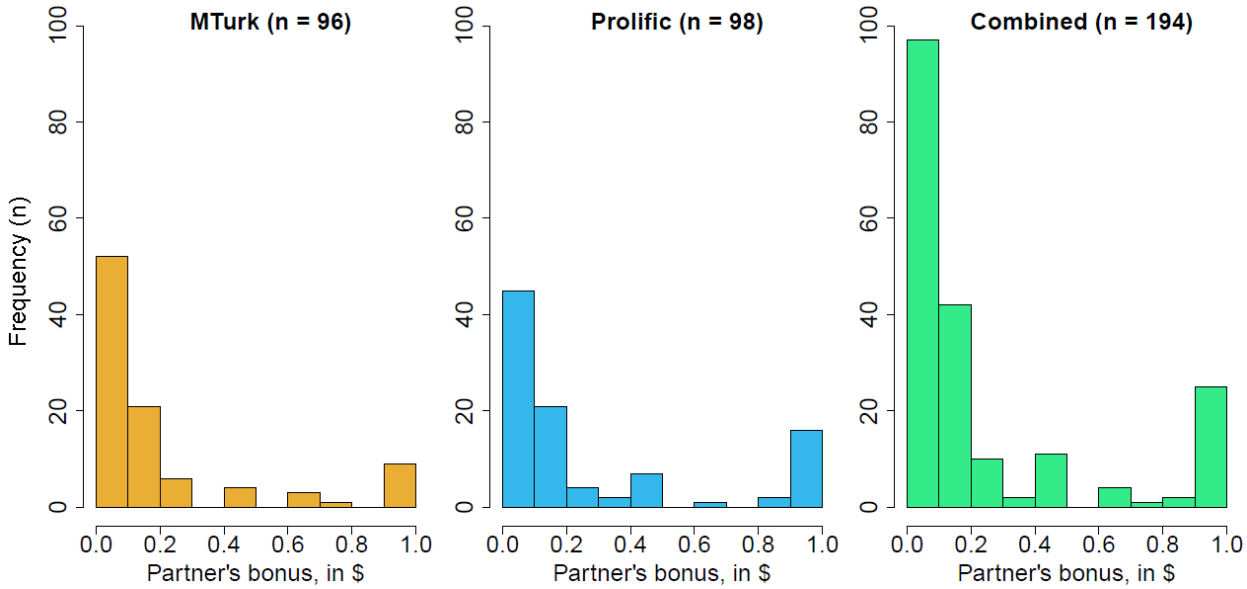


Figure 15: Histograms of punishment choices in Study 9 by subject pool: MTurk (left panel), Prolific (middle panel), and MTurk+Prolific combined (right panel). Lower values along the X-axes indicate harsher punishments, i.e. larger reductions of the transgressor's bonus.

Message choices. The most frequently chosen message was the most informative, "revenge" message ("... Your partner decided to reduce your extra bonus because you were unfair to them."): 81 participants (42%) chose this option. Thirty-seven participants (19%) chose not to send any message (ignorance), while 30 (15%) sent the bonus only message. Finally, 28 participants (14%) sent the justice message, and 18 (9%) sent the suffering message (see Table 11, Column 1). The above proportions reflect the popularity of messages among *all* participants, regardless of their punishment decisions. However, it is more meaningful to separately investigate the distribution of messages chosen by people 1) who reduced their partner's bonus and 2) who did not.

Table 11: Number (and proportion) of participants choosing each message, Study 9.

Message	Participants							
	(1) Everyone (any X)		(2) Did not punish (X = $1)		(3) Punished (X < $1)		(4) Fully punished (X = $0)	
Ignorance (no message)	37	19%	5	20%	32	19%	18	23%
Bonus only ($X)	30	15%	12	48%	18	11%	2	3%
Suffering ($X/$1)	18	9%	5	20%	13	8%	2	3%
Justice ($X/$1 + unfair)	28	14%	1	4%	27	16%	9	12%
Revenge ($X/$1 + unfair + partner responsible)	81	42%	2	8%	79	47%	46	60%
Total	194		25		169		77	

Note: Here we report the pooled data across the two samples, but the distribution of messages by punishment type are similar across the subject pools (Appendix VII.3.2)

Among the 25 participants who did not reduce their partner's payoff (Column 2)., the vast majority (88%) chose either the no message (20%), the bonus only message (48%), or the "suffering" message (20%), the last of which simply informed the partner that they have received the full bonus (thus, receiving this message did not cause any suffering in this case). Only a few participants (12%) chose either the justice or revenge messages.[14] By contrast, among participants who decided to reduce the other person's payoff (Column 3), the majority, 106 sent either the justice (16%) or revenge message (47%), thereby making sure that the other person understood the reason why they received a reduced bonus. Among participants who enacted the harshest punishment possible (i.e., reduced the other person's payoff to $0, Column 4), an even larger proportion, 72% chose these messages. However, it is worth noting that there was a substantial group who decided to keep this severe punishment hidden by not sending any message (23%). The open-ended explanations provided by these participants clarify that these were deliberate decisions, rather than errors: These participants had either a strong preference for fairness but did not want their partner's to know about the reduced bonus (e.g., "I don't think they deserve to receive a bonus. I'm not trying to be malicious so I'd prefer they didn't receive a message along with it") or they thought that their partner's action was so egregious that they *did not even deserve* an explanation (e.g., " I did not choose to send a message because he/she does not deserve an explanation").

[14]Based on the open-ended explanations, these participants wanted to let the other person know that they were unhappy, but they wouldn't want to reduce the other person's payoff (e.g., "I'd want to say something about making me do more work, but I wouldn't reduce the money"). This is consistent with the expressive function of punishment (e.g., Feinberg, 1965).

IV.6 Study 10: Trade-offs in real punishment

In Study 9 we established that in this type of situation (unfair allocation of work in a professional setting), most people would impose a severe financial punishment upon the transgressor, but, more importantly, they would also prefer to send a clear explanation for why they did so. This finding already corroborates the natural observations in Study 8, and highlights that punishers are not indifferent about the transgressor's beliefs, even if those beliefs don't bear any consequences on the material and emotional well-being of the transgressor. However, participants' choices in Study 9 were unconstrained, and they were not forced to make a trade-off between different goals (e.g., distributive justice, retributive justice, belief-based motives), which does not allow us to draw any inference about the relative value punishers place on affecting the transgressor's beliefs compared to the transgressor's monetary outcomes, or whether they would be willing to make compromises between these distinct motives. To test whether people care about the transgressor's affective reactions and beliefs, and whether they are willing to make trade-offs between motives, we set up an incentivized experiment in which participants who believed they had been harmed by a transgressor made choices between different levels of punishment and different communications that conveyed information to the perceived transgressor: Participants could either punish the transgressor severely, moderately, or not at all.

First, in a baseline condition we measured participants' preference for *pure* distributive justice, in the absence of any information about the punishment, i.e., a case in which punishments were completely hidden from the transgressors and transgressors were not even aware of the possibility of receiving a reduced payoff. Then, in three other conditions we gradually added more information (e.g., a notification to the transgressor informing them about the reduction; an explanation of why they were being punished) to the punishment option that was the *least preferred* in the baseline condition (moderate punishment), while keeping constant the other two punishment options. By doing this, we were able to measure whether adding more information to the previously least preferred punishment option shifted participants' preferences towards what would otherwise be the non-preferred option. The study was designed to test between the following four hypotheses:

H1: **Distributive justice only.** There is no difference in punishment behavior across conditions: Punishers are equally likely to choose the moderate punishment options.

H2: **Comparative suffering.** Punishers are more likely to choose the moderate punishment option in those conditions in which the added information makes the transgressor suffer (suffering, justice, and revenge), compared to the ignorance condition in which the offender remains unaware of the fact that they had been punished.

H3: **Understanding.** Punishers are more likely to choose the moderate punishment option in those conditions in which the added information lets the transgressor understand the reason for punishment (justice and revenge), compared to conditions in which the offender remains unaware of the reason for punishment (ignorance and suffering).

H4: **Personal vendetta.** Punishers are more likely to choose the moderate punishment option in the revenge condition, in which the added information reveals the identity of the punisher, compared to the other conditions, in which the offender remains ignorant about the identity of the punisher (ignorance, suffering, and justice).

This experimental design allowed us to test both whether adding new information to the least preferred (moderate) option "crowds in" non-punishers (i.e., the possibility of communicating leads participants who would otherwise not have punished, to punish) and whether more information "crowds out" severe punishers (i.e., participants eschew the more severe punishment they would otherwise have preferred in exchange for the ability to communicate).

Methods

Participants and ethics statement. We recruited 1,959 participants on Prolific.[15] Among recruited participants, 153 (7.8%) quit the experiment before being matched with another person (i.e., before being assigned to an experimental condition). Among the 1,806 participants who were matched with another person (i.e., 903 pairs), we excluded 101 (5.6%) participants from our analyses: Two (0.1%) duplicate responses (people who participated more than once), 17 (0.9%) incomplete responses (people who quit before completing the experiment), and 82 participants (4.5%) who failed at least one comprehension check question or who failed the attention check question. We stopped data collection as soon as we obtained 200 observations (i.e., 200 Recipients) in each of the four conditions, after applying the exclusion criteria. The exclusion criteria, as well as the data collection stopping rule were preregistered at AsPredicted.org (https://aspredicted.org/blind.php?x=p55qc2). The final sample contained 1,705 responses (50.3% female; age: $M = 34.5$ years). The overall exclusion rates (duplicate + incomplete + failed checks) were not significantly different across conditions, $\chi^2(3, N = 903) = 1.126$, $p = .771$. We report the detailed exclusion statistics by condition in Appendix VII.3.3. The study was reviewed and approved by the Institutional Review Board at Carnegie Mellon University and conducted ethically. Informed consent was obtained from all participants.

[15] https://prolific.ac. Eligibility criteria: Participants must be U.S. citizens residing in the U.S., speak English as their first language, and have an approval rating of at least 99%. Participants could complete the experiment only once.

Procedure. Participants were directed to a Qualtrics survey (*Qualtrics*, 2020). After reading the general instructions, participants had to answer three comprehension check questions about the experimental task and their payment. If participants answered all three questions correctly, they proceeded to the matching stage in which they were matched in pairs in real-time using the Qualtrics extension SMARTRIQS (Molnar, 2019; Molnar, 2020). Within each matched pair of participants, one person was randomly assigned to role A (i.e., the Allocator), while the other participant was assigned to role B (i.e., the Recipient).[16] In the first stage of the experiment, the Allocator made a decision about how to divide work between him- or herself and the Recipient. Participants were told that they would receive a fixed compensation for their work ($1.50 each), regardless of how the work would be split between them. The work consisted of a real-effort "slider" task: Participants had to adjust several sliders to preset (random) values (see Figure 16).

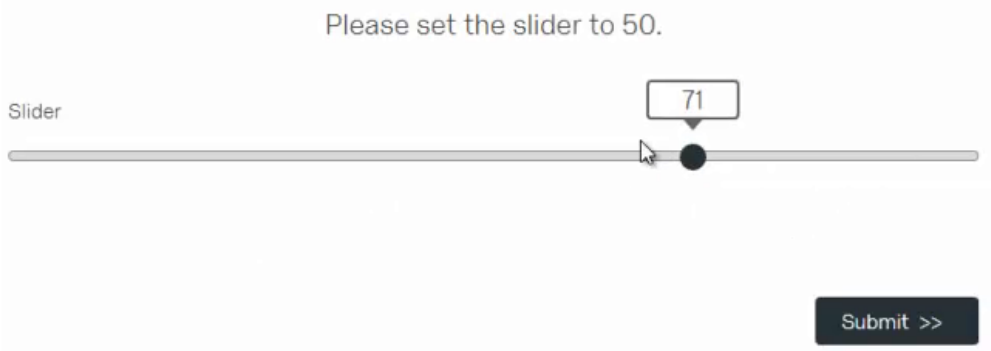


Figure 16: The slider task used in the first stage of Study 10. The task consisted of adjusting sliders to preset random values. Each pair of participants had to complete 50 sliders combined.

This task is adopted from Gill and Prowse (2012), and is frequently used in economic experiments. The task is designed to capture exerted effort only (it is skill-independent) and to minimize intrinsic motivation. Each pair of participants had to complete 50 of these sliders combined, and the Allocator decided how to split this work within the pair. Unbeknownst to the Recipient, however, the Allocator could not allocate any number of sliders: The Allocator had to choose between two options, both of which allocated a disproportionate amount of work to the Recipient. One option left the Allocator with only 10 sliders, assigning the 40 remaining sliders to the Recipient. The other option was even more unequal: It assigned 5 sliders to the Allocator and 45 sliders to the Recipient. Since the Recipient did not know about this constrained choice set, the Allocator's choice was perceived as unfair, regardless of which option was chosen (recall that participants received a fixed compensation).

[16]Throughout the study, participants were simply referred to as "A" and "B"—participants were not referred to as "Allocator" or "Recipient" at any point during the study to avoid potential biases associated with these labels.

This information asymmetry was a necessary element of the experimental design, because it guaranteed that all Recipients would face a very unfair allocation of work. Otherwise, most Allocators would have chosen fair allocations (e.g., 25/25), which would have made the experiment unsuitable for studying punishment behavior. After the Allocator made a choice, the decision was transmitted to the Recipient. Then, the Allocator completed his or her share of work (5 or 10 sliders), after which the survey concluded for them.

The Recipient, however, was presented with a second (surprise) stage immediately after learning the Allocator's decision, but before starting to work on the slider task. This second stage served as the main part of the experiment. In this stage, the Recipient was told that both he or she and the Allocator would receive a surprise bonus of $1.00 (each), in addition to the fixed compensation for the slider task. In addition, the Recipient had the opportunity to decrease the Allocator's surprise bonus, without any cost to the Recipient.[17] The Recipient's choice set was constrained, since we wanted to create trade-offs between different punishment motives. The Recipient could choose between three options:

- Do not decrease the other person's bonus, leaving them the full $1.00.
  [**no punishment**]
- Decrease the other person's bonus by $0.50, leaving them $0.50.[18]
  [**moderate punishment**]
- Decrease the other person's bonus by $0.90, leaving them only $0.10.
  [**severe punishment**]

If the Recipient chose the no punishment or severe punishment options, the Allocator simply received the $1.00 (or $0.10) bonus after the experiment, without any further message or explanation. If the Recipient chose the moderate punishment, the Allocator received a message along with the $0.50 bonus.

[17] Unlike in most other experiments studying punishment behavior, punishment was not costly to the punisher in our experiment, since our hypotheses are agnostic about the trade-off between utility from the punisher's *own payoffs* and the other sources of her utility. Instead, we focus on the trade-offs between the punisher's utility from the *transgressor' payoff* and the punisher's utility from the *transgressor' suffering and beliefs*.

[18] We selected this particular level for the moderate punishment option (−$0.50) based on the results of Study 9, which suggested that most participants (over 80%, see Figure 15) would prefer either the severe punishment (−$0.90) or the no punishment (−$0) over a reduction of −$0.50. This allowed us to introduce a moderate punishment option that was less preferred than the other two options in the baseline (ignorance) condition.

These messages were manipulated between subjects across four conditions:

- No message. [*ignorance* condition]
- "Your bonus has been reduced by $0.50." [*suffering* condition]
- "Your bonus has been reduced by $0.50, because you were unfair to your partner in the previous task." [*justice* condition]
- "Your bonus has been reduced by $0.50. Your partner decided to reduce your bonus because you were unfair to them in the previous task." [*revenge* condition]

Figure 17 depicts the Recipient's decision screen in the revenge condition:

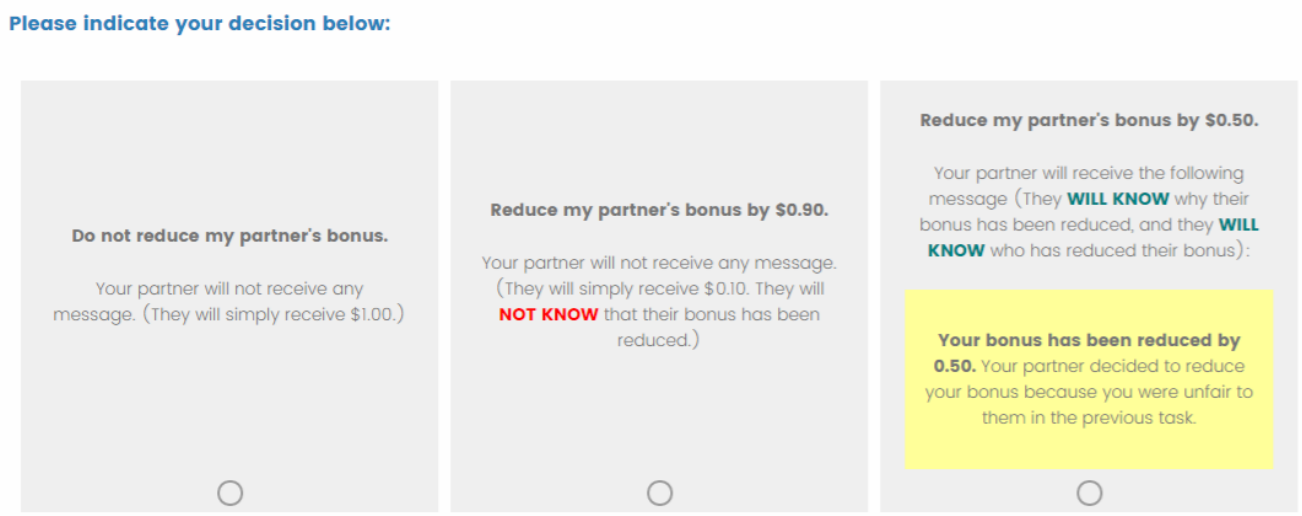


Figure 17: Sample screenshot of the punishment decision in the revenge condition in Study 10. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order. The no and severe punishment options were identical in all four conditions, while the moderate punishment option was manipulated across conditions. Sample screenshots from all four conditions are included in Appendix VII.3.4.

Since the experimental manipulation was implemented between subjects, each Recipient made only a single decision.[19] After making a choice, the Recipient was asked to explain their choice (open-ended response), and then answered three comprehension check questions about the consequences of their choice. Then, the Recipient was asked to recall how he or she felt when they first saw the Allocator's decision. The Recipient was presented with

[19]We decided to use a between-subjects design, instead of asking Recipients to make multiple decisions with different options, to minimize the potential experimental demand effects (Zizzo, 2010; Charness, Gneezy, & Kuhn, 2012).

the following list of feelings (presented in a random order, except for the last two items): *angry, beloved, betrayed, calm, disappointed, happy, relaxed, sad, satisfied, surprised, other (please specify)*, and *none of the above*. He or she first indicated for each, whether he/she had experienced the emotion, then for each identified as experienced, indicated how strongly they had experienced it (on a continuous scale from *1: a little* to *7: extremely*). Following this, we elicited the Recipient's beliefs about four things: the morality of each of the three options that each was faced with, the downstream effect of each option on the Allocator's behavior, the effect of the options on the Allocator's feelings, and the effect of the options on the Allocator's sense of guilt:

- "How MORAL or IMMORAL would it be to choose the following options?" (from −100*: very immoral* to +100*: very moral*)
- "Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?" (from −100*: very bad* to +100*: very good*)
- "Would YOUR PARTNER FEEL GUILTY or would YOUR PARTNER FEEL PROUD about his or her allocation of work?" (from −100*: very guilty* to +100*: very proud*)
- "Would YOUR PARTNER TREAT OTHERS WORSE or would YOUR PARTNER TREAT OTHERS BETTER in the future?" (from −100*: much worse* to +100*: much better*)

To make sure that Recipients were paying sufficient attention to the above questions, we included an attention check question after these items, and excluded everyone who failed to answer the attention check correctly. Then, we recorded the Recipient's age and sex. Finally, the Recipient had to complete the work assigned to him or her: adjusting 40 (or 45) sliders. At the end of the survey, we asked both Allocators and Recipients to indicate the extent to which they believed that they were interacting with a bot or a human partner (continuous scale from −100*: definitely a bot* to +100*: definitely a human*). We used this question to conduct robustness checks on our main analyses, by excluding, in a set of ancillary analyses, those participants who had a strong (but incorrect) belief that their partner was a bot, as this might have affected their decision.

Results

Allocators' task allocation. Out of the 900 Allocators who were included in the final sample, 810 (90%) chose the 10/40 allocation and 90 (10%) chose the 5/45 allocation.

Main results: punishment decision. Out of the 805 Recipients who were included in the final sample, 270 (34%) chose no punishment, 245 (30%) chose moderate punishment, and 290 (36%) chose severe punishment across the four conditions combined.

In the ignorance condition only a minority of Recipients (20%) chose the moderate option, indicating that when only distributive preferences matter, most people prefer the no punishment (37%) or severe punishment (43%) options. Aligned with the hypothesis that people derive utility directly from the transgressor's understanding (i.e., the transgressor's beliefs), a significantly higher proportion of Recipients chose the moderate punishment option in the justice (41%) and revenge (34%) conditions than in the ignorance condition (20%), $\chi^2(1, N = 402) = 20.345$, $p < .001$ and $\chi^2(1, N = 401) = 8.694$, $p = .003$, respectively (see Figure 18). Recipients also chose the moderate punishment significantly more likely in the justice (41%) than in the suffering condition (26%), $\chi^2(1, N = 404) = 9.496$, $p = .002$.

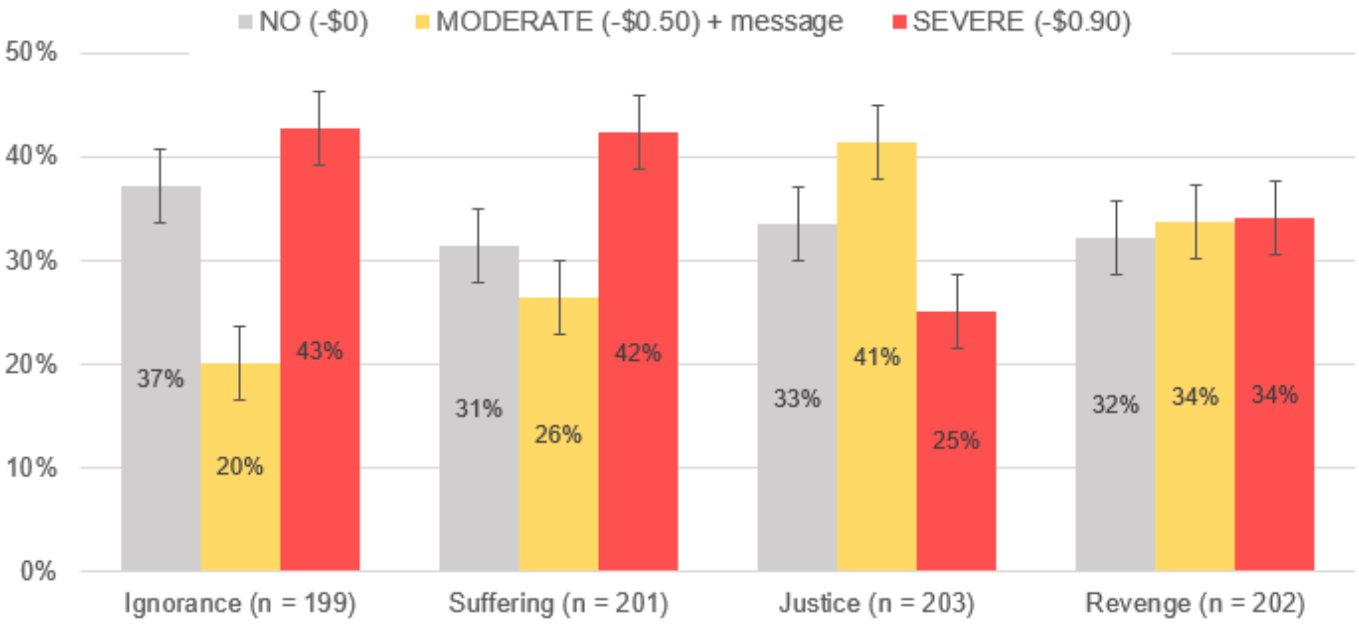


Figure 18: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions in Study 10. Error bars represent ±1 standard error.

Although the suffering condition (26%) was in-between the ignorance (20%) and the revenge conditions (34%), so the results are directionally consistent with the hypothesis that Recipients derive some utility from making the Allocator suffer, neither difference was significant ($\chi^2(1, N = 400) = 1.864$, $p = .172$ between the suffering and ignorance conditions, and $\chi^2(1, N = 403) = 2.217$, $p = .137$ between the suffering and revenge conditions). This result highlights the limited role of pure retributive motives (in absence of belief-based motives). We did not observe a significant difference between the justice (41%) and revenge (34%) conditions either, $\chi^2(1, N = 405) = 2.252$, $p = .133$. If anything, disclosing that the Recipient was responsible for reducing the Allocator's payoff makes the corresponding

message *less* appealing—compared to when the Allocator is informed about the reason of punishment only, i.e., the justice condition. The smaller proportion of Recipients choosing the moderate punishment option in the revenge condition than in the justice condition is consistent with the finding in Study 9, showing that among people who would reduce their partner's payoff *and* send an explanation (i.e., justice and revenge messages, $n = 106$), a significant minority of people ($n = 27$, 25%) prefer *not* to disclose the identity. In that study, participants selected the less informative justice message rather than the revenge message, even though their choice was unconstrained.

Crowding-in and crowding-out of punishment. Next, we conducted regression analyses to see whether the differences in the proportion of Recipients choosing the moderate punishment across conditions was driven by non-punishers "crowding in" (i.e., more Recipients selecting some punishment) or by the "crowding out" of severe punishers (i.e., fewer Recipients selecting severe punishment). To investigate this, we created three dummy variables that correspond to the three pieces of information that were gradually added to the messages in the moderate punishment option. In this way, the coefficients on these terms capture the marginal effect of each of these pieces of information on punishment. The first dummy variable ("Suffer") indicates whether the maximum possible bonus was displayed in the message. This dummy is 0 in the ignorance condition and 1 otherwise. The second dummy ("Explain") indicates whether the message clarifies the reason why the Allocator's payoff has been reduced, and is 0 in the ignorance and suffering conditions and 1 otherwise. Finally, the third dummy ("Identity") indicates whether the Allocator is informed that the Recipient decided to reduce their payoff, and is 1 in the revenge condition and 0 otherwise.

We conducted three separate OLS linear regressions, adding the above dummy variables as independent variables and the likelihood of choosing no, moderate, and severe punishment as dependent variables (see Table 12, Columns 1, 3, and 5). Most importantly, these analyses replicate the results reported in the previous section: Participants were significantly more likely to choose the moderate punishment when the accompanying message clarified the reason for reducing the Allocator's payoff (i.e., $Explain = 1$, in the justice and revenge conditions), $\beta = 0.150$, $t(801) = 3.320$, $p < .001$. Further, participants chose the moderate punishment marginally significantly less when this option also revealed their identity to the Allocator (i.e., $Identity = 1$, in the revenge condition), $\beta = -0.077$, $t(801) = 1.709$, $p = .088$. Adding the maximum possible bonus to the messages, thus merely making the Allocator suffer ($Suffer = 1$), did not lead Recipients to select the moderate punishment more often, $\beta = 0.063$, $t(801) = 1.380$, $p = .168$, which highlights the limited role of pure retributive motives.

Table 12: OLS regressions, Study 10: Likelihood of the Recipient choosing no, moderate, and severe punishment options.

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE punishment		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer[1]	−0.058	−0.057	0.063	0.058	−0.004	−0.0005
	(0.047)	(0.047)	(0.045)	(0.045)	(0.048)	(0.047)
Explain[2]	0.022	0.027	0.150***	0.151***	−0.172***	−0.178***
	(0.047)	(0.047)	(0.045)	(0.045)	(0.047)	(0.047)
Identity[3]	−0.013	−0.016	−0.077†	−0.077†	0.090†	0.092*
	(0.047)	(0.047)	(0.045)	(0.045)	(0.047)	(0.047)
Sex (Female = 1)		0.072*		0.035		−0.107**
		(0.033)		(0.032)		(0.034)
Age (years)		0.003*		−0.002†		−0.001
		(0.001)		(0.001)		(0.001)
Constant	0.372***	0.222***	0.201***	0.271***	0.427***	0.508***
	(0.034)	(0.059)	(0.032)	(0.057)	(0.034)	(0.059)
Observations	805	805	805	805	805	805
R^2	0.002	0.017	0.030	0.035	0.022	0.036
Adjusted R^2	−0.002	0.011	0.026	0.029	0.019	0.030

Note: †$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$

[1] *Dummy: ignorance = 0; suffering = 1; justice = 1; revenge = 1*

[2] *Dummy: ignorance = 0; suffering = 0; justice = 1; revenge = 1*

[3] *Dummy: ignorance = 0; suffering = 0; justice = 0; revenge = 1*

None of the experimental manipulations had a significant effect on the likelihood of choosing no punishment, all $p > .217$, which suggests that having the ability to convey more information did not cause a "crowding in" of non-punishers: The proportion of participants who decided to let the Allocator's unfairness go unpunished remained relatively constant across conditions. By contrast, participants were significantly less likely to choose the severe punishment when they could explain the punishment by choosing the moderate punishment (i.e., $Explain = 1$), $\beta = -0.172$, $t(801) = 3.625$, $p < .001$. In the revenge condition ($Identity = 1$), however, they were marginally significantly more likely to choose the severe punishment, $\beta = 0.090$, $t(801) = 1.911$, $p = .056$. These effects are directionally opposite, and of similar magnitude, to the effects observed for the likelihood of moderate punishment, suggesting that most of the main effects can be explained by the "crowding out" of severe punishers: Once Recipients were able to explain the punishment decision, they enacted a less severe punishment, even though in the absence of an explanation, they would have preferred a severe punishment. This indicates that at least some participants were willing to compromise on distributive justice, to make sure that their partner understood the reason for being punished.

Robustness checks: demographics, anger, and suspicion. In a second set of OLS regressions (see Table 12, Columns 2, 4, and 6) we also included sex and age as demographic controls. Although both sex and age have significant main effects on punishment choices—women are more likely to choose no punishment, while men are more likely to choose severe punishment; older participants are more likely to choose no punishment—the effects of the experimental manipulation are robust to the inclusion of these demographic controls. Furthermore, all of the results are robust if we limit our analyses to those Recipients only who reported that they were angry after seeing the Allocator's decision. Results are also robust to the exclusion of suspicious Recipients (i.e., who did not believe that the Allocator was a human). We report the results of these robustness checks in Appendix VII.3.5.

Alternative explanation 1: "explanation enhances suffering"

Even though the main results strongly support the hypothesis that punishers care about transgressors' beliefs per se, and derive utility from being able to let the transgressors know why they are being punished, the observed pattern of results can be also consistent with an alternative hypothesis. If the transgressor *suffers more* when the punishment is accompanied by an explanation, then the higher willingness to enact a punishment with an explanation is also consistent with retributive motives, i.e., the desire to make the transgressor suffer for their misbehavior. This would imply that punishers who prefer the moderate punishment

in the justice and revenge conditions do not necessarily care about what the transgressor believes per se, but they do want to make these transgressors suffer more by clarifying why they received the punishment. To test whether this alternative hypothesis can explain the main results, we compared Recipients' responses to the post-punishment question "[If you chose this option: ...] Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?" Recall that we asked this question about each of the available options (no, moderate, severe punishment), instead of asking the question only about the option that participants selected, therefore we obtained measures on all three options from everyone (avoiding a self-selection bias).

We found no significant differences in the ratings of the no punishment option between any two options (all $p > .215$, see Figure 19, gray bars). Recipients reported that the Allocator would feel very good in all four conditions if they decided to let them receive the full \$1 bonus (without letting them know anything else about the bonus): $M = 85.4$, 95% CI [81.7, 89.1], $M = 88.0$, 95% CI [84.5, 91.5], $M = 85.4$, 95% CI [81.3, 89.4], and $M = 84.8$, 95% CI [81.1, 88.5] in the ignorance, suffering, justice, and revenge conditions, respectively.

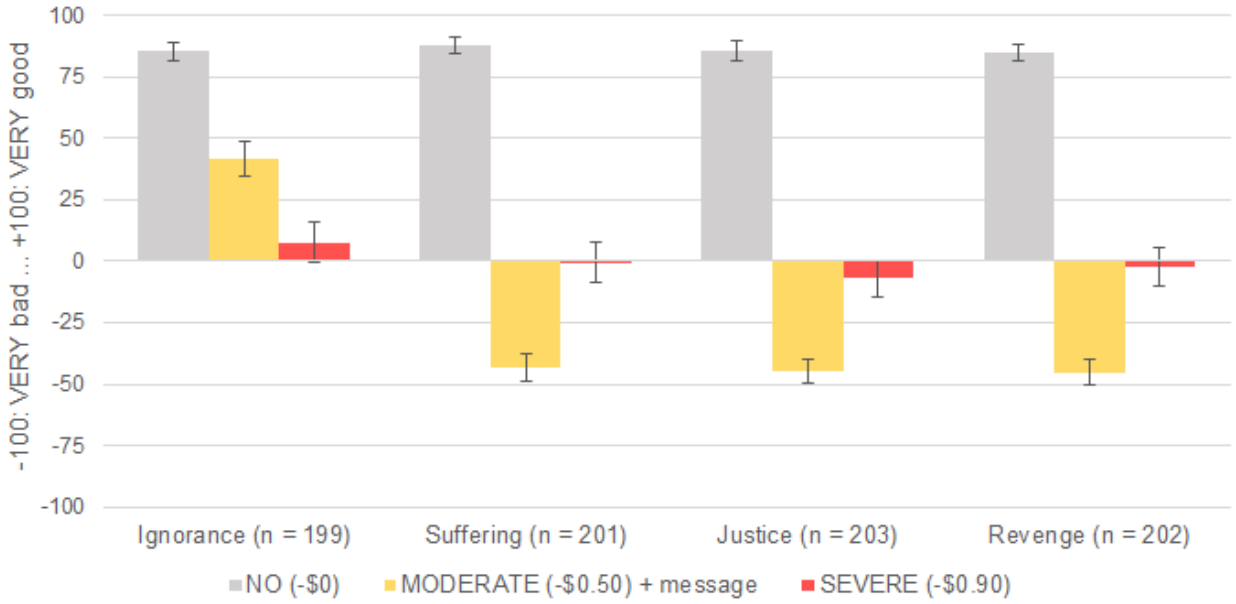


Figure 19: Means of Recipients' responses to the question: "Would YOUR PARTNER feel bad (experience suffering) or feel good (experience joy)?" in Study 10. Error bars: 95% CI.

Similarly, we found no difference in the ratings of the severe punishment between any two conditions (all $p > .074$), except for between the ignorance and justice conditions, where we found a significant but weak effect, $t(400) = 2.633$, $p = .009$, $d = 0.26$ (Figure 19, red bars). More importantly, however, Recipients reported that the Allocator would feel neutral

(i.e., not significantly different from 0) in all four conditions if they received a reduced bonus of \$0.10 (without knowing that the bonus was reduced): $M = 7.66$, 95% CI [−0.34, 15.7], $M = -0.58$, 95% CI [−8.61, 7.46], $M = -7.14$, 95% CI [−14.7, 0.45], $M = -2.59$, 95% CI [−10.4, 5.27] in the ignorance, suffering, justice, and revenge conditions, respectively. This indicates that even though this option reduced the Allocator's material welfare the most, Recipients (correctly) anticipated that their partner would not suffer because they would remain ignorant. By contrast, Recipients reported that the Allocator would feel rather good when receiving the moderately reduced bonus in the ignorance condition ($M = 49.4$, 95% CI [34.8, 48.5]), and significantly better than when receiving the same bonus in any of the other three conditions (all $p < .001$), in all of which they would feel rather bad: $M = -43.2$, 95% CI [−48.6, −37.8], $M = -44.5$, 95% CI [−49.3, −39.7], $M = -45.2$, 95% CI [−50.2, −40.2] in the suffering, justice, and revenge conditions, respectively (Figure 19, yellow bars).

The "explanation enhances suffering" hypothesis implies that Recipients would expect that the Allocator suffers more in the justice and revenge conditions than in the suffering condition. However, we found no such difference in the reported suffering between the suffering, justice, and revenge conditions (all $p > .595$), which indicates that letting the Allocator know the reason why their payoff was reduced was not expected to make them to suffer more. To provide further support against the "explanation enhances suffering" hypothesis, we added Recipients' beliefs about the Allocator's suffering upon receiving the moderately reduced bonus to the regression analyses we conducted in the previous section (see Table 35 in Appendix VII.3.5). Not only does the effect of the Explain dummy remain unaffected by adding this covariate to the model, $\beta = 0.150$, $t(800) = 3.316$, $p < .001$, but the Recipients' belief about the Allocators' suffering is not a significant predictor of the Recipients' likelihood of choosing the moderate punishment, $p = .838$. In other words, participants who believed that their partner would suffer more when enacting the moderate punishment, were not more likely to choose this option, compared to people who thought that their partner would suffer less.

The above results strongly support that the Recipients' higher willingness to choose the moderate punishment in the justice and revenge conditions cannot be explained by their desire to inflict more suffering. This very limited role of pure retributive motives behind punishment decisions is consistent with the finding that introducing the suffering component per se (difference between the ignorance and suffering conditions) did not significantly increase the proportion of participants who chose the moderate option. It is also consistent with the finding in Study 9 showing that among people who would punish their partner ($n = 169$), only a negligible fraction (8%) would make their partner suffer without sending them any explanation.

Alternative explanation 2: preventing future transgressions

The other potential alternative explanation is that Recipients had deterrence motives: They wanted to prevent the Allocator from harming others in the future. While this interaction was one-shot and anonymous—i.e., sending an explanation could not confer any instrumental value to the Recipient—Recipients might still want to send an explanation to the Allocator if they care about how the Allocator behaves towards others in the future and if they believe that sending a message would change the Allocator's future behavior. To investigate whether such considerations could explain the observed results, we compared Recipients' responses to the post-punishment question "[If you chose this option: ...] Would your partner treat others worse / better in the future?" Recall that we asked this question about each of the available options (no, moderate, severe punishment), therefore we obtained measures on all three options from everyone (avoiding a self-selection bias). While Recipients thought that enacting no or severe punishment would not change their partner's future behavior significantly (or make it even worse), they thought that the moderate option would make their partner behave better in the future in all but the ignorance condition (see Figure 20).

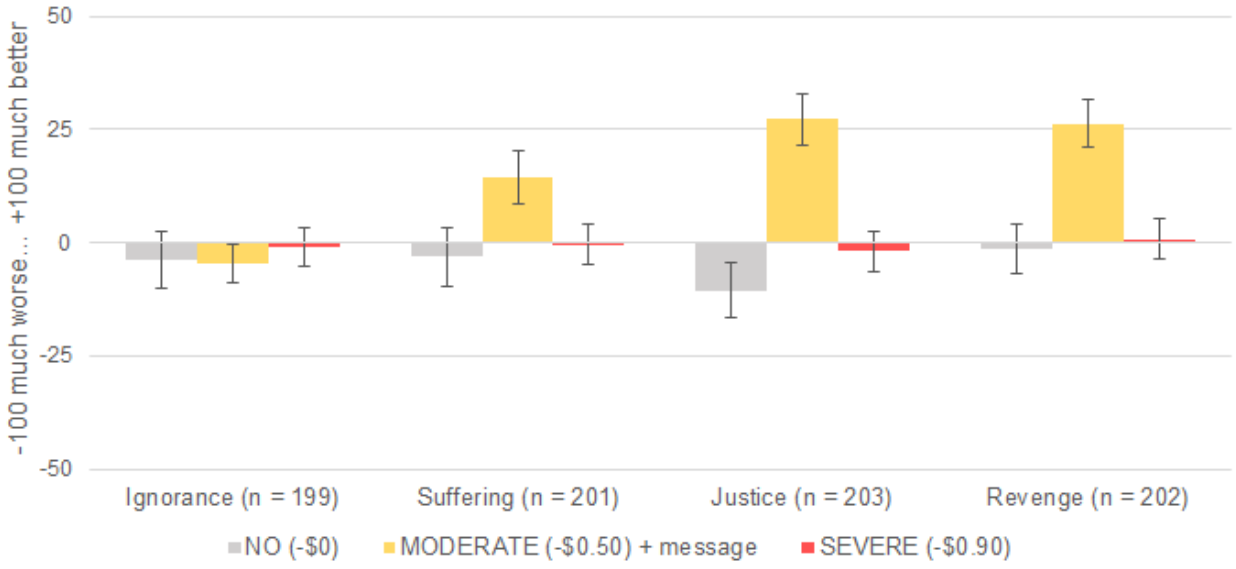


Figure 20: Means of Recipients' responses to the question: "Would YOUR PARTNER TREAT OTHERS WORSE or BETTER in the future?" in Study 10. Error bars: 95% CI.

This suggests that *some* of the increased willingness to choose the moderate punishment in the justice and revenge conditions might be explained by deterrence motives, i.e. to make the other person behave better in the future. To investigate whether this motive was solely responsible for the observed differences between conditions, we added this independent variable to the regression conducted in the previous section. Recipients who thought that

the moderate punishment would make their partner behave better in the future, were more likely to choose the moderate punishment, $\beta = 0.002$, $t(800) = 4.876$, $p < .001$, and including this variable weakens the coefficient of the Explain dummy, which supports the claim that participants were at least partially motivated by deterrence. More importantly, however, the Explain dummy remains significant, $\beta = 0.124$, $t(800) = 2.764$, $p = .006$. We report the details of this regression in Table 36 in Appendix VII.3.5.

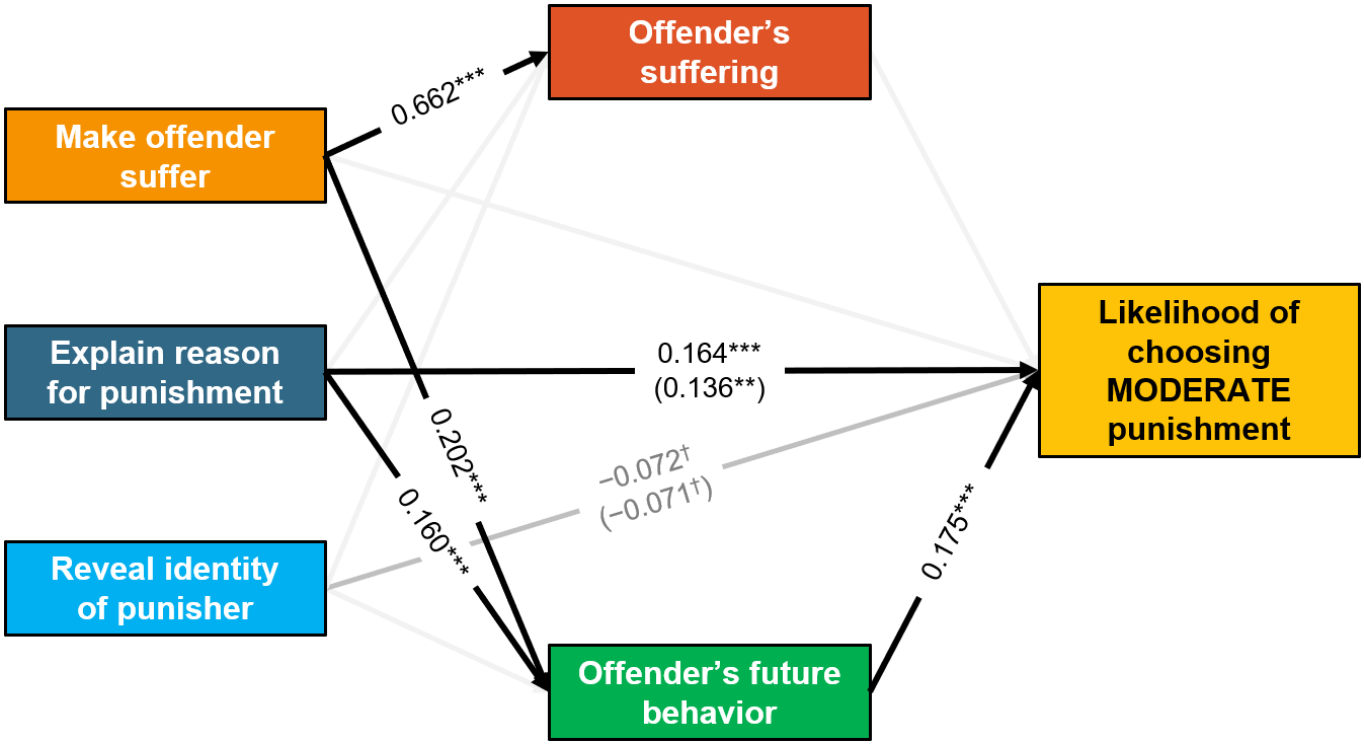
Mediation analysis: combined effects of alternative mechanisms

To investigate the combined effects of the two alternative hypotheses—i.e., "explanation enhances suffering" and deterrence—as well as to establish what proportion of the main effect of explaining the reason for punishment these alternative mechanisms can account for, we conducted a mediation analysis (see Figure 21). We standardized all variables and we also controlled for the sex and age of participants. The outcome variable was the (standardized) likelihood of choosing the moderate punishment, whereas the predictor variables were the three dummy variables (Suffer, Explain, Identity) that represented the information gradually introduced in experimental conditions. We then added two potential mediators: 1) the punisher's beliefs about the offender's suffering if the punisher chose the moderate punishment, and 2) the punisher's belief about the offender's future behavior (whether they will behave better or worse) if the punisher chose the moderate punishment.

A bootstrapped mediation with $5,000$ replications revealed that the punisher's belief about the offender's future behavior partially mediates the effect of explaining the reason for punishment, $\beta = 0.028$, $SE = 0.01$, $p < .001$. More importantly, however, the direct effect of explaining the reason for punishment remained significant after introducing the two potential mediators, $\beta = 0.136$, $SE = 0.05$, $p = .005$, which is 83% of the original (i.e., non-mediated) effect. This shows that *most* of the effect of sending an explanation cannot be explained by participants' desire to change the future behavior of transgressors: Participants have a preference for enacting a punishment with an explanation, beyond merely wanting to deter their partners from transgressing again.

Furthermore, this mediation analysis confirms earlier findings about the weakness of pure retributive motives. Even though participants knew that offenders would suffer more if they revealed the highest possible bonus the offender could have gotten (i.e., choose the moderate punishment in all but the ignorance condition), this increased suffering associated with the moderate punishment *did not* explain punisher's willingness to choose the moderate punishment at all, $p > .10$. In fact, the only effect that revealing the maximum potential bonus (i.e., making the offender suffer) had on the likelihood of choosing the moderate punishment, was through an indirect effect mediated by punishers' belief about offenders' future

behavior, $\beta = 0.035$, $SE = 0.01$, $p = .005$. This shows that the slightly higher willingness to choose the moderate punishment in the suffering condition than in the ignorance condition (see Figure 18) can be *fully* explained by punishers' deterrence motives and their (incorrect) belief that the offender would behave better in the future if they learned that they received a reduced bonus (but without learning the reason why).



Note: $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 21: Mediation analysis, Study 10. Coefficients are standardized Beta coefficients. For better visibility, non-significant paths ($p > .10$) are hidden.

IV.7 General discussion

Across three studies—one observational, one hypothetical choice, and one real choice—we show that punishers take into account what a transgressor thinks after being punished, and that these motives affect the punisher's decision after taking account of purely distributional preferences (i.e., the transgressor's material outcome), the transgressor's suffering, and the potential deterrence effects of the punishment. Our results strongly support the idea that punishment decisions are at least partially motivated by *belief-based preferences*—the punisher's preferences over what the transgressor believes—and that these preferences can often dominate distributional preferences. Specifically, punishers have a strong desire for transgressors to understand that they have been punished (i.e., are experiencing a negative outcome) and for what reason (i.e., as payback for a specific bad behavior in the past). These results are consistent with both the understanding hypothesis (French, 2001; Miller, 2001; Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011; Funk et al., 2014) and the expressive function of punishment (Feinberg, 1965; Masclet et al., 2003; Xiao & Houser, 2005; Sarin et al., 2020), but more importantly, our paper makes four novel contributions and addresses several limitations and open questions of prior work.

First, as summarized in the previous paragraph, we demonstrate a direct causal link between belief-based motives and actual punishment decisions, whereas existing empirical work could provide only correlational evidence between post-punishment satisfaction and the availability of communication. Second, we not only find evidence for the existence of purely belief-based motives behind punishment but also show that these can be even stronger than the desire for retribution, and that some people are willing to compromise on distributive justice to make sure that the offender understands the circumstances of punishment. Third, we address and disentangle the "reason and source" confound (i.e., when the offender learns the reason for punishment, they inevitably learn the source of punishment as well), which had been present in previous work. We show that most of the effects described in the first two points are driven by the motive to reveal the reason for punishment—but not necessarily the source of it. Finally—to our knowledge—the present paper is the first to gather empirical support for the understanding hypothesis in both a real-world observational study and an incentivized experiment featuring real-time interaction between real participants. Our findings also highlight the possibility that most of the previous work likely overestimated the desire for pure retribution or the restoration of distributive justice, because virtually all "classic" experiments were confounded by belief-based motives: Participants—both victims, transgressors, and punishers—had perfect information about the potential payoffs and their actual outcomes (e.g., Fehr & Gächter, 2002 and its replications;

Fehr & Fischbacher, 2004; Henrich et al., 2006). This means that whenever an offender was punished for his or her misbehavior, it was immediately obvious to him or her not only that his or her payoff had been reduced, but also *why* it had been reduced. Furthermore, even in anonymous settings, transgressors were fully aware that their partners (or fellow group members) were responsible for punishing them. Yet, past research has identified retribution as the main driver of punishment decisions in these situations, even though punishers' decisions could have been influenced by considerations for what offenders believe, regardless of retributive or deterrence motives. To our knowledge, Study 10 of the present paper is the first experiment that clearly disentangles these three potential motives behind punishment—material, affective, and cognitive.

The most closely related study that aimed to disentangle distributive preferences from other considerations is by Crockett et al. (2014), who conducted a 3-player one-shot Trust Game in which a third-party observer could punish an untrustworthy participant (who was given funds but did not return them). The authors manipulated whether the punishment was *open* or *hidden*—i.e., whether the transgressor would realize that they had been punished by their payoff being reduced by the third party—and found that people were significantly more willing to punish in the open condition. The authors conclude that the preference to communicate norms through punishment plays an important role for punishment decisions. However, their experimental design did not force participants to make any trade-offs with respect to the various motives behind punishment, and it did not allow for disentangling the comparative suffering hypothesis from the understanding hypothesis: The higher likelihood of punishing transgressors in the open condition of their study is consistent with both a preference for retributive justice and belief-based preferences. By contrast, in our experiment we isolated retributive motives from belief-based preferences, and demonstrated that most of the difference between open and hidden punishment is driven by belief-based preferences, and not by retributive motives. Furthermore, we also show that once we allow participants to choose punishment options that provide an explanation to the transgressor (and thus, provide additional belief-based utility for the punisher), we observe a *crowding-out* effect of severe punishers, but no crowding-in effect of people who would otherwise not punish.

These findings suggest that the availability of explanations can increase overall social welfare by reducing the severity of material punishments, which has valuable implications for institutional policies. For instance, this insight can help in addressing the escalation of punitive damages in the legal system, which legal scholars have argued are getting out of hand (e.g., Huber, 1989, Sunstein, Kahneman, & Schkade, 1998). If victims are provided alternative ways of communicating messages to transgressors, through, for example, mediation or victim impact statements, they may be willing to settle legal suits for lower

amounts or to ask for lower punitive damages. This is consistent with evidence showing that providing legal cover for transgressors and victims to communicate to each other (i.e., through "apology laws") is associated with faster settlements and lower settlement amounts (B. Ho & Liu, 2011). Allowing victims to send a message to the offender could also elevate perceptions that procedural justice was served and thus reduce revenge and increase forgiveness (Aquino, Tripp, & Bies, 2006). Understanding what motivates people to enact revenge also has practical, managerial implications. If, for example, workers are motivated to affect the misbehaving colleague's beliefs, rather than merely wanting them to suffer, then managers could provide platforms for their employees to openly express and communicate their discontent to the transgressor (without the fear of retaliation), to reduce employees' desire to engage in costly vendettas and other kinds of unproductive behavior. The same insight could be also applied to reduce the potential harms caused by public complaints, negative word-of-mouth, and other forms of customer retaliation. For example, as a way of decreasing customer desire to take revenge by writing negative reviews, firms would offer customers the opportunity to send the company a (private) message directly, *before* asking them to write a public review on their product or service (Grégoire, Laufer, & Tripp, 2010).

Finally, this work expands our understanding of belief-based preferences to include preferences over second-order beliefs, i.e., the beliefs of others, that go beyond image motivation and strategic considerations (e.g., Battigalli et al., 2019). Most previous work on belief-based preferences has examined preferences for cognitive states about material outcomes affecting the self, such as whether one has an untreatable illness (Oster et al., 2013; Ganguly & Tasoff, 2017) or whether one's stock portfolio has decreased in value (Karlsson et al., 2009). Research on image motivation (e.g., Ariely, Bracha, & Meier, 2009; Soetevent, 2011) has found that people also have a preference for certain cognitive states of others, but specifically with respect to image. That is, people care what others think about them.

People's concern about what goes on in other people's minds, however, goes well beyond the desire for others to hold a positive image of them, and has diverse consequences for interpersonal relations, politics and economics. For example, research in psychology suggests that people not only have a desire to be perceived positively, but also realistically (as perceived by themselves), even if this implies a negative image (W. B. Swann, Pelham, & Krull, 1989; W. B. Swann, 2011). Our paper finds evidence for an additional interest in others' beliefs that seems to be unrelated to what others perceive about them: People care about what transgressors think about punishment they receive. We have demonstrated that punishers not only want offenders to have a particular understanding of their outcomes, but they are even willing to compromise on distributive and retributive justice to fulfill this goal.

This page intentionally left blank

"As it was, we always misunderstood
ourselves and rarely understood others."

— *Oscar Wilde* (1891)

Chapter V

Mind the Gap (Between Minds)

In this dissertation, I have gathered and presented consistent evidence from a set of ten studies, featuring a broad range of empirical methods and contexts, demonstrating that people directly care about what others *believe*. The pattern of attitudes, preferences, and behaviors documented in the present work is incompatible with standard economic models that assume that people care about others' beliefs only to the extent to which knowing about others' beliefs can improve the quality of decision-making. The contexts studied here were devoid of any such instrumental or strategic aspect, thus standard economics falls short of explaining the results I obtained.

V.1 Summary of findings

In Chapter II, I offered a new alternative to the conventional theory of belief homophily—according to which people have an intrinsic distaste for encountering differences in beliefs, and which is often seen as the main culprit for selective self-exposure to concordant views ("echo chambers") and geographic segregation by political ideology. Instead, I argued that when people face others who hold beliefs different from their own, they do not find these encounters disturbing because others hold different beliefs *per se*, but because they are convinced that others hold *false beliefs*. In five pre-registered studies ($N = 2,835$) featuring self-recalled personal experiences and vignette scenarios, I demonstrated that participants express stronger negative feelings when they are convinced that others hold false beliefs, compared to cases in which others' beliefs are different from their own, even when participants' objective knowledge about the validity of beliefs and the potential negative consequences of holding false beliefs are held constant. I also showed that higher confidence that others hold false beliefs—but not different beliefs—evokes stronger negative emotions and triggers avoidance behaviors. Finally, I demonstrated the moderating role of the identity of the person who might be affected by negative consequences: People are primarily disturbed when false beliefs could affect someone who they care about. These findings highlight the possibility

that most of the behaviors that have been previously attributed to belief homophily, i.e., the desire to reduce differences between one's own and others' beliefs, have been misattributed, and should be attributed to the desire to avoid others holding false beliefs.

In Chapter III, I focused on situations in which making the right choice involves selecting the "lesser of two evils," however, only seeing the chosen option can lead others to misunderstand the decision maker's intentions. I investigated the desire for revealing contextual information to a partner when such information clarifies or corrects the partner's beliefs about the interaction but has no further consequences to the decision-maker (i.e., does not affect outcomes). In two experiments ($N = 448$ pairs), I showed that people are willing to pay ex post to reveal their choice set to their partner, even after a one-shot anonymous interaction with no reputational consequences, and in some cases even when doing so reveals their selfish intentions. I found that the choice to reveal is driven by concern for the beliefs of strangers, but only when revealing signals generous intentions; those who reveal a choice that appears selfish report doing so out of a desire to be or appear honest. And though some people leave a misunderstanding in place when it is self-enhancing to do so, almost no one is willing to *create* a misunderstanding (by hiding the other option), even when it could conceal selfish behavior.

Finally, in Chapter IV, I tested the idea that people care directly about what others believe in the context of moral behavior and punishment—that people care about what transgressors think about punishment they receive, and specifically that they understand they have been punished, and why. Across three studies, including an observational study featuring real-world revenge stories and a high-powered pre-registered experiment involving a stylized workplace interaction, I examined whether such belief-based preferences play a crucial role in punishment decisions. Results from the studies not only demonstrated a direct causal link between belief-based motives and actual punishment decisions but showed that these motives can be even stronger than the desire for retribution. In fact, I found very little evidence for purely retributive preferences (to inflict suffering on transgressors), and showed that a substantial fraction of subjects are willing to compromise on distributive justice (to restore a fair allocation of wealth) to make sure that the offender understands the circumstances of punishment. These effects are robust to a variety of controls (e.g., demographics, anger, suspicion), and I also demonstrated that the preference for affecting offenders' beliefs cannot be explained by deterrence motives only (to make transgressors behave better in the future)—punishers glean value directly from affecting the beliefs of the transgressors in this way. Thus, people punish others not only because they want to achieve something instrumental (i.e., reduce the transgressor's welfare or deter future offenses), but also because they want to affect what offenders believe.

V.2 Limitations and directions for future research

Of course, all scientific research, especially the kind that sets out to propose novel concepts and alternative approaches, is bound to have its own limitations; this dissertation is no exception. Here, I focus on two broad types of limitations—or, rather, areas for improvement—of the present work:

1. A better and more nuanced understanding of the precursor (and boundary) conditions and mechanisms that determine when and why people care about what others believe (i.e., the "psychologist's perspective");

and

2. A better understanding of the applicability and generalizability of these insights; whether belief-based motives are powerful enough to affect real choices in a meaningful way (i.e., the "economist's perspective").

That is, there is a lot of room for improvement, in terms of understanding both the *causes* and the *consequences* of people's propensity to care about what others believe. In the rest of this Chapter, I discuss the above limitations in more detail, and highlight potential avenues for future research. While the present work can be regarded as preliminary in many ways, hopefully it will inspire, and pave the way for, a multitude of future projects.

V.2.1 The nuts and bolts of beliefs: pre-cursors and mechanisms

Though I theorized about, and provided some preliminary evidence for, the underlying psychological constructs and contextual factors that determine when, and to what extent, people care about others' beliefs (e.g., social proximity, potential consequences), admittedly, the present work should be considered as a proof-of-concept, rather than a comprehensive review of *all* potential mechanisms. The aim of this dissertation is to highlight the discrepancies between the predictions of standard economic theory and actual human behavior, when it comes to interacting with others whose beliefs might differ from one's own, creating potential misunderstandings—a *gap between minds*. That is, I focus on demonstrating the *existence* of such purely belief-based considerations, rather than trying to understand all relevant elements that trigger, amplify, attenuate, or moderate these motives. However, understanding what exactly drives people to care about what's going on in others' minds, and identifying the contextual factors that play a vital role in these processes, is the logical and necessary next step along this research agenda.

The role of perspective-taking and mind reading

I see several avenues along which my work can be extended, and built upon, mainly by combining these insights with theories and experimental paradigms in social psychology that investigate perspective taking, empathy, and mind reading. One necessary pre-cursor for the phenomena I investigate in my dissertation is the ability to infer others' mental states. However, this ability is neither universal nor fixed: People vary drastically in how easily and accurately they can read others' minds (e.g., Crone, Bullens, van der Plas, Kijkuit, & Zelazo, 2008; Davis, 1983; Van der Graaff et al., 2014), and mind reading skills can be improved with training, both during adolescence (Chandler, 1973; Goldstein & Winner, 2012) and adulthood (Marangoni, Garcia, Ickes, & Teng, 1995). What if, for instance, people who are good at mentalizing would feel more disturbed when encountering others who hold false beliefs, as opposed to people who are less proficient mind readers? Better perspective-taking ability is usually associated with better outcomes in conflict resolution (e.g., Galinsky, Ku, & Wang, 2005; Leith & Baumeister, 1998; Stephan & Finlay, 1999), but what if there is also a negative side of being able to put ourselves in someone else's shoes, that could make us *more* upset, because it enables us to realize *how different* our views are? These are empirical questions, and can be tested, both in laboratory settings and in the field.

Boundary conditions: psychological (and social) distance

Beyond the ability to read minds—which is a necessary but not sufficient condition for caring about beliefs—what are the contextual factors and cues that trigger caring about others' beliefs? While my work demonstrates that even relatively subtle experimental manipulations can elicit such motives (e.g., a misunderstanding of the intent behind allocating a few dollars during an anonymous online interaction), we certainly do not care about *all* beliefs that others hold. There are countless others out there who firmly hold beliefs and narratives that we would deem blatantly false, and at least hundreds of people who might have a different *view of us* than the image we construe for ourselves; yet, we seem to ignore *most* of these differences *most* of the time. So, what are the boundary conditions that are necessary for making people care about what others believe? For instance, are there simple spatiotemporal effects that drive how much people care about others' beliefs, e.g., by paying more attention to the beliefs of those who are closer to them, both physically and in time? That is, the more *psychologically distant* someone seems to be (see Trope & Liberman, 2010), the less we care about what they believe? This could be a potentially promising candidate mechanism, since there is ample of evidence that people process psychological distance automatically (e.g., Bar-Anan, Liberman, Trope, & Algom, 2007), which could allow them to rely on cues of psychological distance to determine whose beliefs they should be paying attention to.

What about group belonging? Do people tend to monitor the mental states of "in-group" members, who they share some similarities with? Is it necessary to have some degree of (indirect) instrumentality—as we demonstrated in Study 4—that the person holding false beliefs might take an action that might affect someone we care about? For instance, do Americans care about what fellow citizens believe, but not so much about what foreigners do? If so, could this effect be explained by simple instrumental considerations (i.e., a compatriot is more likely to affect one's well-being in some way than a foreigner), or would this be beyond instrumentality? (If the group is large enough, such as an entire nation, and its members are rather distant from each other, then probably it *is* beyond any instrumentality).

Incidental internal processes: mood and emotions

In addition to psychological and social distance (i.e., mostly external factors), are there *internal* psychological processes that are incidental to belief-formation, but may nevertheless trigger caring about beliefs (e.g., mood, emotions, drive states)? For instance, according to the "feelings-as-information theory" (Schwarz, 2012), people in a sad mood tend to rely more on bottom-up processing, by paying more attention to details and adopting a more analytical mindset. Similarly, would people in a bad mood tend to pay more attention to others' mental states—and more readily conclude that these others hold incorrect beliefs? Indeed, there is some evidence that, when experimentally induced, sadness increases perspective-taking and theory of mind abilities (Converse, Lin, Keysar, & Epley, 2008). But then, could this effect drive a vicious cycle of experiencing negative feelings and being cognizant of others' false beliefs?: If negative mood triggers people people to pay more critical attention to what others believe, they will more likely identity falsities, which could make them feel ever worse.

Moralizing minds

Another limitation of the present work is that each context that I conducted experiments in, involved some aspect of morality (i.e., falseness, altruism, honesty, moralistic punishment). This raises the question whether having a moral component is also a necessary condition for caring about what is going on in others' minds. Do people have intrinsic preferences for what others (should) believe only in those situations which are related to morality? Would people still be upset when encountering false beliefs, if they could not pass any moral judgment on their beholder (e.g., because the other person was bound to hold that belief, given circumstances)? Would they be still motivated to disambiguate misunderstandings, when these conflicts in beliefs are not related to morality at all? Would people care *more* about someone's belief once the morality of the context is highlighted (e.g., in religious or otherwise "morally salient" settings)?

The desire to be understood

Throughout this dissertation, I frequently referred to "the desire to be understood," which can be loosely described as the motive to be perceived by others in the same way how one perceives oneself, including intentions, beliefs, and feelings (see self-verification, Sedikides, 1993; W. B. J. Swann, Rentfrow, & Guinn, 2003). While this concept is closely related to the propensity of caring about what others believe, it is also broader, as it encompasses other (non-belief) psychological constructs, such as emotions. However, more importantly, to what extent is our preference for others to have a particular set of beliefs stem from our need to be understood? That is, are these motives simply yet another manifestation of the search for self-verification, or are they fundamentally different? Although the present work cannot answer this question directly, I suspect that these accounts can potentially be empirically distinguished (i.e., caring about others' beliefs vs. self-verification).

V.2.2 Consequences: applications and generalizability

Admittedly, one major limitation of the present work is its external validity and generalizability. All but one of the ten studies reported here were conducted in some kind of online laboratory setting: We either asked participants to recall real events from their own lives, or we put them into an experimentally manipulated situation and observed their reactions and behavior. Only Study 8 relied on "field" data (self-reported revenge stories on Quora), but even that context had limited practical relevance (beyond entertainment value). Therefore, one obvious way how the present work could be extended, and improved upon, would be to look at whether the same phenomenon—that is, considerations for others' beliefs affect choices in a meaningful way—occurs in real word settings.

While in our studies we demonstrated that punishers are willing to make a trade-off between sending a message and enacting a harsh punishment, the stakes in these online experiments were just a few dollars, and the initial offense was relatively mild: allocating a bit more of a simple task to another participant (i.e., only a few minutes of extra effort). But in the real world, vengeful behavior is often triggered by much more severe and consequential offenses (e.g., cheating on someone's partner, damaging someone's property, sabotaging someone's work). By moving away from abstract laboratory settings, one could look at actual punishment behavior in the workplace or in the legal system. Would the "victims" of these far more malicious crimes and actions also be willing to make a compromise on the severity of the punishment, if by doing so they can ensure that the offender will *understand*?

In the concluding section I review three areas of potential applications, in which gaining a deeper understanding of the causes and consequences of caring about others' beliefs might prove to be essential: 1) political behavior; 2) mental health; and 3) the future of work.

V.3 Applications: fear and loathing of other minds

Perhaps the most obvious application of the concepts outlined in this dissertation—belief-based utility and caring about others beliefs—is political preferences and ideological polarization. As highlighted in the general discussion of Chapter II, if we can improve our understanding of what exactly makes people upset when encountering differences in political views (and as a result, secluding themselves in echo chambers and avoiding others who disagree with them), then we can implement interventions and re-design institutions in a way that would combat belief polarization, and improve the quality of public discourse—and with it, strengthen democratic processes.

V.3.1 We all got left behind: anti-establishment sentiments

A highly relevant application in the field of political science could be investigating the "feelings of being left behind," and how it relates to caring about what others' believe. It has become a trope among political pundits to associate the feeling of being left behind with reactance and anti-establishment sentiments: People who feel that they had been left behind and disenfranchised by their leaders—and society, in general—are more likely to support candidates who promise a drastic overhaul of the status quo. The conventional explanation for this phenomenon is almost purely economic, that is, people were left behind *economically*, and are facing *economic hardship*, thus demand systemic change.

Indeed, this narrative is so compelling, that it has emerged as one of the primary explanations for Donald J. Trump's unanticipated win in the 2016 U.S. presidential election (Porter, 2016). However, as other researchers have pointed out, there is little connection between economic hardship and support for Trump, and the feeling of left behind has more to do with subjective perceptions of social status (e.g., Mutz, 2018). As the sociologist Robert Wuthnow puts it his book, in which he investigates the woes of rural America:

> *"Washington [D.C.] is distant from their communities, geographically and culturally. As far as they can see, the federal government hasn't the least interest in trying to understand rural communities' problems [...] The sentiment is so pervasive, so vehement, that it is hard to get underneath it so see what it means."*
>
> (Wuthnow, 2019, p. 98)

What Wuthnow highlights is the challenge of identifying what exactly triggers the feeling of being misunderstood. Perhaps, this is another manifestation of the negative consequences of caring about what others believe. With the rapid acceleration of communication technologies and the emergence of the Internet, now the entire world is in our living room—and you could be exposed to drastically different beliefs day and night, *especially* online. This could be a perfect hotbed for constantly encountering ideas that we deem blatantly false (including "fake news"), that only provide fuel to existing ideological conflicts. Furthermore, feelings of left behind (and misunderstood) might also drive people towards extremist groups, which lure in new supporters by promising them instant connection to like-minded individuals and a strong sense of belonging. For instance, a meta-analysis on violent extremism (both jihadist radicalization and Western extremism), found that the most important personal factor that explains why someone becomes prone to radicalization, is the individual's overall *mental health*, and the key psychological issues are described, among others, in terms of personal alienation, isolation, loneliness, and misfit (Vergani, Iqbal, Ilbahar, & Barton, 2020). Therefore, there is a possibility that the effects of belief-based considerations on political behavior are mostly indirect, and are primarily driven by indirect effects through mental health.

V.3.2 Misunderstood, morose, and miserable: the mental health consequences

According to a recent large-scale review on the correlations between the quality of social relationships and physiological determinants of longevity in representative samples of U.S. adults,[1] researchers found that the majority of adults, between 50–60%, do not feel understood by even those *who are the closest to them*, that is, by their spouses and close family members (Yang et al., 2016). This stylized finding holds in all demographic groups, regardless of gender, age, or ethnicity—there seems be an epidemic of misunderstanding raging across the country. Similarly, survey data suggests that the feeling of loneliness in the U.S. has also reached historic levels in recent years (Cigna, 2018). Other work demonstrated strong links between the feeling of being understood and greater life satisfaction, even during time periods as short as two weeks (Lun, Kesebir, & Oishi, 2008), and showed that felt misunderstanding even affects the perception of pain (Oishi, Schiller, & Gross, 2013). Moreover, studies investigating the neural bases of feeling misunderstood identified regions related to negative affect and social pain (i.e., anterior insula), which suggests that feeling misunderstood evokes strong emotional responses (Morelli, Torre, & Eisenberger, 2014).

[1] MIDUS: aged 25–64, 1995–96 and 2004–2009; HRS: aged 50 and above, 2004–2006

Taking these into account, it does not sound far-fetched to hypothesize about the mental health consequences of being exposed to, and paying attention to, different views and beliefs. Furthermore, as I highlighted in Section V.2.1 (on mechanisms), the feeling of being understood is closely related to the phenomena I investigate in the present work. Therefore, understanding the links between when and why people care about what others believe, and studying how they react when they encounter differences, or falsities, in others' beliefs, could be key for addressing this ever-growing epidemic of misunderstanding (and loneliness), that might contribute to the overall poor mental health conditions prevalent in our society. Furthermore, it could be also worthwhile to investigate what individual psychopathologies contribute to a stronger focus on what others believe (e.g., neuroticism, social anxiety, depression).

The future of work: A.I. think, therefore A.I. am

Finally, another interesting application of the concepts covered in this dissertation is the "future of work" (i.e., artificial intelligence, automatization, and job displacement). There has been a recent expansion of research that looks at the multitude of effects of A.I., automatization, and job displacement, focusing primarily on the economic impact of these phenomena (e.g., Acemoglu & Restrepo, 2018, 2020), as well as highlighting their legal implications (e.g., Chander, 2016; Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018). However, there is much less emphasis on the potential psychological effects that will inevitably occur when people will be displaced or will have to interact with these new technologies. For example, if belief-based motives impact people's well-being and choices, then what will be the psychological consequence of relying more and more on non-human interactions? Machines—to our current and best knowledge—do not have mental states, but in the near future they might *appear* to have one.[2] Would people react differently when interacting with advanced an A.I., which could mimic behavior that seems to be driven by beliefs and intentions? Would people adapt to these new situations, and start caring about what machines "believe," as if they were friends and colleagues? For instance, would people become frustrated when a machine would disagree with them, or when it wouldn't seem to understand their beliefs—just like in those interactions with other people that I highlighted in Chapter II? If so, researchers who are developing new technologies in the field of robotics and human-computer interaction should take into account that users also have certain belief-based preferences, and might prefer their machine peers to fulfill these needs.

[2]In fact, that moment might not even be that distant from the present, given the rapid advances in the fields of robotics, engineering, and human-computer interaction (see e.g., Traeger, Strohkorb Sebo, Jung, Scassellati, & Christakis, 2020).

Another aspect of the future—or even the present—of work, that concerns beliefs, is the ever-increasing need for specialization of workers. In the modern society, in order to gain comparative advantage, people have to acquire progressively specialized skills, and pursue increasingly more specialized careers—a phenomenon that has been termed as "hyperspecialization" (Johns, Laubscher, & Malone, 2011). The philosopher Elijah Millgram, rather aptly, drew a parallel between this phenomenon and the myth of the Tower of Babel, and argued that society is facing similarly dire consequences if the negative externalities of hyperspecialization remain unaddressed (Millgram, 2015). But hyperspecialization has implications to belief-based motives as well: Once people become more and more specialized, it is increasingly more difficult for them to have shared beliefs and a common understanding of the world with others. A few centuries ago, when the vast majority of people worked in agriculture, completing the exact same routine tasks every day, finding common ground was easy. By contrast, in modern society, with the increasing demand of unique and niche skills, it is increasingly more likely that people are surrounded by others who live completely different lives and experience drastically different realities—simply because of their divergent career paths and distinct qualifications. This poses another challenge to society: How can we combat the epidemic of misunderstanding, and its dreadful consequences, if this phenomenon is deeply rooted in *what we do* and *how we do it*, that is, in the very nature of modern work?

I expect that addressing the issues outlined in this chapter will be immensely complex, but in order to take the first steps towards any satisfactory resolution, we must develop a more nuanced understanding of when and why people care about what others believe around them, for which the present dissertation will, hopefully, pave the way.

Chapter VI

Epilogue

"The purpose of economics, as stated in Paul Samuelson and William Nordhaus's popular textbook Economics, *is 'to improve the living conditions of people in their everyday lives.' While the authors associate 'living conditions' with material concepts such as gross domestic product or income, this traditional economic concept of consumption describes only a fraction of what brings pleasure and pain. Instead, we dwell on successes and failures, the past and the future, relationships, fears, regrets, disappointments and triumphs, whether we have fulfilled our goals, and whether other people like and respect us.*

[... but] is utility all in the mind? Samuelson and Nordhaus state that 'centuries of human history also show that warm hearts alone will not feed the hungry or heal the sick.' While true, this claim underestimates the importance of belief-based approaches and fails to shed light on many of the most significant ills of contemporary life: political polarization; spread of distrust, misinformation and ignorance; transforming labour markets; drug abuse and mental health problems. To address these issues, economics must wake up from its centuries-long diversion and re-embrace the Benthamite concept of belief-based utility."

— *Excerpt from* Loewenstein & Molnar, 2018

This page intentionally left blank

Chapter VII

Appendices

VII.1 Appendix for Chapter II

VII.1.1 Principal Component Analysis, Study 1

Table 13: Correlation coefficients (Pearson's r) between emotions (p values in parentheses)

	Upset	Frustrated	Disturbed	Calm	Relaxed
Frustrated	0.741 (< .001)				
Disturbed	0.676 (< .001)	0.680 (< .001)			
Calm	−0.634 (< .001)	−0.530 (< .001)	−0.455 (< .001)		
Relaxed	−0.604 (< .001)	−0.585 (< .001)	−0.533 (< .001)	0.764 (< .001)	
Happy	−0.526 (< .001)	−0.554 (< .001)	−0.486 (< .001)	0.588 (< .001)	0.640 (< .001)

Table 14: Importance of components, Study 1

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	2.001	0.861	0.677	0.578	0.516	0.447
Eigenvalues	4.005	0.741	0.458	0.334	0.267	0.195
% of VAR explained	0.667	0.124	0.076	0.056	0.044	0.032
Cumulative % VAR	0.667	0.791	0.867	0.923	0.967	1.000

Table 15: Component loadings, Study 1

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Upset	−0.428	−0.283	−0.306	−0.380	−0.494	−0.505
Frustrated	−0.418	−0.385		−0.473	0.596	0.306
Disturbed	−0.390	−0.528	0.111	0.708	−0.156	0.180
Calm	0.406	−0.471	0.448		0.243	−0.595
Relaxed	0.422	−0.399	0.165	−0.334	−0.509	0.516
Happy	0.384	−0.336	−0.814	0.139	0.241	

VII.1.2 Mean emotion ratings, Study 2

Table 16: Mean emotion ratings, Study 2

	Recalled a situation involving...		
Emotion	different beliefs M [95% CI]	incorrect beliefs M [95% CI]	Test statistics
Disturbed	54.00 [49.37, 58.63]	67.70 [63.78, 71.62]	$t(385) = 4.425$, $p < .001$ Cohen's $d = 0.44$
Frustrated	61.43 [57.04, 65.81]	73.31 [69.51, 77.11]	$t(388) = 4.013$, $p < .001$ Cohen's $d = 0.40$
Upset	51.76 [47.40, 56.12]	63.95 [59.92, 67.97]	$t(393) = 4.029$, $p < .001$ Cohen's $d = 0.40$
Calm	43.52 [39.43, 47.6]	36.88 [32.59, 41.16]	$t(395) = 2.197$, $p = .029$ Cohen's $d = 0.22$
Happy	29.69 [25.79, 33.59]	20.04 [16.52, 23.55]	$t(391) = 3.604$, $p < .001$ Cohen's $d = 0.36$
Relaxed	36.52 [32.39, 40.65]	29.78 [25.6, 33.96]	$t(396) = 2.249$, $p = .025$ Cohen's $d = 0.23$
Affective valence	0.20 [0.06, 0.34]	−0.20 [−0.33, −0.07]	$t(395) = 4.127$, $p < .001$ Cohen's $d = 0.41$

VII.1.3 Principal Component Analysis, Study 2

Table 17: Correlation coefficients (Pearson's r) between emotions (p values in parentheses)

	Upset	Frustrated	Disturbed	Calm	Relaxed
Frustrated	0.731 (< .001)				
Disturbed	0.761 (< .001)	0.696 (< .001)			
Calm	−0.625 (< .001)	−0.546 (< .001)	−0.514 (< .001)		
Relaxed	−0.653 (< .001)	−0.607 (< .001)	−0.573 (< .001)	0.842 (< .001)	
Happy	−0.531 (< .001)	−0.530 (< .001)	−0.511 (< .001)	0.634 (< .001)	0.705 (< .001)

Table 18: Importance of components, Study 2

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	2.040	0.875	0.634	0.554	0.465	0.384
Eigenvalues	4.159	0.765	0.405	0.307	0.217	0.147
% of VAR explained	0.693	0.128	0.067	0.051	0.036	0.025
Cumulative % VAR	0.693	0.821	0.888	0.939	0.975	1.000

Table 19: Component loadings, Study 2

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Upset	−0.424	−0.346	−0.208	−0.172	−0.788	
Frustrated	−0.403	−0.388		0.800	0.179	
Disturbed	−0.398	−0.484	0.114	−0.571	0.514	
Calm	0.410	−0.434	0.491			−0.627
Relaxed	0.432	−0.379	0.233		−0.201	0.758
Happy	0.382	−0.404	−0.799		0.182	−0.124

VII.1.4 Vignettes used in Study 3

MAYOR scenario, Study 3

[Vignette, both conditions] While mowing the lawn on a Saturday afternoon, you see Jill, your neighbor, removing banners from her front yard. The banners are in support of the current mayor, who is running for re-election. When you confront Jill and ask why she has decided to remove the banners, she tells you that she has read an article in a local newspaper about the mayor's alleged corruption, and that now **she is convinced that the mayor is deeply corrupt**. However, you know from a very reliable source that the information in the article is fabricated, and is just a blatant attempt to discredit an extremely popular, and to the best of your knowledge honest, mayor. **You are virtually certain that the mayor is innocent**.

[INCORRECT BELIEFS CONDITION]
So, you know that Jill has **incorrect beliefs** about the mayor.

[DIFFERENT BELIEFS CONDITION]
So, you know that you and Jill have **different beliefs** about the mayor.

[Attention check, both conditions] Jill believes that the mayor is innocent. [*False*]

PROMOTION scenario, Study 3

[Vignette, both conditions] Your colleague at a startup company, John, who you are really close with, tells you over lunch that he is looking forward to the next performance evaluation. John tells you that his boss seems to be very satisfied with his work, and that **he is convinced that he is getting promoted soon**. However, you know from a reliable source that John's boss is rather disappointed with his work, and **you are virtually certain that John is not getting promoted anytime soon**.

[INCORRECT BELIEFS CONDITION]
So, you know that John has **incorrect beliefs** about his promotion.

[DIFFERENT BELIEFS CONDITION]
So, you know that you and John have **different beliefs** about his promotion.

[Attention check, both conditions] John thinks that he is getting promoted soon. [*True*]

SAVINGS scenario, Study 3

[Vignette, both conditions] You meet up with your best friend, Jackie, in a café. After chatting about a TV show that you both binge-watched over last weekend, she tells you that she is very excited to share some news with you. Jackie tells you that she has decided to invest some of her savings in a biotech stock, BioZyme, that she had heard good things about. **She is convinced that BioZyme has excellent economic prospects.** However, you know someone who works at BioZyme, who told you that the company has no products in the pipeline, and **you are virtually certain that BioZyme is on the verge of bankruptcy.**

[INCORRECT BELIEFS CONDITION]
So, you know that Jackie has **incorrect beliefs** about BioZyme.

[DIFFERENT BELIEFS CONDITION]
So, you know that you and Jackie have **different beliefs** about BioZyme.

[Attention check, both conditions] Jackie has decided to invest some money in the biotech company BioZyme. [*True*]

WEIGHT LOSS scenario, Study 3

[Vignette, both conditions] Your brother, James, who is a software developer, just finished a very important project, for which he had to work overtime almost every day for three months. As a result, he has neglected his diet and put on some extra weight, about 15 lbs. James decided that he would adopt a healthier diet to regain his fitness. **He is convinced that a low-fat diet is ideal for people who want to lose weight**, so he completely stopped eating fatty foods such as nuts, cheese, fish, or olive oil. However, you have seen conclusive evidence on different diets in medical journals, and **you are virtually certain that a low-carb, and not a low-fat diet, is ideal for people who wish to lose weight**; so he should reduce eating rice, bread, and chips instead.

[INCORRECT BELIEFS CONDITION]
So, you know that James has **incorrect beliefs** about diets.

[DIFFERENT BELIEFS CONDITION]
So, you know that you and James have **different beliefs** about diets.

[Attention check, both conditions]: James wants to gain some weight. [*False*]

VII.1.5 OLS regression results, Study 3

Table 20: OLS regression results, Study 3

	Dependent variable:		
	Disturbed		
	(1)	(2)	(3)
Condition: Incorrect beliefs	10.492***	9.102***	9.267***
	(1.832)	(1.817)	(1.787)
Confidence: self correct (0–100)		0.086	0.067
		(0.056)	(0.055)
Confidence: other incorrect (0–100)		0.210***	0.226***
		(0.051)	(0.050)
Confidence: different beliefs (0–100)		−0.033	−0.048
		(0.052)	(0.052)
Sex (female = 1)			3.749*
			(1.783)
Age (years)			0.339***
			(0.071)
Scenario: Promotion	0.286	1.572	0.966
	(2.590)	(2.610)	(2.570)
Scenario: Savings	17.671***	18.582***	18.264***
	(2.593)	(2.577)	(2.537)
Scenario: Weight loss	−17.993***	−15.176***	−15.159***
	(2.593)	(2.604)	(2.563)
Constant (i.e., Mayor scenario)	49.963***	29.688***	16.681**
	(2.048)	(5.006)	(5.510)
Observations	829	829	829
R^2	0.212	0.247	0.273
Adjusted R^2	0.208	0.240	0.265
Note:			$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

VII.1.6 Vignettes used in Study 4

MAYOR scenario, Study 4

[CARES ABOUT CONDITION]
While mowing the lawn on a Saturday afternoon, you see Jill, your neighbor,

[DOES NOT CARE ABOUT CONDITION]
While visiting your friend who lives in a different state, you see Jill, your friend's neighbor,

[BOTH CONDITIONS]

removing banners from her front yard. The banners are in support of the current mayor, who is running for re-election. When you confront Jill and ask why she has decided to remove the banners, she tells you that she has read an article in a local newspaper about the mayor's alleged corruption, and that now **she is convinced that the mayor is deeply corrupt**. However, you know from a very reliable source that the information in the article is fabricated, and is just a blatant attempt to discredit an extremely popular, and to the best of your knowledge honest, mayor. **You are virtually certain that the mayor is innocent**.

[CARES ABOUT CONDITION]
You have lived in this town your entire life, and you care a lot about local politics.

[DOES NOT CARE ABOUT CONDITION]
You are driving home tomorrow, and you don't really care about local politics.

[BOTH CONDITIONS]
So, you know that Jill has **incorrect beliefs** about the mayor.

[Attention check, BOTH CONDITIONS]
Jill believes that the mayor is innocent. [*False*]

PROMOTION scenario, Study 4

[CARES ABOUT CONDITION]
Your colleague at a small startup company, John,

[DOES NOT CARE ABOUT CONDITION]
One of your colleagues at a large call center, John,

[BOTH CONDITIONS]

tells you over lunch that he is looking forward to the next performance evaluation. John tells you that his boss seems to be very satisfied with his work, and that **he is convinced that he is getting promoted soon**. However, you know from a reliable source that John's boss is rather disappointed with his work, and **you are virtually certain that John is not getting promoted anytime soon**.

[CARES ABOUT CONDITION]
You have known John for more than a decade, and you are really close with him.

[DOES NOT CARE ABOUT CONDITION]
John is one of the newest hires, and you barely know him.

[BOTH CONDITIONS]
So, you know that John has **incorrect beliefs** about his promotion.

[Attention check, BOTH CONDITIONS]
John thinks that he is getting promoted soon. [*True*]

SAVINGS scenario, Study 4

[CARES ABOUT CONDITION]
You meet up with your best friend, Jackie, in a café.

[DOES NOT CARE ABOUT CONDITION]
You meet up with one of your former classmates from high school, Jackie, in a café.

[BOTH CONDITIONS]

After chatting about a TV show that you both binge-watched over last weekend, she tells you that she is very excited to share some news with you. Jackie tells you that she has decided to invest some of her savings in a biotech stock, BioZyme, that she had heard good things about. **She is convinced that BioZyme has excellent economic prospects.** However, you know someone who works at BioZyme, who told you that the company has no products in the pipeline, and **you are virtually certain that BioZyme is on the verge of bankruptcy.**

[CARES ABOUT CONDITION]
You have known Jackie since childhood, and you are really close with her.

[DOES NOT CARE ABOUT CONDITION]
You haven't seen Jackie since high school, and you were never really close with her.

[BOTH CONDITIONS]
So, you know that Jackie has **incorrect beliefs** about BioZyme.

[Attention check, BOTH CONDITIONS]
Jackie has decided to invest some money in the biotech company BioZyme. [*True*]

WEIGHT LOSS scenario, Study 4

[CARES ABOUT CONDITION]
Your brother, James,

[DOES NOT CARE ABOUT CONDITION]
A passenger on a flight sitting next to you, James,

[BOTH CONDITIONS]

who is a software developer, just finished a very important project, for which he had to work overtime almost every day for three months. As a result, he has neglected his diet and put on some extra weight, about 15 lbs. James decided that he would adopt a healthier diet to regain his fitness. **He is convinced that a low-fat diet is ideal for people who want to lose weight**, so he completely stopped eating fatty foods such as nuts, cheese, fish, or olive oil. However, you have seen conclusive evidence on different diets in medical journals, and **you are virtually certain that a low-carb, and not a low-fat diet, is ideal for people who wish to lose weight**; so he should reduce eating rice, bread, and chips instead.

[CARES ABOUT CONDITION]
James is your only sibling, and you are really close with him.

[DOES NOT CARE ABOUT CONDITION]
You have not met James before, and he lives very far from you.

[BOTH CONDITIONS]
So, you know that James has **incorrect beliefs** about diets.

[Attention check, BOTH CONDITIONS]
James wants to gain some weight. [*False*]

VII.1.7 Mean disturbance ratings, Study 4

Table 21: Mean disturbance ratings and test statistics in Study 4, by scenario and condition

Scenario	Condition		Test statistics[1]
	Does not care M [95% CI]	Cares about M [95% CI]	
Mayor	59.15 [54.09, 64.21]	68.09 [63.34, 72.84]	$t(198) = 2.524$, $p = .012$ Cohen's $d = 0.356$
Promotion	40.54 [35.27, 45.82]	65.95 [61.39, 70.51]	$t(195) = 7.144$, $p < .001$ Cohen's $d = 1.007$
Savings	59.84 [54.58, 65.11]	76.51 [72.11, 80.92]	$t(197) = 4.761$, $p < .001$ Cohen's $d = 0.665$
Weight loss	29.68 [24.36, 35.00]	47.29 [41.79, 52.80]	$t(200) = 4.509$, $p < .001$ Cohen's $d = 0.430$
ALL (standardized)	-0.31 $[-0.41, -0.22]$	0.31 [0.22, 0.40]	$t(800) = 9.327$, $p < .001$ Cohen's $d = 0.656$

[1] *Pairwise comparisons between the means in the corresponding rows (two-tailed t-tests).*

VII.1.8 OLS regression results, Study 4

Table 22: OLS regression results, Study 4

	Dependent variable:		
	Standardized disturbance rating		
	(1)	(2)	(3)
Condition: Cares about	0.623^{***}	0.618^{***}	0.604^{***}
	(0.067)	(0.066)	(0.065)
Confidence: other incorrect		0.159^{***}	0.160^{***}
		(0.033)	(0.033)
Sex (female = 1)			0.217^{***}
			(0.065)
Age (years)			0.011^{***}
			(0.003)
Constant	-0.311^{***}	-0.309^{***}	-0.832^{***}
	(0.047)	(0.047)	(0.114)
Observations	808	808	808
R^2	0.097	0.123	0.156
Adjusted R^2	0.096	0.120	0.151
Note:			$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

VII.1.9 Dependent measures, Study 5

1. HOW LIKELY is it that YOU would AVOID WORKING with this person?
2. HOW LIKELY is it that YOU would AVOID TALKING to this person?
3. HOW LIKELY is it that YOU would AVOID HANGING OUT with this person?
4. HOW LIKELY is it that YOU would AVOID TRUSTING this person?
5. HOW LIKELY is it that YOU would BLOCK this person on Twitter?
6. Would YOU PREFER or NOT PREFER this person to be YOUR FAMILY MEMBER?
7. Would YOU PREFER or NOT PREFER this person to be YOUR NEIGHBOR?
8. Would YOU PREFER or NOT PREFER this person to be YOUR PARTNER / SPOUSE (romantic relationship)?
9. Would YOU PREFER or NOT PREFER this person to be YOUR COLLEAGUE / CLASSMATE (professional relationship)?

Participants provided their responses to Questions 1–5 on continuous slider scales from 0: *not at all* to 100: *very*.

Participants provided their responses to Questions 6–9 on continuous slider scales from −100: *prefer not* to 100: *prefer*.

VII.1.10 OLS regression results, Study 5

Table 23: OLS regression results, Study 5

	Dependent variable:	
	Disturbed	
	(1)	(2)
Confidence: different beliefs	0.028	0.006
	(0.065)	(0.065)
Confidence: other's beliefs is incorrect	0.456^{***}	0.471^{***}
	(0.058)	(0.059)
Confidence: own belief is correct	0.326^{***}	0.317^{***}
	(0.065)	(0.065)
Sex: female		6.806^{**}
		(2.427)
Age: years		0.086
		(0.095)
Ideology: conservative		−0.013
		(0.021)
Constant	-11.095^{*}	-16.562^{*}
	(5.438)	(6.507)
Observations	600	598
R^2	0.273	0.281
Adjusted R^2	0.269	0.274
Note:	$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$	

Additional OLS regression results, Study 5

Table 24: Regression results: avoidance behaviors, Study 5

	Dependent variable:									
	Avoid working		Avoid talking		Avoid hanging out		Avoid trusting		Block on Twitter	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Confidence: difference	−0.022	−0.010	0.010	0.031	0.048	0.062	0.024	0.044	−0.025	−0.018
	(0.070)	(0.069)	(0.072)	(0.071)	(0.070)	(0.069)	(0.066)	(0.066)	(0.074)	(0.074)
Confidence: other incorrect	0.536***	0.527***	0.551***	0.536***	0.621***	0.608***	0.597***	0.578***	0.511***	0.506***
	(0.063)	(0.062)	(0.064)	(0.064)	(0.063)	(0.062)	(0.060)	(0.059)	(0.066)	(0.067)
Confidence: self correct	0.066	0.063	0.042	0.041	−0.007	−0.009	0.064	0.064	−0.017	−0.019
	(0.070)	(0.069)	(0.072)	(0.070)	(0.070)	(0.068)	(0.066)	(0.066)	(0.074)	(0.074)
Sex: female		1.813		−0.313		1.355		−2.919		−1.342
		(2.575)		(2.634)		(2.559)		(2.457)		(2.770)
Age: years		−0.309**		−0.360***		−0.287**		−0.225*		−0.107
		(0.101)		(0.103)		(0.100)		(0.096)		(0.109)
Ideology: conservative		−0.069**		−0.073**		−0.086***		−0.060**		−0.004
		(0.022)		(0.022)		(0.022)		(0.021)		(0.023)
Constant	2.136	12.171′	3.271	15.883*	6.831	16.014*	7.974	16.930*	3.246	8.052
	(5.840)	(6.904)	(5.990)	(7.060)	(5.825)	(6.861)	(5.546)	(6.587)	(6.163)	(7.424)
Observations	600	598	600	598	600	598	600	598	600	598
R^2	0.199	0.227	0.204	0.236	0.256	0.287	0.269	0.287	0.152	0.151
Adjusted R^2	0.195	0.219	0.200	0.229	0.252	0.280	0.265	0.280	0.147	0.143

Note: ′$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$

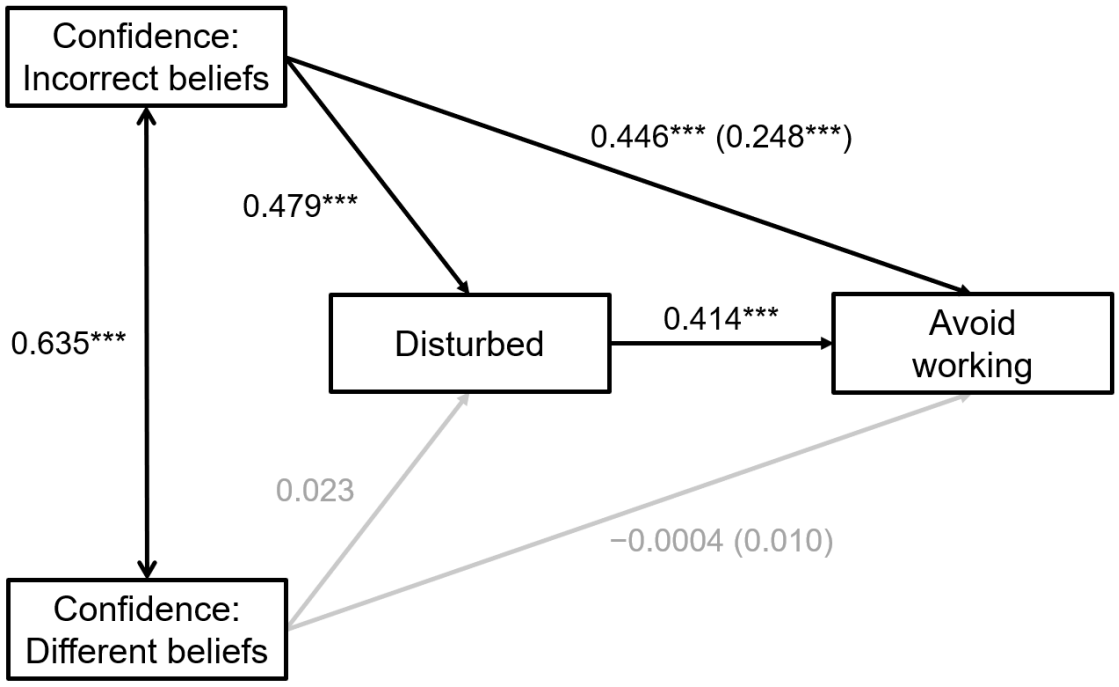
Table 25: Regression results: relationship preferences, Study 5

	Dependent variable:							
	Prefer: neighbor		Prefer: colleague		Prefer: family member		Prefer: romantic partner	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Confidence: difference	−0.144	−0.192′	−0.089	−0.125	−0.337**	−0.368***	−0.420***	−0.431***
	(0.105)	(0.102)	(0.105)	(0.102)	(0.106)	(0.104)	(0.104)	(0.102)
Confidence: other incorrect	−0.693***	−0.641***	−0.727***	−0.684***	−0.710***	−0.671***	−0.651***	−0.625***
	(0.094)	(0.092)	(0.094)	(0.091)	(0.095)	(0.094)	(0.093)	(0.092)
Confidence: self correct	0.191′	0.179′	0.137	0.129	0.160	0.152	0.105	0.101
	(0.105)	(0.101)	(0.105)	(0.101)	(0.106)	(0.104)	(0.104)	(0.102)
Sex: female		7.199′		2.914		3.225		−1.231
		(3.794)		(3.788)		(3.888)		(3.809)
Age: years		0.278′		0.244		0.163		−0.013
		(0.149)		(0.149)		(0.153)		(0.149)
Ideology: conservative		0.197***		0.205***		0.181***		0.180***
		(0.032)		(0.032)		(0.033)		(0.032)
Constant	24.639**	14.722	26.122**	19.128′	25.740**	21.267*	15.234′	18.759′
	(8.767)	(10.169)	(8.753)	(10.155)	(8.884)	(10.422)	(8.686)	(10.210)
Observations	600	598	600	598	600	598	600	598
R^2	0.157	0.219	0.163	0.225	0.209	0.251	0.224	0.262
Adjusted R^2	0.153	0.211	0.159	0.217	0.205	0.243	0.220	0.254

Note: $'p < .10$; $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

VII.1.11 Detailed mediation analyses, Study 5

Avoiding working with the author of the Tweet



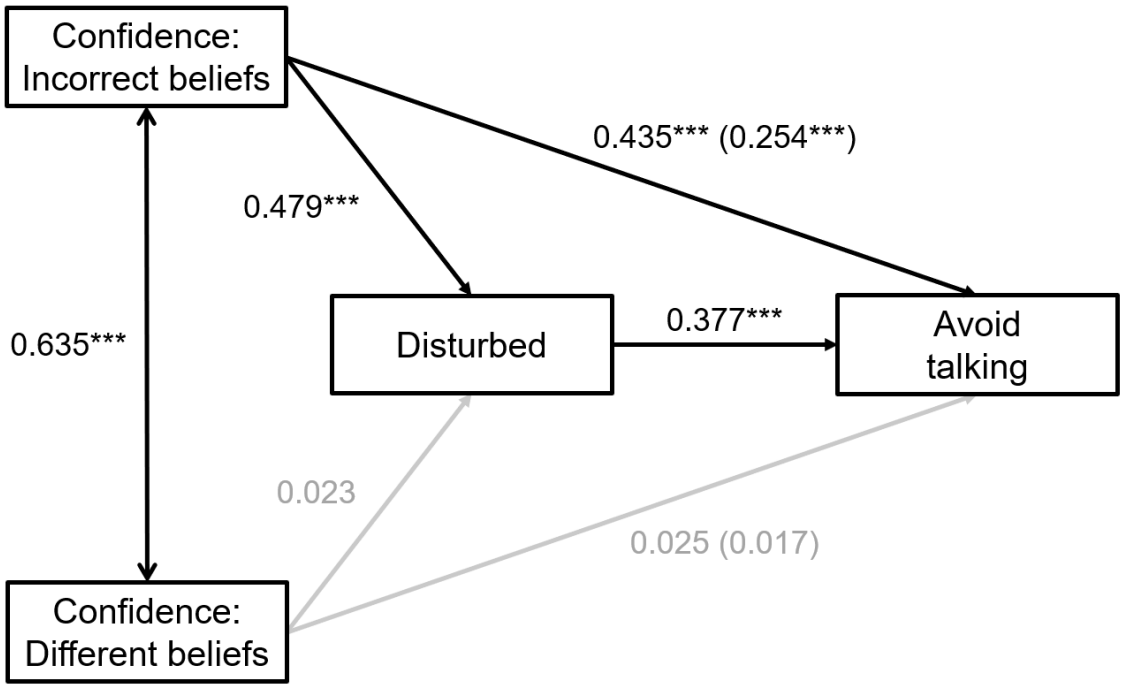
Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 22: Mediation: avoiding working with, Study 5. Coefficients: standardized Beta coefficients.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = 0.198$, 95% CI [0.148, 0.256], $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = 0.010$, 95% CI [−0.032, 0.054], $p > .05$.

Avoiding talking to the author of the Tweet



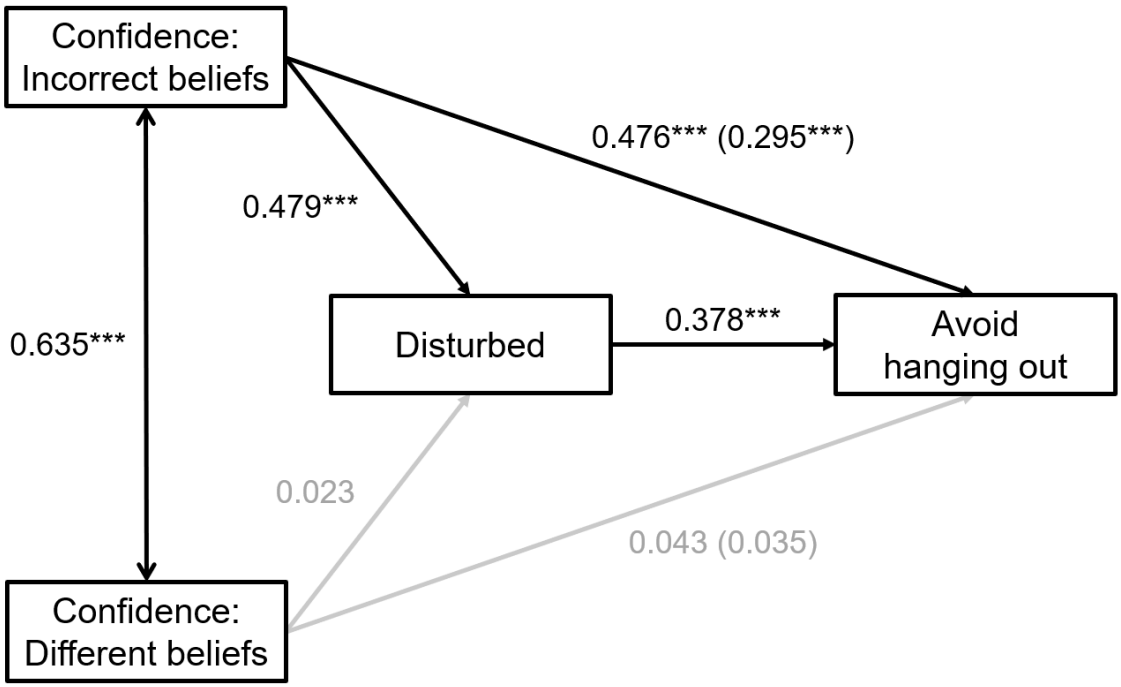
Note: $^{*}p < .05$*;* $^{**}p < .01$*;* $^{***}p < .001$

Figure 23: Mediation: avoiding talking to, Study 5. Coefficients: standardized Beta coefficients.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = 0.181$, 95% CI [0.131, 0.238], $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = 0.009$, 95% CI [−0.030, 0.049], $p > .05$.

Avoiding hanging out with the author of the Tweet



Note: $^{*}p < .05$*;* $^{**}p < .01$*;* $^{***}p < .001$

Figure 24: Mediation: avoiding hanging out, Study 5. Coefficients: standardized Beta coefficients.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = 0.181$, 95% CI [0.135, 0.238], $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = 0.009$, 95% CI [−0.030, 0.049], $p > .05$.

Avoiding trusting the author of the Tweet

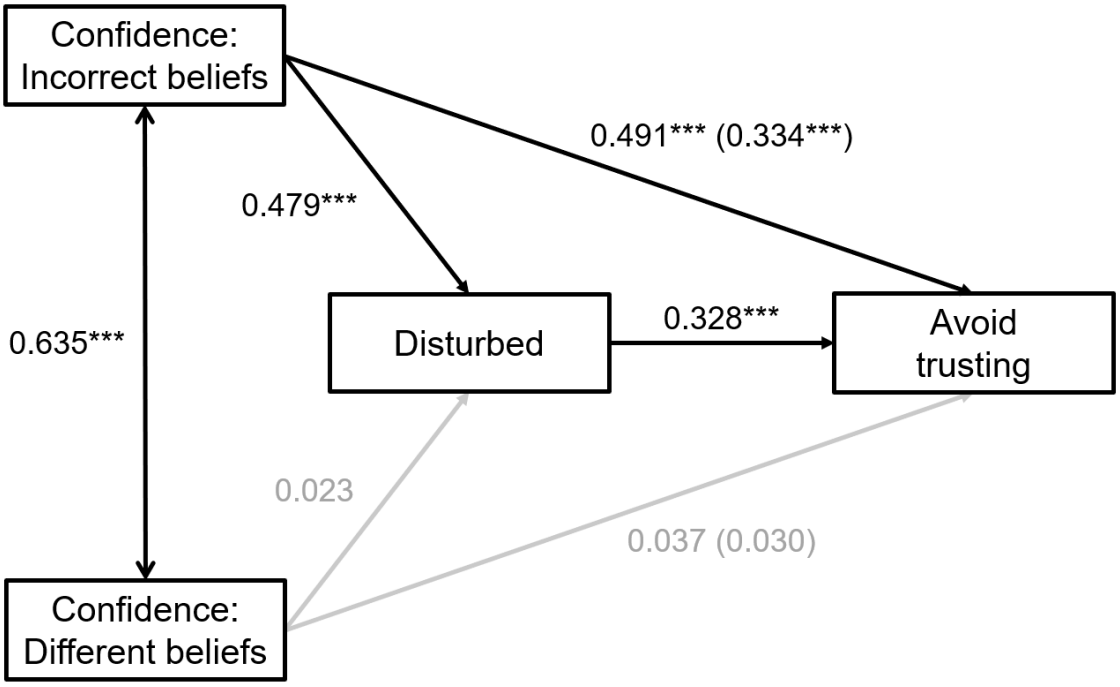
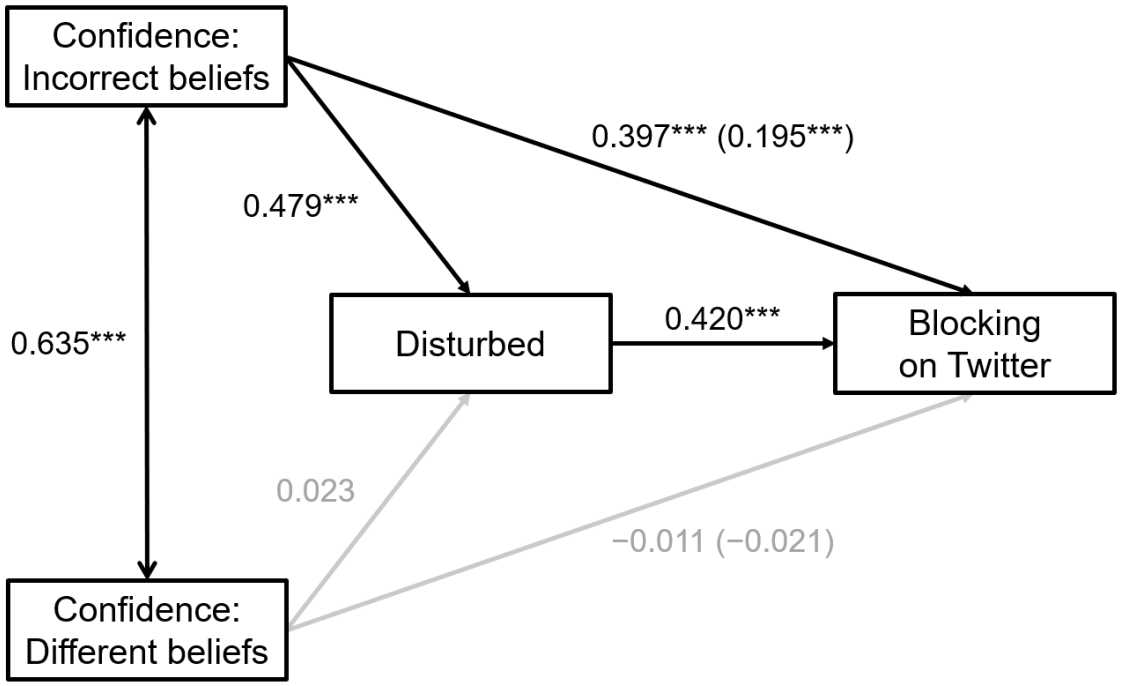


Figure 25: Mediation: avoiding trusting, Study 5. Coefficients: standardized Beta coefficients.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = 0.157$, 95% CI [0.112, 0.210], $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = 0.008$, 95% CI [−0.025, 0.044], $p > .05$.

Blocking the author of the Tweet



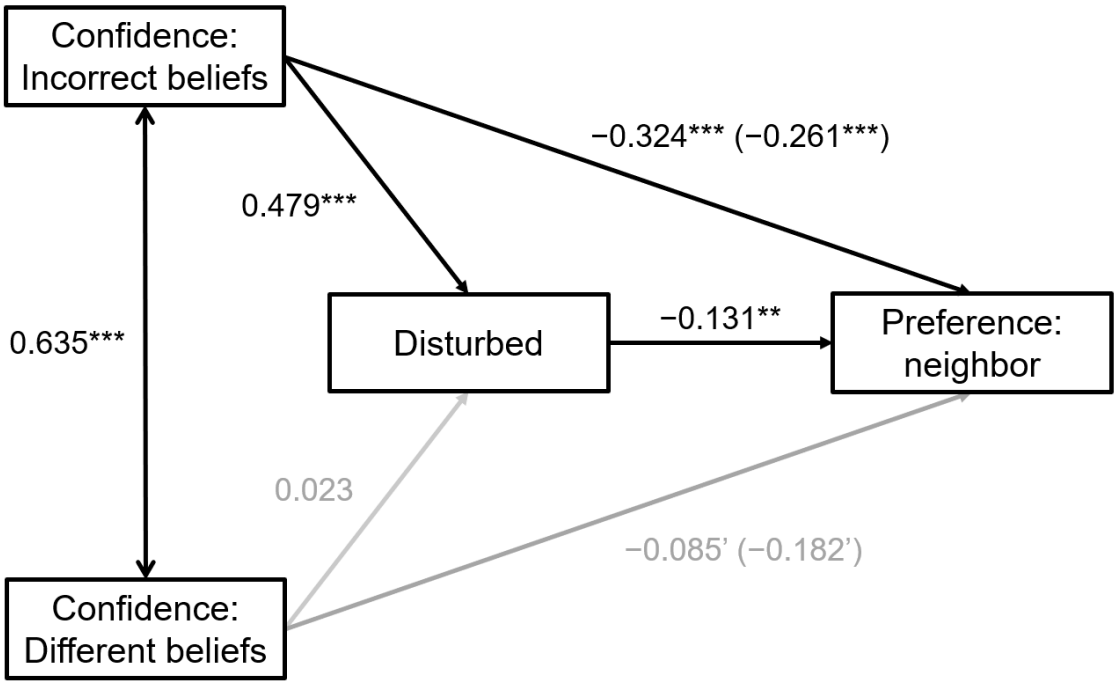
Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 26: Mediation: blocking on Twitter, Study 5. Coefficients: standardized Beta coefficients.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = 0.201$, 95% CI [0.149, 0.257], $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = 0.010$, 95% CI [−0.032, 0.054], $p > .05$.

Preference for having the author of the Tweet as a neighbor



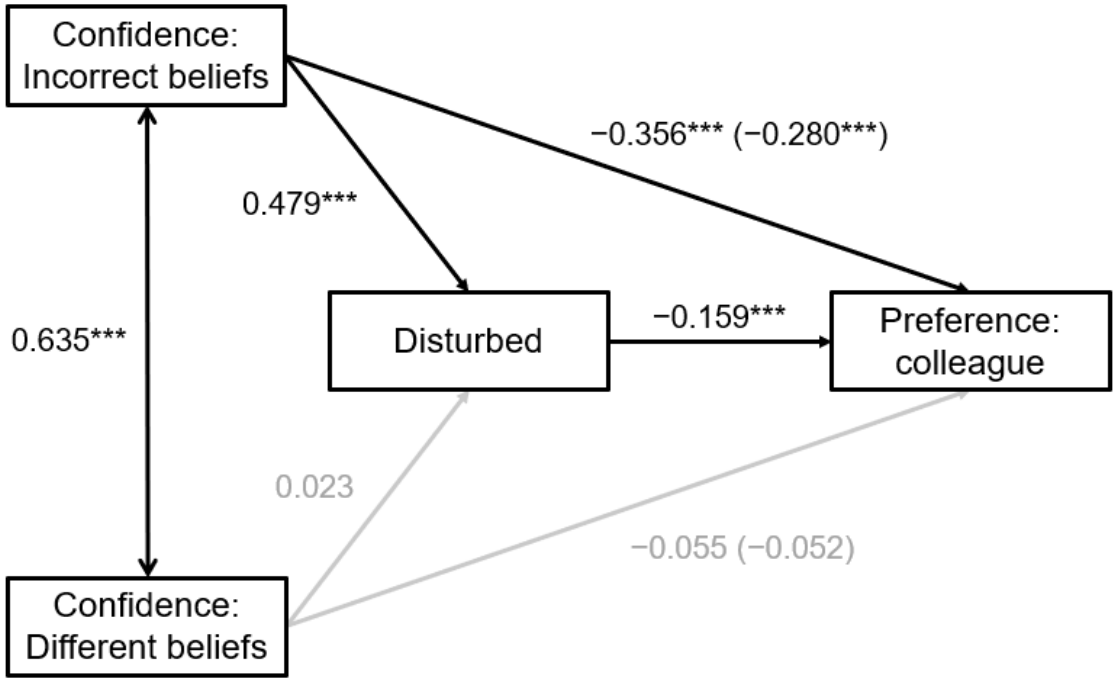
Note: $'p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 27: Mediation: preference as a neighbor, Study 5. Coefficients: standardized Beta.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = -0.063$, 95% CI $[-0.115, -0.018]$, $p = .005$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = -0.003$, 95% CI $[-0.020, 0.011]$, $p > .05$.

Preference for having the author of the Tweet as a colleague



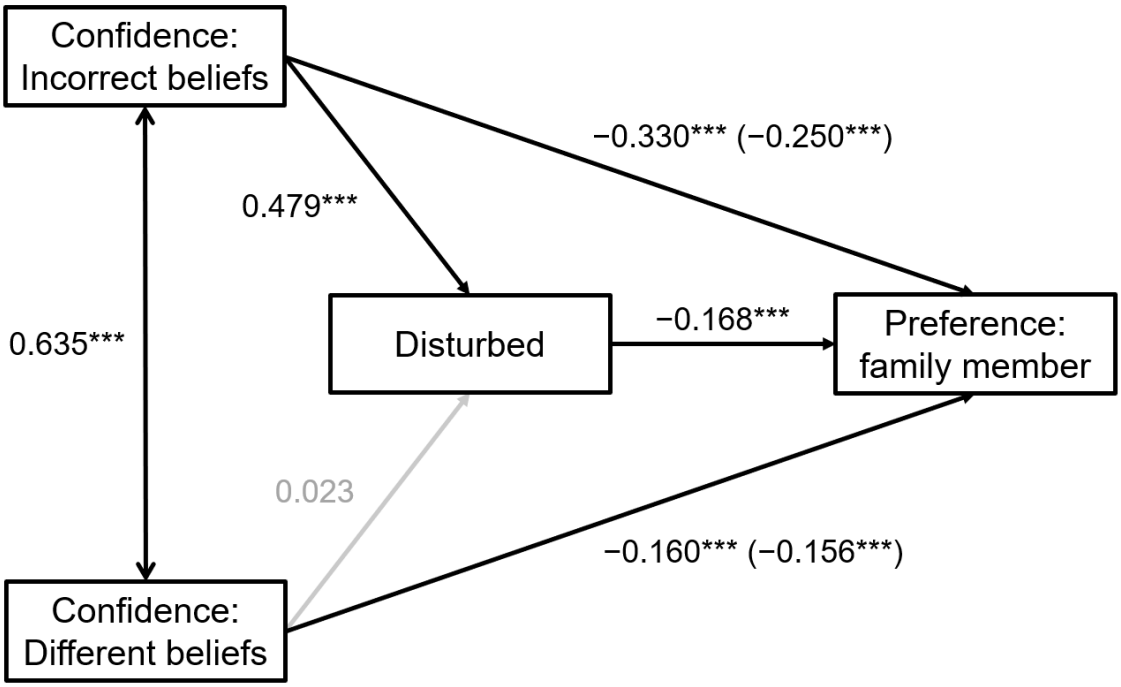
Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 28: Mediation: preference as a colleague, Study 5. Coefficients: standardized Beta.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = -0.076$, 95% CI $[-0.130, -0.033]$, $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = -0.004$, 95% CI $[-0.023, 0.014]$, $p > .05$.

Preference for having the author of the Tweet as a family member



Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 29: Mediation: preference as a family member, Study 5. Coefficients: standardized Beta.

Reported disturbance significantly mediates the effect of confidence that the other person has incorrect beliefs, $\beta = -0.080$, 95% CI $[-0.135, -0.034]$, $p < .001$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = -0.004$, 95% CI $[-0.023, 0.015]$, $p > .05$.

Preference for having the author of the Tweet as a romantic partner

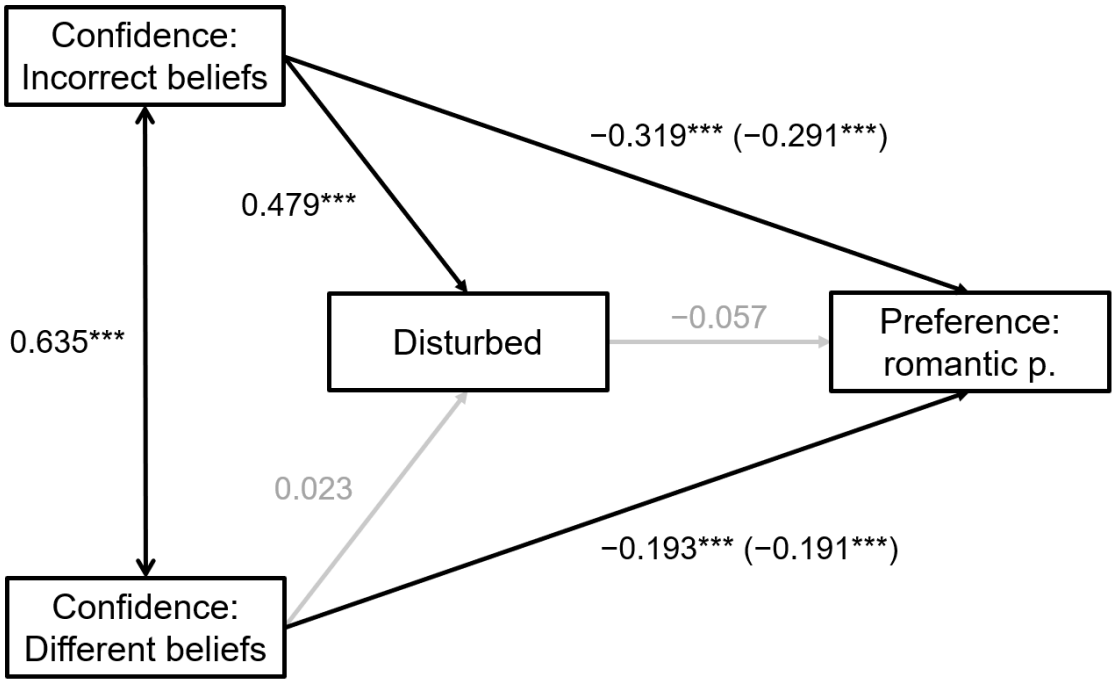


Figure 30: Mediation: preference as a romantic partner, Study 5. Coefficients: standardized Beta.

Reported disturbance does not mediate the effect of confidence that the other person has incorrect beliefs, $\beta = -0.027$, 95% CI [−0.0775, 0.018], $p > .05$.

Reported disturbance does not mediate the effect of confidence that the other person has different beliefs, $\beta = -0.001$, 95% CI [−0.011, 0.007], $p > .05$.

VII.2 Appendix for Chapter III

VII.2.1 Belief elicitation methods in Study 6

Recipients were told that they had to make a guess about both parties' alternative payoff (i.e. their own and their partner's): "How much would Option B have given to your partner / to you?" They were also told that they had a chance to win a bonus of \$0.25 per guess (\$0.50 in total) if they guessed the unchosen alternative payoffs correctly. The probability of getting these bonuses increased with accuracy:

Within an error of ±\$0.05, they received the bonus with 100% probability (±\$0.10: 75%, ±\$0.20: 50%, ±\$0.30: 25%).

If the difference between the guess and the actual value was higher than \$0.30, participants did not have a chance to win the bonus. After participants made their guesses, we calculated the probability of winning the bonus based on their accuracy, and then a random number generator algorithm determined whether the recipient has won \$0.50 (for correctly guessing both payoffs) or \$0.25 (for correctly guessing one payoff) or \$0. Recipients were not informed directly about their accuracy, or probability to win, but they were told whether they received any bonuses at the end of the experiment. This belief elicitation technique may seem overly complicated, but it was necessary in our design. We had to make sure that recipients could not determine the unchosen alternative option based on the feedback from the guessing stage. Disclosing their accuracy or probability to win the prize would have indirectly revealed the alternative option, even if the allocator decided to keep it hidden.

VII.2.2 Sample decision screen, Study 6

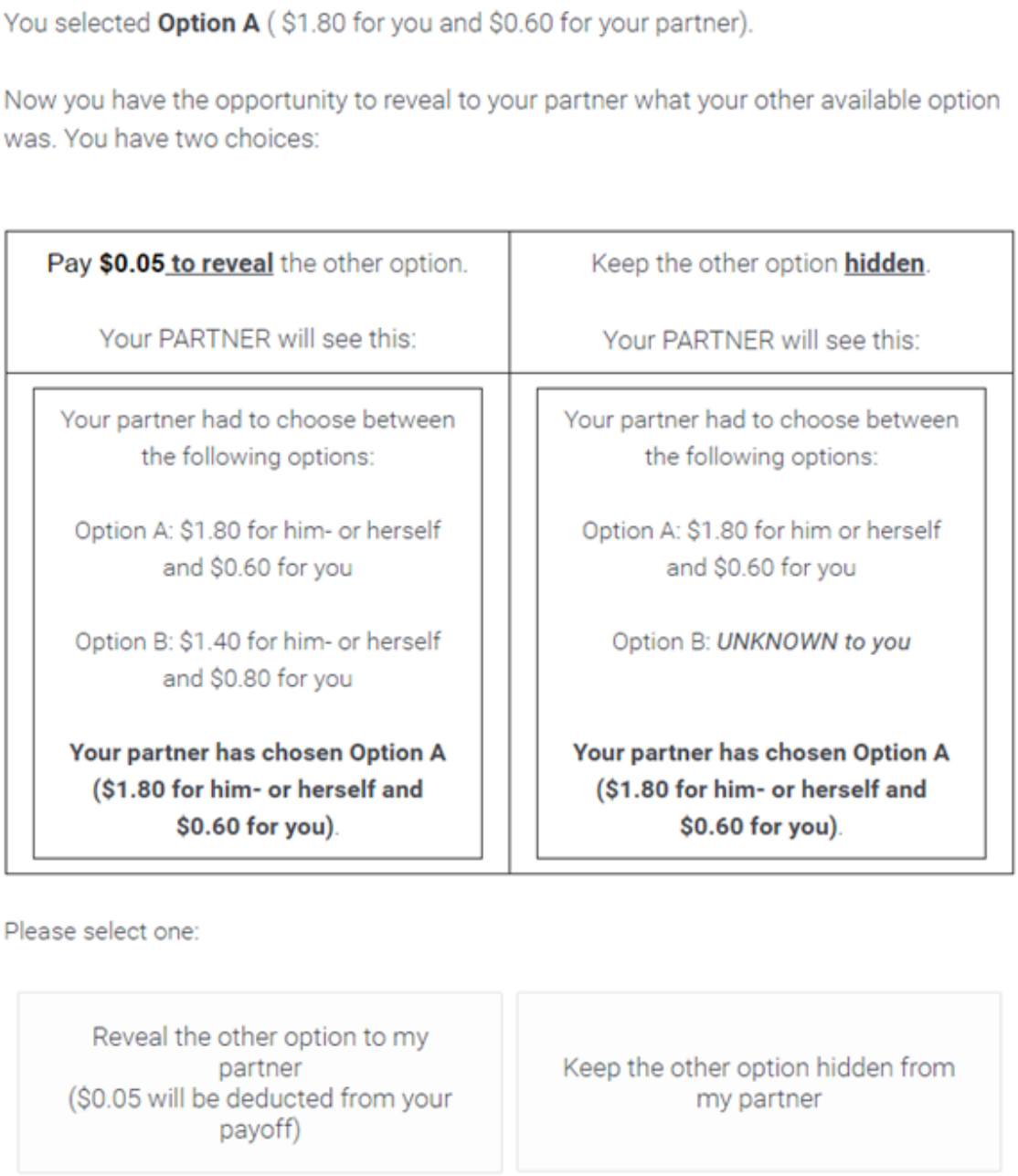


Figure 31: Sample screenshot of the allocator's decision screen, Study 6.

VII.2.3 Recipients' guesses and ratings of the transfer, Study 6

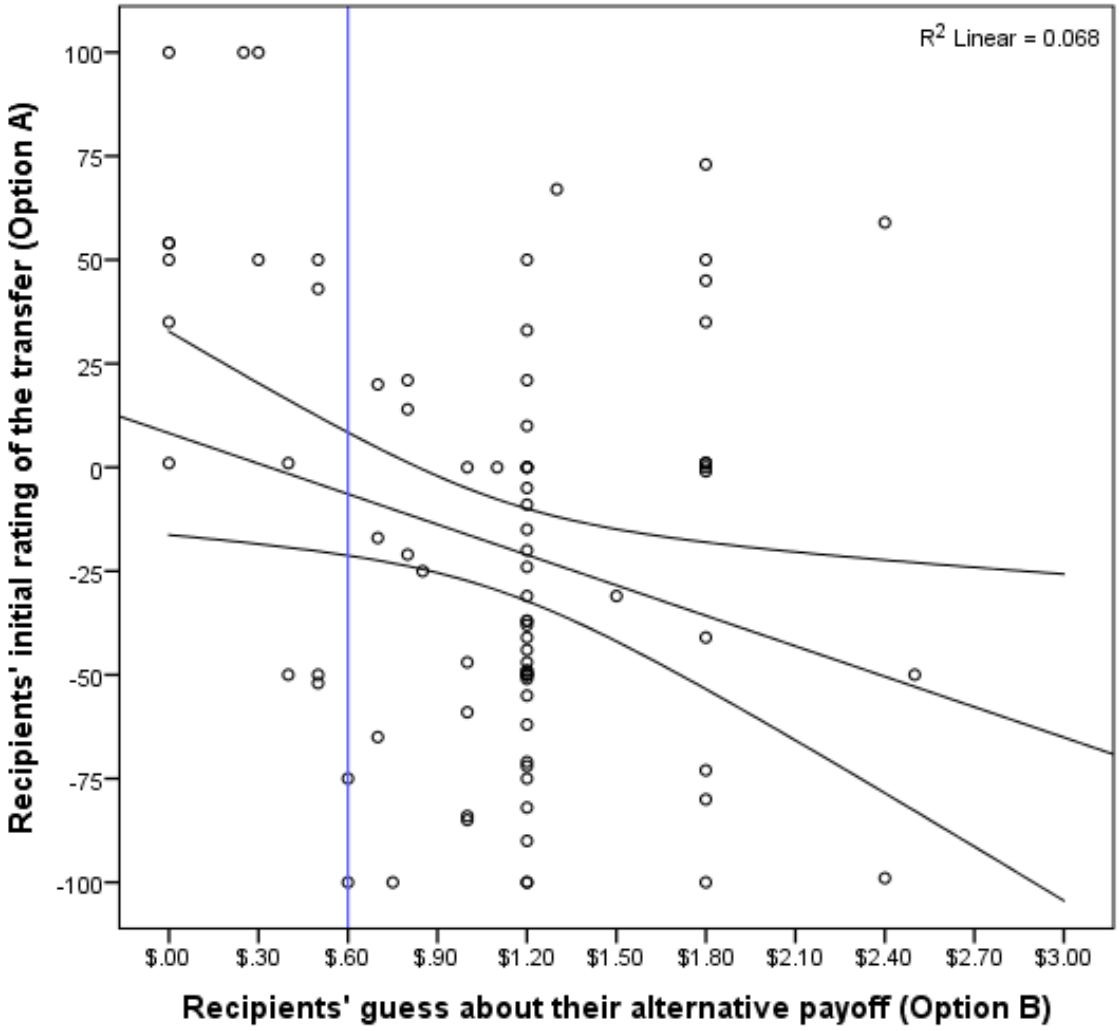


Figure 32: Recipients' initial ratings of their partner's transfer in function of the expected amount of the alternative option. The vertical blue indicates the amount of money recipients actually received (Option A = \$0.60). The negative slope indicates that recipients evaluated their partner's transfer more negatively when their expectations about the alternative option were higher. The linear regression is significant at $p = .017$.

Table 26: Recipients' ratings of the transfer before and after the allocator revealed (or did not reveal) the alternative option, Study 6

				Rating of transfer (before)				Rating of transfer (after)		Change of rating	
Condition	*n*	Received option	*n*	*M*	95% CI	Allocator revealed	*n*	*M*	95% CI	*M*	sig.[2]
Generous	40	A	23	-27.26	[-43.47, -11.06]	no	11	-17.82	[-42.72, 7.08]	5.64	.586
						yes	12	65.00	[38.73, 91.27]	95.75	< .001
		B	17	-82.76	[-95.5, -70.03]	no	16	-75.88	[-92.52, -59.23]	7.06	.146
						yes	1		[N/A][1]		
Reasonable	40	A	39	-23.92	[-39.34, -8.5]	no	26	-28.38	[-46.97, -9.78]	-2.13	.677
						yes	13	45.69	[7.2, 84.18]	64.23	< .001
		B	1		[N/A][1]	no	0		[N/A][1]		
						yes	1		[N/A][1]		
Selfish	40	A	22	-0.45	[-27.45, 26.54]	no	21	-14.57	[-38.8, 9.65]	-9.33	.430
						yes	1		[N/A][1]		
		B	18	-6.39	[-32.33, 19.55]	no	12	7.42	[-20.71, 35.55]	1.67	.706
						yes	6	90.00	[75.46, 104.54]	120.67	.010

[1] Sample size was too small to calculate CI, [2] Significance levels were obtained by performing paired sample *t*-tests on ratings of transfers.

VII.2.4 Additional dependent measures and PCA, Study 7

The following were asked of allocators after the choice to reveal or not had already been made.

1. "In your option, to what extent does your partner want to know Option [unselected]? [slider: 0 does not want to know at all ... 100 strongly wants to know]

"Indicate response for the questions below [slider: 0 not important at all ... 100 very important]:

2. "How important is it for you that people who are close to you (family, friends, colleagues) perceive you as fair?" [fairness / close]
3. "How important is it for you that people who are close to you (family, friends, colleagues) perceive you as honest?" [honesty / close]
4. "How important is it for you to make sure that people who are close to you (family, friends, colleagues) don't feel bad?" [empathy / close]
5. "How important is it for you that people who you do not know (strangers) perceive you as fair?" [fairness / stranger]
6. "How important is it for you that people who you do not know (strangers) perceive you as honest?" [honesty / stranger]
7. "How important is it for you that people who you do not know (strangers) don't feel bad?" [empathy / stranger]"

We observed significant positive correlations between all of the six measures that captured allocator's social concerns towards strangers and close others, all $r > .279$, all $p < .001$, suggesting reasonable factorability. Secondly, the Kaiser-Meyer-Olkin measure of sampling adequacy was .683, above the commonly recommended value of .6, and Bartlett's test of sphericity was significant, $\chi^2(15) = 243.32$, $p < .001$. Therefore, we conducted a principal components analysis to identify and compute composite scores for the factors underlying the six measures. Initial eigen values indicated that the first two factors explained 58%, and 20% of the variance, respectively. The third, fourth, fifth, and sixth factors had eigen values below one, and explained 13%, 5%, 2%, and 2% of the variance, respectively (Table 27).

The component matrix (Table 28) indicated that the first two components correspond to the three close others-related, and the three stranger-related questions, respectively, therefore we created two new variables ("attitudes towards close others" and "attitudes towards

strangers") for the two components by taking the arithmetic mean of the corresponding three measurements for each. These compound measurements indicated to what extent allocators cared about the beliefs and feelings of close others or strangers, respectively.

Table 27: Principal Component Analysis, Study 7

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.583	59.715	59.715	3.583	59.715	59.715	2.563	42.718	42.718
2	1.121	18.676	78.391	1.121	18.676	78.391	2.14	35.673	78.391
3	0.707	11.783	90.174						
4	0.26	4.341	94.515						
5	0.24	4.004	98.519						
6	0.089	1.481	100						

Extraction Method: Principal Component Analysis.

Table 28: Component matrix, Study 7

Component Matrix[a]

	Component	
	1	2
Fair to close others	0.756	0.542
Honest to close others	0.702	0.593
Feel good: close others	0.695	0.173
Fair to strangers	0.865	-0.379
Honest to strangers	0.829	-0.351
Feel good: strangers	0.775	-0.423

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

VII.3 Appendix for Chapter IV

VII.3.1 Detailed results of hierarchical linear regression analyses, Study 8.

Table 29: Results of the hierarchical linear regression analysis: Number of views (log)

	Dependent variable:						
	Number of Views (log-transformed)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Transgressor suffers	0.249			0.137	0.223		0.102
	(0.186)			(0.209)	(0.209)		(0.210)
Transgressor learns reason		0.274*		0.256*		0.394**	0.375**
		(0.109)		(0.114)		(0.135)	(0.141)
Transgressor learns identity			0.013		−0.0002	−0.206	−0.203
			(0.104)		(0.106)	(0.129)	(0.131)
Story age (days since post)	−0.00004	−0.0001	−0.0001	−0.0001	−0.0001	−0.0001	−0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	4.324***	4.379***	4.550***	4.261***	4.349***	4.435***	4.351***
	(0.194)	(0.113)	(0.110)	(0.222)	(0.232)	(0.119)	(0.229)
Observations	85	81	82	80	81	79	78
R^2	0.022	0.076	0.002	0.079	0.017	0.103	0.104
Adjusted R^2	−0.002	0.053	−0.024	0.042	−0.022	0.067	0.055
Note:							$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 30: Results of the hierarchical linear regression analysis: Number of upvotes (log)

	Dependent variable:						
	Number of upvotes (log-transformed)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Transgressor suffers	0.318			0.056	0.200		0.042
	(0.258)			(0.282)	(0.286)		(0.288)
Transgressor learns reason		0.469**		0.456**		0.501**	0.489*
		(0.147)		(0.154)		(0.185)	(0.193)
Transgressor learns identity			0.209		0.193	−0.060	−0.061
			(0.142)		(0.145)	(0.176)	(0.180)
Story age (days since post)	−0.001*	−0.001*	−0.001*	−0.001*	−0.001*	−0.001*	−0.001*
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Constant	2.997***	2.998***	3.166***	2.958***	2.992***	3.018***	2.992***
	(0.269)	(0.153)	(0.150)	(0.300)	(0.318)	(0.163)	(0.314)
Observations	85	81	82	80	81	79	78
R^2	0.069	0.165	0.079	0.164	0.086	0.164	0.163
Adjusted R^2	0.047	0.144	0.055	0.132	0.050	0.130	0.117
Note:							$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 31: Results of hierarchical linear regression analysis: Upvotes per View (log)

	Dependent variable:						
	Number of upvotes per View (log)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Transgressor suffers	0.036			−0.006	0.014		−0.003
	(0.037)			(0.040)	(0.040)		(0.040)
Transgressor learns reason		0.065**		0.065**		0.053*	0.053†
		(0.021)		(0.022)		(0.026)	(0.027)
Transgressor learns identity			0.047*		0.045*	0.019	0.019
			(0.020)		(0.020)	(0.025)	(0.025)
Story age (days since post)	−0.0001**	−0.0001***	−0.0001**	−0.0001***	−0.0001**	−0.0001**	−0.0001**
	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)
Constant	0.685***	0.678***	0.690***	0.684***	0.678***	0.673***	0.677***
	(0.038)	(0.021)	(0.021)	(0.042)	(0.044)	(0.023)	(0.044)
Observations	85	81	82	80	81	79	78
R^2	0.116	0.212	0.164	0.213	0.166	0.217	0.217
Adjusted R^2	0.095	0.192	0.143	0.181	0.134	0.186	0.174

Note: $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

VII.3.2 Robustness check, Study 9

Number (and proportion) of participants choosing each message, MTurk only

Message	Participants							
	Everyone (any X)		Did not punish (X = $1)		Punished (X < $1)		Fully punished (X = $0)	
Ignorance (no message)	19	20%	2	22%	17	20%	8	20%
Bonus only ($X)	12	13%	3	33%	9	10%	1	2%
Suffering ($X/$1)	8	8%	2	22%	6	7%	2	5%
Justice ($X/$1 + unfair)	15	16%	1	11%	14	16%	2	5%
Revenge ($X/$1 + unfair + partner responsible)	42	44%	1	11%	41	47%	28	68%
Total	96		9		87		41	

Figure 33: Message choices in Study 9, MTurk sample only.

Number (and proportion) of participants choosing each message, Prolific only

Message	Participants							
	Everyone (any X)		Did not punish (X = $1)		Punished (X < $1)		Fully punished (X = $0)	
Ignorance (no message)	18	18%	3	19%	15	18%	10	28%
Bonus only ($X)	18	18%	9	56%	9	11%	1	3%
Suffering ($X/$1)	10	10%	3	19%	7	9%	0	0%
Justice ($X/$1 + unfair)	13	13%	0	0%	13	16%	7	19%
Revenge ($X/$1 + unfair + partner responsible)	39	40%	1	6%	38	46%	18	50%
Total	98		16		82		36	

Figure 34: Message choices in Study 9, Prolific sample only.

VII.3.3 Detailed exclusion report, Study 10

Table 32: Exclusion report, Study 10.

Role	Condition	Matched	Excluded	Included	Excluded %	χ^2 test
Allocator	ignorance	219	0	219	0.0%	
	suffering	229	0	229	0.0%	NA
	justice	228	2	226	0.9%	(expected $n = 0$)
	revenge	227	1	226	0.4%	
Recipient	ignorance	219	20	199	9.1%	
	suffering	229	28	201	12.2%	$\chi^2(3, N = 903) =$
	justice	228	25	203	11.0%	1.126, $p = .771$
	revenge	227	25	202	11.0%	

Note: Allocators had to pass only three comprehension checks, whereas Recipients had to pass six comprehension checks and an attention check, leading to higher exclusion rates for Recipients.

VII.3.4 Punishment decision screens, Study 10

Ignorance condition

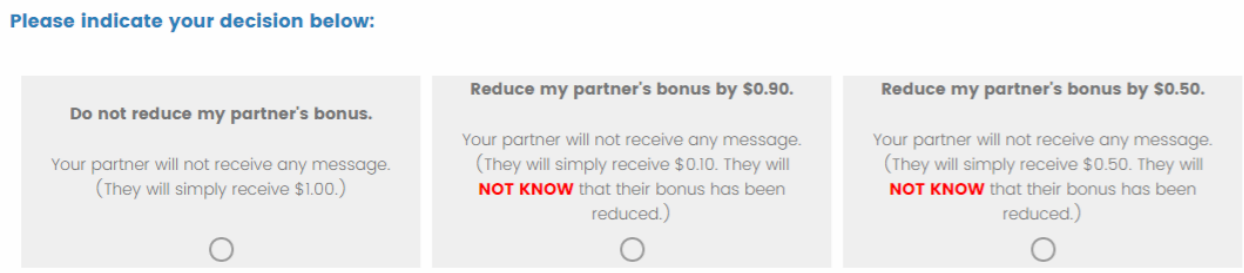


Figure 35: Screenshot of the punishment decision in the ignorance condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Suffering condition

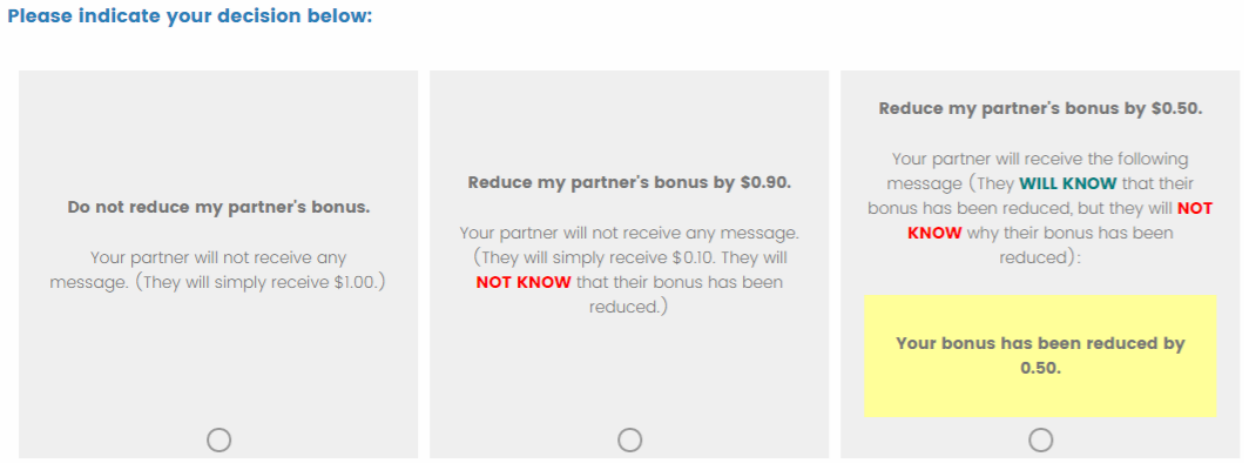


Figure 36: Screenshot of the punishment decision in the suffering condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Justice condition

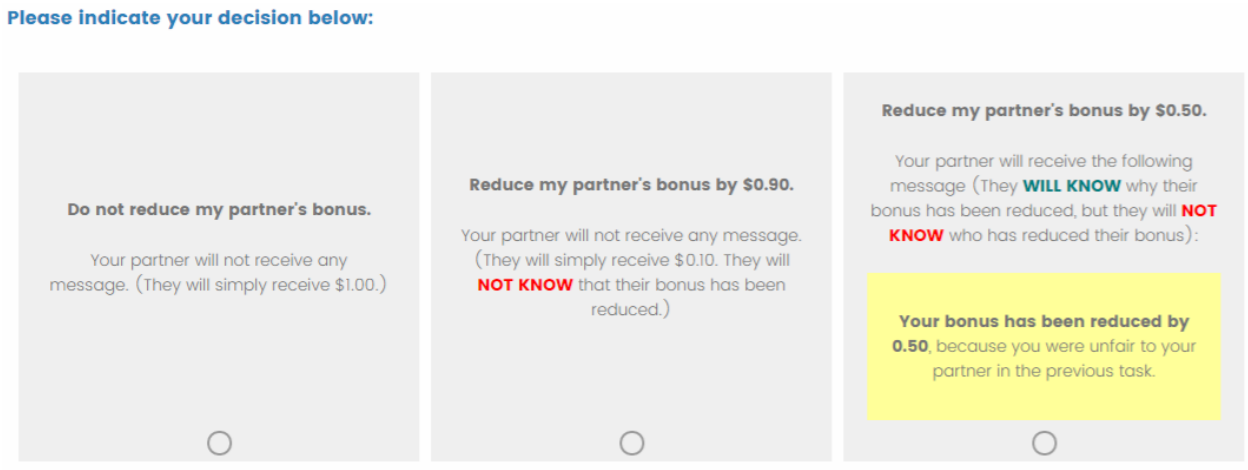


Figure 37: Screenshot of the punishment decision in the justice condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

Revenge condition

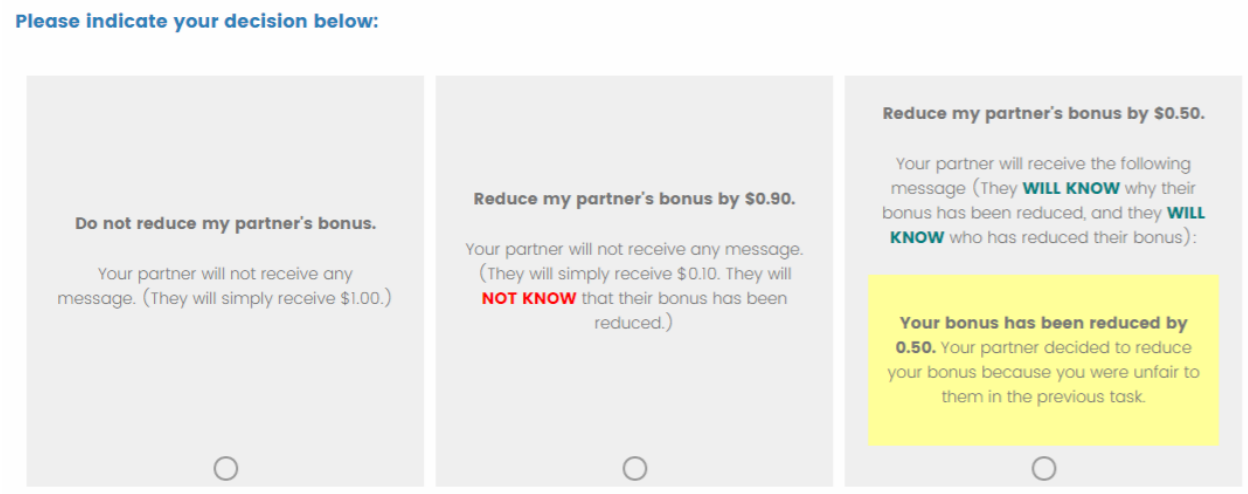


Figure 38: Screenshot of the punishment decision in the revenge condition. The no punishment (left), moderate punishment (right), and severe punishment (middle) options were presented in a random order.

VII.3.5 Robustness Checks, Study 10

Robustness check: Anger

In this robustness check we looked at the role of anger. It can be argued that when people are in a "hot" state, they have stronger retributive motives, and think less about the consequences of their actions, thus might have weaker belief-based preferences, and would prefer to inflict greater pain to the transgressor. If this is the case, we should expect a shift towards the moderate punishment option when we introduce the suffering component (ignorance → suffering condition), but not when we introduce the explanation component (suffering → justice condition). To test this,we analyzed the choices of Recipients who reported that they were angry when they saw the Allocator's decision, i.e., who selected *angry* from the list of 12 feelings ($n = 242$). While these participants were more likely to choose the severe punishment overall, we still observe a significant shift towards the moderate punishment option (crowding out), once this option allowed for sending an explanation (see Figure 39).

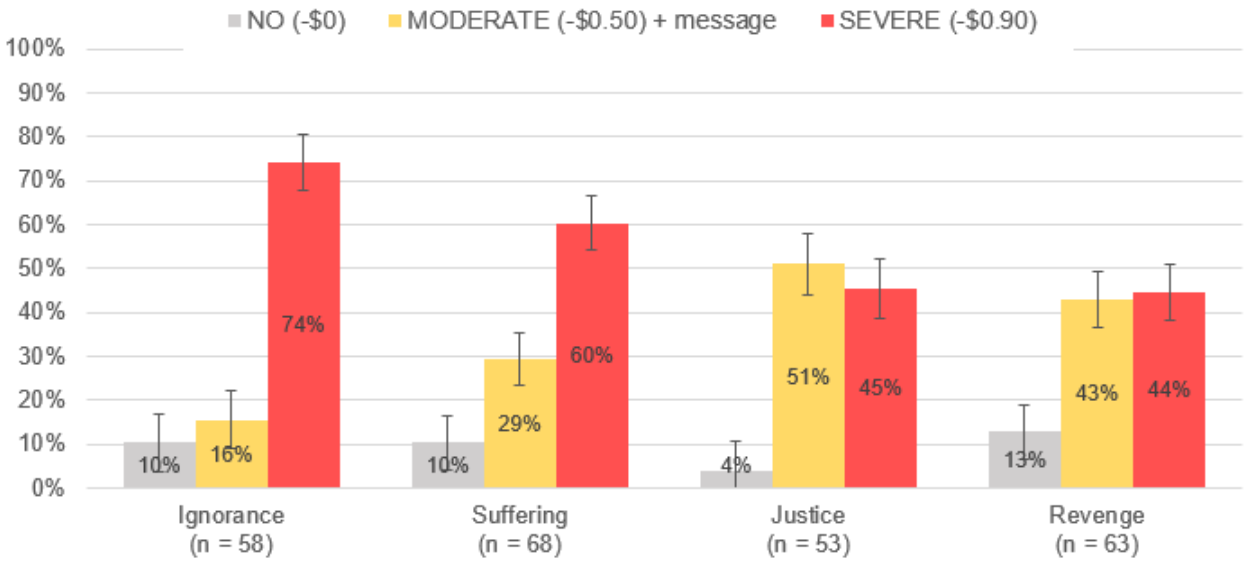


Figure 39: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions, among those Recipients who were angry. Error bars: $\pm 1SE$.

If anything, the main effect between conditions is even more pronounced for angry Recipients than for the full sample. All of the main results are robust: A significantly higher proportion of Recipients chose the moderate punishment option in the justice condition (51%) and in the revenge condition (43%) than in the ignorance condition (16%), $\chi^2(1, N = 111) = 14.285$, $p < .001$, and $\chi^2(1, N = 121) = 9.532$, $p = .002$, respectively. The proportion of moderate-choosers was also significantly higher in the justice condition (51%) than in the suffering condition (29%), $\chi^2(1, N = 121) = 4.942$, $p = .026$.

There is no significant difference between the revenge and suffering, ignorance and suffering, and justice and revenge conditions, $\chi^2(1, N = 131) = 2.019$, $p = .155$, $\chi^2(1, N = 126) = 2.672$, $p = .102$, and $\chi^2(1, N = 116) = 0.466$, $p = .495$, respectively. The main results of the OLS regression hold as well: When limiting the analysis to angry Recipients only, the justice and revenge conditions are associated with a significantly higher likelihood of choosing the moderate option (Table 33). Interestingly, we also observe a marginally significant effect of the suffering component, which suggests that angry Recipients had at least some retributive motives, in addition to distributive and belief-based preferences.

Table 33: Regression results: Likelihood of choosing punishment options, among those Recipients who reported being angry ($n = 242$)

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE pun.		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer	−0.001	−0.009	0.139†	0.142†	−0.138	−0.133
	(0.053)	(0.052)	(0.082)	(0.083)	(0.087)	(0.087)
Explain	−0.065	−0.066	0.215*	0.207*	−0.150†	−0.141
	(0.054)	(0.054)	(0.084)	(0.085)	(0.089)	(0.089)
Identity	0.089	0.089	−0.081	−0.088	−0.008	−0.002
	(0.055)	(0.055)	(0.086)	(0.086)	(0.090)	(0.090)
Sex (Female = 1)		0.090*		0.043		−0.133*
		(0.038)		(0.060)		(0.062)
Age (years)		−0.001		−0.003		0.004
		(0.002)		(0.003)		(0.003)
Constant	0.103**	0.084	0.155*	0.234*	0.741***	0.682***
	(0.039)	(0.068)	(0.060)	(0.107)	(0.064)	(0.112)
Observations	242	242	242	242	242	242
R^2	0.012	0.035	0.076	0.082	0.058	0.081
Adjusted R^2	−0.001	0.015	0.064	0.063	0.047	0.061
Note:						$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Robustness check: Suspicion (believing the interaction was fake).
In this robustness check we excluded all participants who indicated a strong (but incorrect) belief that they were interacting with a robot, as opposed to interacting with another human. In the following analyses we excluded everyone who responded below -50 to the following question (i.e., indicated a serious doubt that their partner was real): "Using the slider below, please indicate the extent to which you believed you were interacting with a bot or human partner" (continuous scale from *-100: definitely a bot* to *+100: definitely a human*). By applying the above criteria, we excluded 251 Recipients (31%), while 554 (69%) were retained in the sample. Figure 40 summarizes the main results:

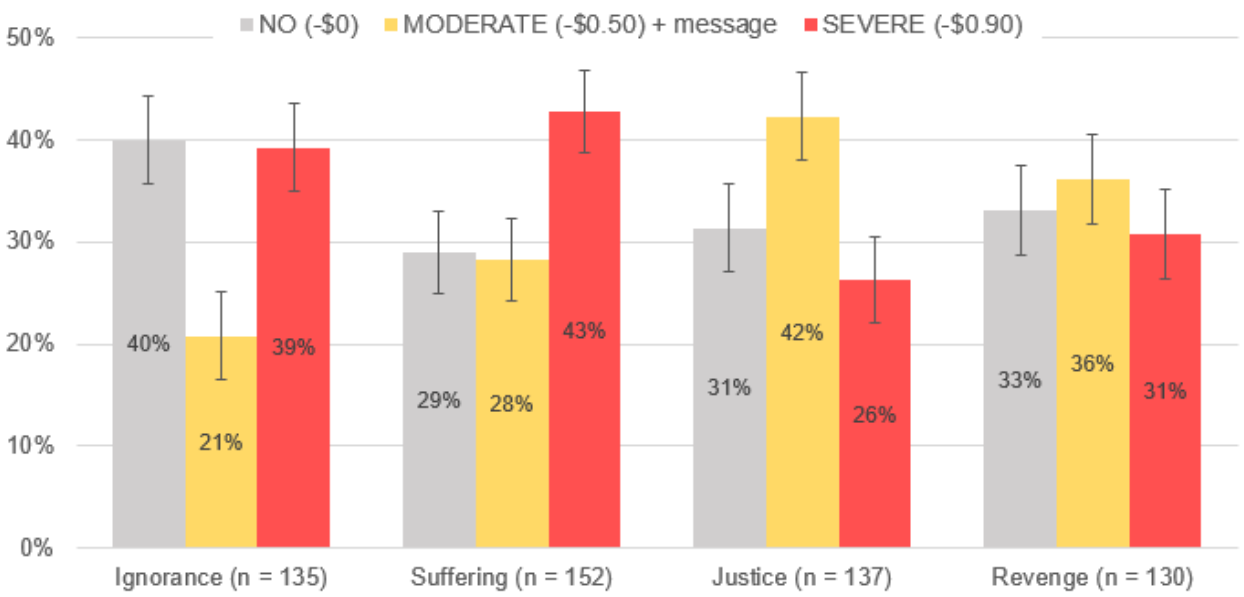


Figure 40: Proportion of Recipients choosing NO (grey), MODERATE (yellow), and SEVERE (red) punishment across conditions, among only those Recipients who had at least some reasonable confidence that they were interacting with another human. Error bars represent ±1 standard error.

All of the main results are robust to excluding suspicious Recipients: A significantly higher proportion of Recipients chose the moderate punishment option in the justice condition (42%) and in the revenge condition (36%) than in the ignorance condition (21%), $\chi^2(1, N = 272) = 13.684$, $p < .001$, and $\chi^2(1, N = 265) = 7.012$, $p = .008$, respectively.

The proportion of moderate-choosers was also significantly higher in the justice condition (42%) than in the suffering condition (28%), $\chi^2(1, N = 289) = 5.651$, $p = .002$. There is no significant difference between the revenge and suffering, ignorance and suffering, and justice and revenge conditions, $\chi^2(1, N = 282) = 1.649$, $p = .199$, $\chi^2(1, N = 287) = 1.802$, $p = .180$, and $\chi^2(1, N = 267) = 0.825$, $p = .364$, respectively. The main results of the OLS regression hold as well: Adding the explanatory component to the moderate option (i.e., justice and revenge conditions) is associated with a significantly higher likelihood of

choosing the moderate option and a significantly lower likelihood of choosing the severe option (Table 34). In addition, adding the suffering component to the moderate option (i.e., all but the ignorance condition) is associated with a significantly lower likelihood of choosing no punishment, but this effect holds only if we exclude suspicious participants (cf. Table 12).

Table 34: Regression results: Likelihood of choosing punishment options, excluding Recipients who believed that they were interacting with a robot.

	Dependent variable:					
	Likelihood of choosing NO punishment		Likelihood of choosing MODERATE pun.		Likelihood of choosing SEVERE punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Suffer	-0.111^{*}	$-0.108^{\dagger}$	0.075	0.071	0.035	0.037
	(0.056)	(0.056)	(0.054)	(0.054)	(0.056)	(0.056)
Explain	0.024	0.026	0.140^{**}	0.140^{*}	-0.165^{**}	-0.165^{**}
	(0.055)	(0.055)	(0.054)	(0.054)	(0.056)	(0.056)
Identity	0.017	0.018	-0.062	-0.062	0.045	0.044
	(0.058)	(0.057)	(0.056)	(0.056)	(0.058)	(0.058)
Sex (Female = 1)		0.052		0.038		-0.089^{*}
		(0.040)		(0.039)		(0.040)
Age (years)		$0.003^{\dagger}$		$-0.003^{\dagger}$		0.0002
		(0.002)		(0.002)		(0.002)
Constant	0.400^{***}	0.271^{***}	0.207^{***}	0.298^{***}	0.393^{***}	0.431^{***}
	(0.041)	(0.071)	(0.040)	(0.070)	(0.041)	(0.072)
Observations	554	554	554	554	554	554
R^2	0.008	0.017	0.030	0.038	0.019	0.028
Adjusted R^2	0.002	0.008	0.025	0.029	0.014	0.019

Note: $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Test for the "Explanation enhances suffering" hypothesis

Table 35: Regression results: Likelihood of the Recipient choosing MODERATE punishment in Study 10,controlling for the Recipient's belief about the Allocator's suffering upon receiving the moderately reduced bonus (moderate punishment)

	Dependent variable:		
	Likelihood of choosing MODERATE punishment		
	(1)	(2)	(3)
Suffer (ign. = 0; suff., just., rev. = 1)	0.063	0.056	0.063
	(0.045)	(0.057)	(0.057)
Explain (ign., suff. = 0; just., rev. = 1)	0.150^{***}	0.150^{***}	0.151^{***}
	(0.045)	(0.045)	(0.045)
Identity (ign., suff., just. = 0; rev. = 1)	$-0.077^{\dagger}$	$-0.077^{\dagger}$	$-0.077^{\dagger}$
	(0.045)	(0.045)	(0.045)
Recipient's belief about Allocator's suffering		−0.0001	0.0001
		(0.0004)	(0.0004)
Sex (Female = 1)			0.036
			(0.032)
Age (years)			$-0.002^{\dagger}$
			(0.001)
Constant	0.201^{***}	0.204^{***}	0.269^{***}
	(0.032)	(0.036)	(0.058)
Observations	805	805	805
R^2	0.030	0.030	0.035
Adjusted R^2	0.026	0.025	0.028
Note:	$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$		

Test for the "Deterrence" alternative hypothesis

Table 36: Regression results: Likelihood of the Recipient choosing the MODERATE punishment option in Study 10, controlling for the Recipient's motive for improving the Allocator's future behavior.

	Dependent variable:		
	Likelihood of choosing MODERATE punishment		
	(1)	(2)	(3)
Suffer (ign. = 0; suff., just., rev. = 1)	0.063	0.024	0.020
	(0.045)	(0.045)	(0.045)
Explain (ign., suff. = 0; just., rev. = 1)	0.150^{***}	0.124^{**}	0.125^{**}
	(0.045)	(0.045)	(0.045)
Identity (ign., suff., just. = 0; rev. = 1)	$-0.077^{\dagger}$	$-0.075^{\dagger}$	$-0.075^{\dagger}$
	(0.045)	(0.045)	(0.044)
Future behavior of the Allocator		0.002^{***}	0.002^{***}
		(0.0004)	(0.0004)
Sex (Female = 1)			0.032
			(0.032)
Age (years)			$-0.002^{\dagger}$
			(0.001)
Constant	0.201^{***}	0.210^{***}	0.271^{***}
	(0.032)	(0.032)	(0.056)
Observations	805	805	805
R^2	0.030	0.058	0.062
Adjusted R^2	0.026	0.053	0.055
Note:			$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

This page intentionally left blank

References

Abelson, R. P. (1986). Beliefs Are Like Possessions. *Journal for the Theory of Social Behaviour*, *16*(3), 223–250. doi: 10.1111/j.1468-5914.1986.tb00078.x

Acemoglu, D., & Restrepo, P. (2018). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review*, *108*(6), 1488–1542. doi: 10.1257/aer.20160696

Acemoglu, D., & Restrepo, P. (2020). Robots and Jobs: Evidence from US Labor Markets. *Journal of Political Economy*, *128*(6), 2188–2244. doi: 10.1086/705716

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity*. *Quarterly Journal of Economics*, *115*(3), 715–753. doi: 10.1162/003355300554881

Andreoni, J., & Rao, J. M. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics*, *95*(7-8), 513–520. doi: 10.1016/j.jpubeco.2010.12.008

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. doi: 10.1086/691703

Aquino, K., Tripp, T. M., & Bies, R. J. (2006). Getting even or moving on? Power, procedural justice, and types of offense as predictors of revenge, forgiveness, reconciliation, and avoidance in organizations. *Journal of Applied Psychology*, *91*(3), 653–668. doi: 10.1037/0021-9010.91.3.653

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. doi: 10.1257/aer.99.1.544

Ariely, D., Kamenica, E., & Prelec, D. (2008). Man's search for meaning: The case of Legos. *Journal of Economic Behavior & Organization*, *67*(3-4), 671–677. doi: 10.1016/j.jebo.2008.01.004

Ariely, D., & Norton, M. I. (2009). Conceptual Consumption. *Annual Review of Psychology*, *60*(1), 475–499. doi: 10.1146/annurev.psych.60.110707.163536

Babcock, L., Loewenstein, G., & Issacharoff, S. (1997). Creating Convergence: Debiasing

Biased Litigants. *Law & Social Inquiry*, *22*(04), 913–925. doi: 10.1111/j.1747-4469.1997.tb01092.x

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. doi: 10.1126/science.aaa1160

Bar-Anan, Y., Liberman, N., Trope, Y., & Algom, D. (2007). Automatic processing of psychological distance: Evidence from a Stroop task. *Journal of Experimental Psychology: General*, *136*(4), 610–622. doi: 10.1037/0096-3445.136.4.610

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. doi: 10.1016/j.evolhumbehav.2006.01.003

Bateman, A. W., & Fonagy, P. (2012). *Handbook of mentalizing in mental health practice*. American Psychiatric Publishing, Inc.

Battigalli, P., Corrao, R., & Dufwenberg, M. (2019). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, *167*(February), 185–218. doi: 10.1016/j.jebo.2019.04.009

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, *97*(2), 170–176. doi: 10.1257/aer.97.2.170

Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *The Review of Economic Studies*, *80*(2), 429–462. doi: 10.1093/restud/rds030

Bénabou, R., & Tirole, J. (2011). Identity, Morals, and Taboos: Beliefs as Assets. *The Quarterly Journal of Economics*, *126*(2), 805–855. doi: 10.1093/qje/qjr002

Bénabou, R., & Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, *30*(3), 141–164. doi: 10.1257/jep.30.3.141

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*.

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart's Dilemma: On the Social Rewards and Penalties of Advertising Prosocial Behavior. *Journal of Marketing Research*, *52*(1), 90–104. doi: 10.1509/jmr.14.0002

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica:*

Journal of the Econometric Society, *22*(1), 23–36.

Bessi, A., Petroni, F., Vicario, M. D., Zollo, F., Anagnostopoulos, A., Scala, A., ... Quattrociocchi, W. (2016). Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics*, *225*(10), 2047–2059. doi: 10.1140/epjst/e2015-50319-0

Bierbrauer, G. (1979). Why did he do it? attribution of obedience and the phenomenon of dispositional bias. *European Journal of Social Psychology*, *9*(1), 67–84. doi: 10.1002/ejsp.2420090106

Bies, R. J., & Tripp, T. M. (1998). Revenge in organizations: The good, the bad, and the ugly. In R. W. Griffin, A. O'Leary-Kelly, & J. M. Collins (Eds.), *Monographs in organizational behavior and industrial relations, vol. 23, parts a & b. dysfunctional behavior in organizations: Violent and deviant behavior* (pp. 49–67). Elsevier Science/JAI Press.

Bies, R. J., & Tripp, T. M. (2006). The Study of Revenge in the Workplace: Conceptual, Ideological, and Empirical Issues. In *Counterproductive work behavior: Investigations of actors and targets.* (pp. 65–81). Washington: American Psychological Association. doi: 10.1037/10893-003

Bishop, B. (2009). *The big sort: Why the clustering of like-minded America is tearing us apart*. Houghton Mifflin Harcourt.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193. doi: 10.1257/aer.90.1.166

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. *Science*, *328*(5978), 617–620. doi: 10.1126/science.1183665

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*(2), 135–143. doi: 10.1002/ejsp.744

Camerer, C., Loewenstein, G., & Weber, M. (1989). The Curse of Knowledge in Economic Settings: An Experimental Analysis. *Journal of Political Economy*, *97*(5), 1232–1254. doi: 10.1086/261651

Cappelen, A. W., Halvorsen, T., Soerensen, E. O., & Tungodden, B. (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association*, *15*(3), 540–557. doi: 10.1093/jeea/jvw014

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. doi: 10.1016/j.jesp.2005.06.007

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. doi: 10.1037/0022-3514.83.2.284

Carpenter, J. P., & Matthews, P. H. (2003). Beliefs, intentions, and evolution: Old versus new psychological game theory. *Behavioral and Brain Sciences*, *26*(02), 158–159. doi: 10.1017/S0140525X03270059

Carpenter, J. P., Matthews, P. H., & Ong'ong'a, O. (2004). Why Punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, *14*(4), 407–429. doi: 10.1007/s00191-004-0212-1

Carter, S. L. (1996). *Integrity*. New York: Basic Books.

Chander, A. (2016). The Racist Algorithm. *Michigan Law Review*, *115*(6), 1023–1045.

Chandler, M. J. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology*, *9*(3), 326–332. doi: 10.1037/h0034974

Charness, G., & Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, *74*(6), 1579–1601. doi: 10.1111/j.1468-0262.2006.00719.x

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. doi: 10.1016/j.jebo.2011.08.009

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

Cigna. (2018). *New Cigna Study Reveals Loneliness at Epidemic Levels in America.* Retrieved from https://www.cigna.com/newsroom/news-releases/2018/new

-cigna-study-reveals-loneliness-at-epidemic-levels-in-america

Converse, B. A., Lin, S., Keysar, B., & Epley, N. (2008). In the mood to get over yourself: Mood affects theory-of-mind use. *Emotion*, *8*(5), 725–730. doi: 10.1037/a0013283

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279–2286. doi: 10.1037/xge0000018

Crombag, H., Rassin, E., & Horselenberg, R. (2003). On vengeance. *Psychology, Crime & Law*, *9*(4), 333–344. doi: 10.1080/1068316031000068647

Crone, E. A., Bullens, L., van der Plas, E. A. A., Kijkuit, E. J., & Zelazo, P. D. (2008). Developmental changes and individual differences in risk and perspective taking in adolescence. *Development and Psychopathology*, *20*(4), 1213–1229. doi: 10.1017/S0954579408000588

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97–103. doi: 10.1016/j.copsyc.2015.06.003

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. doi: 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. doi: 10.1007/s00199-006-0153-z

Darley, J. M., & Pittman, T. S. (2003). The Psychology of Compensatory and Retributive Justice. *Personality and Social Psychology Review*, *7*(4), 324–336. doi: 10.1207/S15327957PSPR0704{_}05

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. doi: 10.1073/pnas.1517441113

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*,

1(4), 568–570. doi: 10.1017/S0140525X00076664

Dickinson, D. L., & Masclet, D. (2015). Emotion venting and punishment in public good experiments. *Journal of Public Economics*, *122*, 55–67. doi: 10.1016/j.jpubeco.2014.10.008

Dorison, C. A., Minson, J. A., & Rogers, T. (2019). Selective exposure partly relies on faulty affective forecasts. *Cognition*, *188*(May 2018), 98–107. doi: 10.1016/j.cognition.2019.02.010

Dufwenberg, M., & Gneezy, U. (2000). Measuring Beliefs in an Experimental Lost Wallet Game. *Games and Economic Behavior*, *30*(2), 163–182. doi: 10.1006/game.1999.0715

Dumsday, T. (2009). On Cheering Charles Bronson: The Ethics of Vigilantism. *The Southern Journal of Philosophy*, *47*(1), 49–67. doi: 10.1111/j.2041-6962.2009.tb00131.x

Eadeh, F. R., Peak, S. A., & Lambert, A. J. (2017). The bittersweet taste of revenge: On the negative and positive consequences of retaliation. *Journal of Experimental Social Psychology*, *68*, 27–39. doi: 10.1016/j.jesp.2016.04.007

EPA. (1993). *U.S. Environmental Protection Agency Oral History Interview1 - William D. Ruckelshaus*. U.S. Environmental Protection Agency.

Fabrizio, T. (2016). *"Lesser of Two Evils" Takes On New Meaning in 2016.*

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. doi: 10.3758/BRM.41.4.1149

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. doi: 10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, *14*(3), 159–182. doi: 10.1257/jep.14.3.159

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. doi: 10.1038/415137a

Feinberg, J. (1965). The Expressive Function of Punishment. *Monist*, *49*(3), 397–423. doi: 10.5840/monist196549326

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. doi: 10.1016/j.tics.2006.11.005

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. doi: 10.1093/poq/nfw006

French, P. (2001). *The virtues of vengeance*. Kansas: The University Press of Kansas.

Frijda, N. H. (1994). The lex talionis: On vengeance. In *Emotions: Essays on emotion theory* (pp. 263–289).

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*(July 2016), 1–12. doi: 10.1016/j.jesp.2017.04.003

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 459–473. doi: 10.1098/rstb.2002.1218

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the Message. *Personality and Social Psychology Bulletin*, *40*(8), 986–997. doi: 10.1177/0146167214533130

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-Taking and Self-Other Overlap: Fostering Social Bonds and Facilitating Social Coordination. *Group Processes & Intergroup Relations*, *8*(2), 109–124. doi: 10.1177/1368430205051060

Ganguly, A., & Tasoff, J. (2017). Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future. *Management Science*, *63*(12), 4037–4060. doi: 10.1287/mnsc.2016.2550

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79. doi: 10.1016/0899-8256(89)90005-5

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38. doi: 10.1037/0033-2909.117.1.21

Gill, D., & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, *102*(1), 469–503. doi: 10.1257/

aer.102.1.469

Glaeser, E. L., & Sacerdote, B. (2000). The Determinants of Punishment: Deterrence, Incapacitation and Vengeance. *SSRN Electronic Journal*(April). doi: 10.2139/ssrn.236443

Goldstein, T. R., & Winner, E. (2012). Enhancing Empathy and Theory of Mind. *Journal of Cognition and Development*, *13*(1), 19–37. doi: 10.1080/15248372.2011.573514

Gollwitzer, M., & Bushman, B. J. (2012). Do Victims of Injustice Punish to Improve Their Mood? *Social Psychological and Personality Science*, *3*(5), 572–580. doi: 10.1177/1948550611430552

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, *45*(4), 840–844. doi: 10.1016/j.jesp.2009.03.001

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*(3), 364–374. doi: 10.1002/ejsp.782

Golman, R. (2016). Good manners: signaling social preferences. *Theory and Decision*, *81*(1), 73–88. doi: 10.1007/s11238-015-9527-7

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information Avoidance. *Journal of Economic Literature*, *55*(1), 96–135. doi: 10.1257/jel.20151245

Golman, R., Loewenstein, G., Moene, K. O., & Zarri, L. (2016). The Preference for Belief Consonance. *Journal of Economic Perspectives*, *30*(3), 165–188. doi: 10.1257/jep.30.3.165

Golman, R., Loewenstein, G. F., Molnar, A., & Saccardo, S. (2019). The Demand for, and Avoidance of, Information. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2149362

Goodwin, G. P. (2015). Moral Character in Person Perception. *Current Directions in Psychological Science*, *24*(1), 38–44. doi: 10.1177/0963721414550709

Grégoire, Y., Laufer, D., & Tripp, T. M. (2010). A comprehensive model of customer direct and indirect revenge: understanding the effects of perceived greed and customer power. *Journal of the Academy of Marketing Science*, *38*(6), 738–758. doi: 10.1007/

s11747-009-0186-5

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. doi: 10 .1093/jeea/jvw001

Hauser, O. P., Gino, F., & Norton, M. I. (2018). Budging beliefs, nudging behaviour. *Mind & Society*, *17*(1-2), 15–26. doi: 10.1007/s11299-019-00200-9

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243. doi: 10.2307/1416950

Heintz, C., Karabegovic, M., & Molnar, A. (2016). The Co-evolution of Honesty and Strategic Vigilance. *Frontiers in Psychology*, *7*(October), 1–13. doi: 10.3389/fpsyg .2016.01503

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... Ziker, J. (2006). Costly Punishment Across Human Societies. *Science (New York, N.Y.)*, *312*(5781), 1767–70. doi: 10.1126/science.1127333

Hertwig, R., & Engel, C. (2016). Homo Ignorans. *Perspectives on Psychological Science*, *11*(3), 359–372. doi: 10.1177/1745691616635594

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091–1243091. doi: 10.1126/science.1243091

Ho, B. (2012). Apologies as Signals: With Evidence from a Trust Game. *Management Science*, *58*(1), 141–158. doi: 10.1287/mnsc.1110.1410

Ho, B., & Liu, E. (2011). Does sorry work? The impact of apology laws on medical malpractice. *Journal of Risk and Uncertainty*, *43*(2), 141–167. doi: 10.1007/s11166 -011-9126-0

Ho, R., ForsterLee, L., ForsterLee, R., & Crofts, N. (2002). Justice versus vengeance: motives underlying punitive judgements. *Personality and Individual Differences*, *33*(3), 365–377. doi: 10.1016/S0191-8869(01)00161-1

Huber, P. (1989). No-Fault Punishment. *Alabama Law Review*, *40*(3), 1037–1052.

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence

on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. doi: 10.1111/ajps.12152

Jacoby, S. (1983). *Wild justice: The evolution of revenge*. Harper Collins.

Johns, T., Laubscher, R. J., & Malone, T. W. (2011). The Big Idea: The Age of Hyperspecialization. *Harvard Business Review*, *89*(7-8), 56.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, *102*(3), 192–194. doi: 10.1016/j.econlet.2009.01.003

Jones, E. E., & Davis, K. E. (1965). From Acts To Dispositions The Attribution Process In Person Perception. In *Advances in experimental social psychology* (pp. 219–266). doi: 10.1016/S0065-2601(08)60107-0

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. doi: 10.1016/0022-1031(67)90034-0

Jordan, J., McAuliffe, K., & Rand, D. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*(4), 741–763. doi: 10.1007/s10683-015-9466-8

Kahan, D. M. (2015). Climate-Science Communication and the Measurement Problem. *Political Psychology*, *36*(S1), 1–43. doi: 10.1111/pops.12244

Kant, I. (1952). The science of right (W. Hastie, Trans.). In R. Hutchins (Ed.), *Great books ofthe western world: Vol. 42. kant* (p. 397–446). Chicago: Encyclopedia Brittanica.

Kant, I. (1991). *The metaphysics of morals*. Cambridge University Press.

Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, *38*(2), 95–115. doi: 10.1007/s11166-009-9060-6

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204. doi: 10.1111/j.2044-8295.1998.tb02680.x

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the

Age of Algorithms. *Journal of Legal Analysis*, *10*(2005), 113–174. doi: 10.1093/jla/laz001

Leith, K. P., & Baumeister, R. F. (1998). Empathy, Shame, Guilt, and Narratives of Interpersonal Conflicts: Guilt-Prone People Are Better at Perspective Taking. *Journal of Personality*, *66*(1), 1–37. doi: 10.1111/1467-6494.00001

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind.". *Psychological Review*, *94*(4), 412–426. doi: 10.1037/0033-295X.94.4.412

Litman, J. A., & Jimerson, T. L. (2004). The Measurement of Curiosity As a Feeling of Deprivation. *Journal of Personality Assessment*, *82*(2), 147–157. doi: 10.1207/s15327752jpa8202{_}3

Loewenstein, G. (1992). The fall and rise of psychological explanation in the economics of intertemporal choice. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 3–34). New York: Russell Sage Foundation.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. doi: 10.1037/0033-2909.116.1.75

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, *2*(3), 166–167. doi: 10.1038/s41562-018-0301-z

Lun, J., Kesebir, S., & Oishi, S. (2008). On feeling understood and feeling well: The role of interdependence. *Journal of Research in Personality*, *42*(6), 1623–1628. doi: 10.1016/j.jrp.2008.06.009

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology*, *68*(5), 854–869. doi: 10.1037/0022-3514.68.5.854

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, *93*(1), 366–380. doi: 10.1257/000282803321455359

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities. *American Behavioral Scientist*, *62*(1), 16–42. doi: 10.1177/0002764218759576

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(01), 1–15. doi: 10.1017/S0140525X11002160

Miller, D. T. (2001). Disrespect and the Experience of Injustice. *Annual Review of Psychology*, *52*(1), 527–553. doi: 10.1146/annurev.psych.52.1.527

Miller, D. T., & Nelson, L. D. (2002). Seeing approach motivation in the avoidance behavior of others: Implications for an understanding of pluralistic ignorance. *Journal of Personality and Social Psychology*, *83*(5), 1066–1075. doi: 10.1037/0022-3514.83.5.1066

Millgram, E. (2015). *The great endarkenment: philosophy for an age of hyperspecialization*. Oxford University Press.

Minson, J. A., Chen, F. S., & Tinsley, C. H. (2019). Why Won't You Listen to Me? Measuring Receptiveness to Opposing Views. *Management Science*(May 2020), mnsc.2019.3362. doi: 10.1287/mnsc.2019.3362

Mischel, W. (1977). On the future of personality measurement. *American Psychologist*, *32*(4), 246–254. doi: 10.1037/0003-066X.32.4.246

Mitchell, A., Gottfried, J., Kiley, J., & Matsa, K. E. (2014). Political polarization & media habits. *Pew Research Center*, *21*. doi: 202.419.4372

Molnar, A. (2019). SMARTRIQS: A Simple Method Allowing Real-Time Respondent Interaction in Qualtrics Surveys. *Journal of Behavioral and Experimental Finance*, *22*, 161–169. doi: 10.1016/j.jbef.2019.03.005

Molnar, A. (2020). How to implement real-time interaction between participants in online surveys: A practical guide to SMARTRIQS. *The Quantitative Methods for Psychology*, *16*(4), 334–354. doi: 10.20982/tqmp.16.4.p334

Molnar, A., Chaudhry, S., & Loewenstein, G. F. (2020). 'It's Not About the Money. It's About Sending a Message!': Unpacking the Components of Revenge. *SSRN Electronic Journal*. Retrieved from https://www.ssrn.com/abstract=3524910 doi: 10.2139/ssrn.3524910

Molnar, A., & Chaudhry, S. J. (2020). The lesser of two evils: Explaining a bad choice by revealing the choice set. *PsyArXiv Preprints*. Retrieved from https://psyarxiv.com/

8sdme/ doi: 10.31234/osf.io/8sdme

Molnar, A., & Loewenstein, G. F. (2020). The Othello Effect: People Are More Disturbed by Others' Wrong Beliefs Than by Different Beliefs. *SSRN Electronic Journal*. Retrieved from https://www.ssrn.com/abstract=3524651 doi: 10.2139/ssrn.3524651

Morelli, S. A., Torre, J. B., & Eisenberger, N. I. (2014). The neural bases of feeling understood and not understood. *Social Cognitive and Affective Neuroscience*, *9*(12), 1890–1896. doi: 10.1093/scan/nst191

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1–14. doi: 10.1016/j.jesp.2013.10.010

Mutz, D. C. (2018). Status threat, not economic hardship, explains the 2016 presidential vote. *Proceedings of the National Academy of Sciences*, *115*(19), E4330-E4339. doi: 10.1073/pnas.1718155115

Neumann, J. V., & Morgenstern, O. (1947). Theory of games and economic behavior: the Notion of Utility. In (pp. 15–29).

Nisbett, R. E., & Ross, L. (1991). *The person and the situation*. New York: McGraw-Hill.

Oishi, S., Schiller, J., & Gross, E. B. (2013). Felt Understanding and Misunderstanding Affect the Perception of Pain, Slant, and Distance. *Social Psychological and Personality Science*, *4*(3), 259–266. doi: 10.1177/1948550612453469

Osgood, J. M. (2017). Is revenge about retributive justice, deterring harm, or both? *Social and Personality Psychology Compass*, *11*(1), 1–15. doi: 10.1111/spc3.12296

Oster, E., Shoulson, I., & Dorsey, E. R. (2013). Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease. *American Economic Review*, *103*(2), 804–830. doi: 10.1257/aer.103.2.804

Parker, L. (1983). Tacoma Must Weigh Clean Air Against Jobs. *The Washington Post*. Retrieved from https://www.washingtonpost.com/archive/politics/1983/12/11/tacoma-must-weigh-clean-air-against-jobs/5c1b77f0-7108-4d04-be3a-e24b97ca502b/?utm_term=.b71ccd1880c6

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution

of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471. doi: 10.1016/0022-0965(85)90051-7

Porter, E. (2016). Where Were Trump's Votes? Where the Jobs Weren't. *The New York Times*. Retrieved from https://www.nytimes.com/2016/12/13/business/economy/jobs-economy-voters.html

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. doi: 10.1017/S0140525X00076512

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243–256. doi: 10.1037/0022-3514.64.2.243

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*, *111*(3), 781–799. doi: 10.1037/0033-295X.111.3.781

Qualtrics. (2020). Provo, Utah, USA: Qualtrics. Retrieved from https://www.qualtrics.com

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology*, *42*(4), 593–607. doi: 10.1037/0022-3514.42.4.593

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*(2), 98–103. doi: 10.1016/j.tree.2014.12.003

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters*, *8*(5), 802–804. doi: 10.1098/rsbl.2012.0470

Rand, D. G., Fudenberg, D., & Dreber, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization*, *116*, 481–499. doi: 10.1016/j.jebo.2015.05.013

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2005). On Attributing Negative Motives to Others Who Disagree With Our Opinions. *Personality and Social Psychology Bulletin*, *31*(11), 1498–1510. doi: 10.1177/0146167205277093

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Samuelson, P. A. (1948). Consumption Theory in Terms of Revealed Preference. *Economica*, *15*(60), 243. doi: 10.2307/2549561

Santos, M. d., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371–377. doi: 10.1098/rspb.2010.1275

Sarin, A., Ho, M., Martin, J., & Cushman, F. (2020). *Punishment is Organized around Principles of Communicative Inference.* doi: 10.31234/osf.io/2cyf7

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Saxe, L. (1991). Lying: Thoughts of an applied social psychologist. *American Psychologist*, *46*(4), 409–415. doi: 10.1037/0003-066X.46.4.409

Schneider, D., Slaughter, V. P., & Dux, P. E. (2015). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*, *22*(1), 1–12. doi: 10.3758/s13423-014-0644-z

Schumann, K., & Ross, M. (2010). The Benefits, Costs, and Paradox of Revenge. *Social and Personality Psychology Compass*, *4*(12), 1193–1205. doi: 10.1111/j.1751-9004.2010.00322.x

Schwarz, N. (2012). Feelings-as-Information Theory. In *Handbook of theories of social psychology: Volume 1* (Vol. 1, pp. 289–308). SAGE Publications Ltd. doi: 10.4135/9781446249215.n15

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*(2), 317–338. doi: 10.1037/0022-3514.65.2.317

Sicherman, N., Loewenstein, G., Seppi, D. J., & Utkus, S. P. (2016). Financial Attention. *Review of Financial Studies*, *29*(4), 863–897. doi: 10.1093/rfs/hhv073

Soetevent, A. R. (2011). Payment Choice, Image Motivation and Contributions to Charity: Evidence from a Field Experiment. *American Economic Journal: Economic Policy*,

3(1), 180–205. doi: 10.1257/pol.3.1.180

Sommers, T. (2009). The two faces of revenge: moral responsibility and the culture of honor. *Biology & Philosophy*, *24*(1), 35–50. doi: 10.1007/s10539-008-9112-3

Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind & Language*, *17*(1-2), 3–23. doi: 10.1111/1468-0017.00186

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105. doi: 10.1016/0022-1031(91)90011-T

Stephan, W. G., & Finlay, K. (1999). The Role of Empathy in Improving Intergroup Relations. *Journal of Social Issues*, *55*(4), 729–743. doi: 10.1111/0022-4537.00144

Sunstein, C. R. (2001). *Republic.com*. Princeton university press.

Sunstein, C. R., Kahneman, D., & Schkade, D. (1998). Assessing Punitive Damages (With Notes on Cognition and Valuation in Law). *The Yale Law Journal*, *107*(7), 2071. doi: 10.2307/797417

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*(1), 69–78. doi: 10.1016/j.joep.2006.09.001

Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*(4), 457–477. doi: 10.1037/0033-295X.91.4.457

Swann, W. B. (2011). Self-verification theory. In *Handbook of theories of social psychology (vol. 2)* (pp. 23–42).

Swann, W. B., Pelham, B. W., & Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *Journal of Personality and Social Psychology*, *57*(5), 782–791. doi: 10.1037/0022-3514.57.5.782

Swann, W. B. J., Rentfrow, P. J., & Guinn, J. S. (2003). Self-verification: The Search for Coherence. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 367–383). New York, NY: Guilford Press.

Swift, J. (1801). Thoughts on Various Subjects. In *The works of the rev. jonathan swift*

(vol. 5) (pp. 453–465).

te Velde, V. L. (2018). Beliefs-based altruism as an alternative explanation for social signaling behaviors. *Journal of Economic Behavior & Organization*, *152*, 177–191. doi: 10.1016/j.jebo.2018.06.011

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, *117*(12), 6370–6375. doi: 10.1073/pnas.1910402117

Tripp, T. M., Bies, R. J., & Aquino, K. (2002). Poetic justice or petty jealousy? The aesthetics of revenge. *Organizational Behavior and Human Decision Processes*, *89*(1), 966–984. doi: 10.1016/S0749-5978(02)00038-9

Trope, Y., & Liberman, A. (1993). The Use of Trait Conceptions to Identify Other People's Behavior and to Draw Inferences about their Personalities. *Personality and Social Psychology Bulletin*, *19*(5), 553–562. doi: 10.1177/0146167293195007

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. doi: 10.1037/a0018963

Van Boven, L., Ehret, P. J., & Sherman, D. K. (2018). Psychological Barriers to Bipartisan Public Support for Climate Policy. *Perspectives on Psychological Science*, *13*(4), 492–507. doi: 10.1177/1745691617748966

Van Boven, L., Judd, C. M., & Sherman, D. K. (2012). Political polarization projection: Social projection of partisan attitude extremity and attitudinal processes. *Journal of Personality and Social Psychology*, *103*(1), 84–100. doi: 10.1037/a0028145

Van der Graaff, J., Branje, S., De Wied, M., Hawk, S., Van Lier, P., & Meeus, W. (2014). Perspective taking and empathic concern in adolescence: Gender differences in developmental changes. *Developmental Psychology*, *50*(3), 881–888. doi: 10.1037/a0034325

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*(1), 128–133. doi: 10.1016/j.cognition.2013.10.004

Vergani, M., Iqbal, M., Ilbahar, E., & Barton, G. (2020). The Three Ps of Radicalization:

Push, Pull and Personal. A Systematic Scoping Review of the Scientific Evidence about Radicalization Into Violent Extremism. *Studies in Conflict & Terrorism*, *43*(10), 854–854. doi: 10.1080/1057610X.2018.1505686

Vonnegut, K. (1987). *Bluebeard*. Delacorte Press.

Wilde, O. (1891). *The Picture of Dorian Gray* (1985th ed.). Harmondsworth: Penguin Books.

Wood, W., Pool, G. J., Leck, K., & Purvis, D. (1996). Self-definition, defensive processing, and influence: The normative impact of majority and minority groups. *Journal of Personality and Social Psychology*, *71*(6), 1181–1193. doi: 10.1037/0022-3514.71.6.1181

Wuthnow, R. (2019). *The left behind: Decline and rage in small-town America*. Princeton University Press.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*(20), 7398–7401. doi: 10.1073/pnas.0502399102

Yang, Y. C., Boen, C., Gerken, K., Li, T., Schorpp, K., & Harris, K. M. (2016). Social relationships and physiological determinants of longevity across the human life span. *Proceedings of the National Academy of Sciences*, *113*(3), 578–583. doi: 10.1073/pnas.1511085112

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. doi: 10.1037/pspa0000056

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98. doi: 10.1007/s10683-009-9230-z

This dissertation has 198 references.