Social Decision-Making Under Uncertainty

Anne-Marie Nussberger
Brasenose College
University of Oxford

Thesis Submitted for the Degree of
Doctor of Philosophy
May 2021

Abstract

Social decisions are fraught with uncertainty. In my thesis, I explore how uncertainty shapes social decision-making across three empirical projects (9 studies, total *N* = 3,346).

I start by comparing social and individual decisions under risk (i.e., choices over outcomes with known probabilities) across a series of online studies. I find that choices for others are markedly more risk-averse than choices for self because they are more attuned to disadvantageous prospects. Supporting the hypothesis that enhanced uncertainty in social decisions might produce such a precautionary pattern, I observe greater indecision and less confidence in social relative to individual choices, while motivational differences cannot explain their divergence.

As a next step, I set out to experimentally manipulate uncertainty about others' experiences of outcomes in a laboratory experiment. My observations suggest that enhanced uncertainty about others' experience plays a supporting role in a particular manifestation of precautionary preferences for others, namely a stronger aversion to harm others than oneself.

Finally, I consider how uncertainty shapes interactions between humans and artificially intelligent machines as a broader category of social decisions. I find that people's need to explain others' minds and actions extends to interactions with machines. This demand to reduce uncertainty about machines' inner workings is more pronounced for settings involving high stakes and/or scarcity – a pattern paralleling demand for explanation in realms evading machines.

Two central insights emerge: first, enhanced uncertainty about others' experience of outcomes in social choices elicits precautionary preferences whereby decision-makers minimise potential harm to others. Second, humans' notorious need to reduce uncertainty about others' internal states extends to interactions with artificial social agents, where it translates into a robust demand for interpretability of algorithmic decisions.

Acknowledgments

Living up to its subject, my doctoral research has been characterised as much by *uncertainty* as by *others* who supported me along the way. The unexpected permanent closure of my departmental building, or both supervisors moving to different universities, might have made for a volatile environment – but numerous persistent, faithful, and unconditionally reliable people have helped me to embrace these twists as opportunities to learn:

My supervisor Molly Crockett inspired me to explore how uncertainty pertaining to others shapes our decision-making. She nurtured this interest, from close and afar, by nudging me towards different methods and realms of exploration. I have also learned an invaluable lot just from observing Molly's brilliant mind at work. When the Crockett Lab moved across the pond, Nadira Faber readily adopted me in her group and sympathetically chaperoned me through grant and funding applications, project ups and downs, and the twists and turns of life. I feel lucky and immensely grateful for being supervised by two stirring female scientists from whom I feared no judgment but felt wholehearted supported, not at last while becoming a mother.

Andreas Kappes' encouragement and role-modelling for the compatibility of being an early career scientist and a parent has been as much invaluable, as his omnipresent support in times of struggle, and his firing ideas about research. Julian Savulescu gave me his generous support across various stages of my doctorate – in basic currencies (money for research and an office roof above the head), but also in inspirational multidisciplinary collaborations, and the opportunity to teach in South Korea. Geoff Bird took me under his wings as departmental supervisor when both Molly and Nadira relocated. Jim Everett always has been ready to endure me moaning about the latest stats struggles and to have a polishing look at my writing. Jen Siegel's and Hongbo Yu's sharp minds gently introduced me to methodological territories I had not charted before. Talia Longthorne and Eloise Copeland excelled in assisting me to collect data in laborious experiments involving electric shocks, and cognitive enhancement drugs. Stefan Schubert and Lucius Caviola were stimulating labmates, also after deducing Stefan's and my caffein intake. Anthony Gabay,

Luis Sebastian Contreras-Huerta, Patricia Lockwood, and Matthew Apps were never amiss for a nerdy and (yet) cool drink, combined with addictive climbing in Anthony's case.

I am grateful for the generous support of the Studienstiftung des deutschen Volkes. This bursary has been essential for affording my living during the doctorate, and for paying university fees. The Studienstiftung's flexibility when it comes to unexpected changes in timelines and their unconditional support for graduate students who become parents might be unique when it comes to doctorate funding. I feel very lucky for their continued support since my undergraduate days and look forward to contributing my share in the future.

I am most grateful to my friends and family, who have never lost sight of me – no matter how far away I was physically, no matter how absorbed I was mentally. Ben, my partner in life and crime (as far as you get with a lawyer): ever since our days as undergraduate students in Heidelberg, you believed in my ability to work towards something like this thesis, at a place something like Oxford. You are the one who dares me going one step further; who picks me up when I get stuck; and who reassures me in the face of uncertainty. Lately, you are also the one who lovingly distracts our daughter Felicitas from the fact that I sometimes must work, and who does everything to make this possible in the middle of a raging pandemic, childcare outages, and anything that comes our way. Grand-mère et grand-père, my parents, sisters and their families, my family-in-law, and my friends: you have always been on my side and have encouraged me to follow my curiosity – no matter how far away it might have led me away from you. Your support has been manifold: from nurturing care parcels, over spaces and times to unwind, to unfailing trust that, at some point, I would complete my doctorate. If it wasn't for you, this moment would not have come into reach.

Anne-Marie Nussberger

Berlin, May 2021

Contents

1 Introduction 1

Varieties of uncertainty: Existing distinctions 2
Varieties of uncertainty: The social dimension 5
Uncertainty in a different variety of "social": Human-machine interactions 10
Contribution to the literature 13
Overview of empirical chapters 14
Nomenclature 18

2 Social versus individual decision-making under risk 19

Study 2.1: Self-other differences in hypothetical choices under risk 25
Study 2.2: Self-other differences in incentivised choices under risk 37
Study 2.3: Dissociating the roles of uncertainty and motivation for self-other differences in choices under risk 46
Overall discussion 62

3 Uncertainty's role in harm aversion for others versus self 66

Study 3: Hyperaltruism under experimentally enhanced impact uncertainty 68
Discussion 92

4 Uncertainty in human-machine interactions 98

Study 4.1A: Demand for AI interpretability across a variety of applications 103
Study 4.1B: Replicating demand for AI interpretability in representative US and UK samples 108
Study 4.2: How stakes and scarcity drive demand for interpretability in AI 112
Study 4.3: How stakes and scarcity impact tradeoffs between AI interpretability and accuracy 117
Study 4.4: Stakes and scarcity impact interpretability-accuracy tradeoffs across stakeholder perspectives 121
Overall discussion 126

5 Concluding discussion....132

Summary of key findings 132
Uncertainty's contribution to self-other differences in human decision-making 134
Uncertainty in interactions with non-human agents 136
Varieties of uncertainty in social decision-making 139
Limitations and critical reflection 141
Future directions 143
Implications 145

References....147

Appendices....165

Appendix A: List of Figures 165
Appendix B: List of Tables 166
Appendix C: Additional procedures and materials 167
Appendix D: Additional results 182

1 Introduction

Uncertainty is a fact with which all forms of life must be prepared to contend.
(Kahneman & Tversky, 1982, p. 144)

Understanding how human decision-makers grapple with uncertainty has been a central endeavour in research spanning psychology, neuroscience, economics, and philosophy among other disciplines. The ensuing work has focused on a particular form of human life: life as individuals, where our choices bear outcomes only for ourselves (e.g., Behrens et al., 2007; Platt & Huettel, 2008). Meanwhile, many – if not most – of our decisions also bear outcomes for others or are responses to others' decisions; they are social in nature, as much as we, as humans, are social by nature (Henrich & Muthukrishna, 2021; Heyes, 2018). Indeed, socially embedded decisions seem key to pursuing a healthy (Berkman et al., 2000), long (Holt-Lunstad et al., 2015), satisfied (Rohrer et al., 2018), and happy life (Huppert et al., 2004). This raises the question how we navigate uncertainty in social settings, where the inescapable reality that we can never be sure about others' subjective experiences (Harsanyi, 1955; Nagel, 1974) magnifies any uncertainty encountered in solitary choices – from perceptual uncertainty about features of our environment, over uncertainty about possible outcomes, to uncertainty about feasible actions. My thesis explores uncertainty in the social domain by considering different varieties of uncertainty and varieties of social interactions.

Varieties of uncertainty: Existing distinctions

The point in distinguishing different varieties of uncertainty goes beyond terminological correctness: it allows identifying and understanding the motivational underpinnings of human behaviour, which hinges on successfully detecting and adaptively responding to uncertainty (Bland & Schaefer, 2012; Camerer & Weber, 1992; Platt & Huettel, 2008). Take, for instance, the early and pivotal distinction between *risk*, referring to situations where the probability distribution over outcomes is known, and *ambiguity*, referring to situations where information about the distribution of outcomes is unknown (Keynes, 1921; F. H. Knight, 1921). Paradoxically, but maybe not surprisingly, people prefer betting on risky rather than ambiguous events – even when an ambiguous event comes with better prospects than a risky event (Camerer & Weber, 1992; Ellsberg, 1961; Furnham & Marks, 2013). As an illustration, imagine I ask you to bet on one of two coins, A and B. You observe me flipping coin A dozens of times and, from observing the outcomes, you conclude that coin A is fair. After flipping coin B twice, where it showed up once with heads and once with tails, I ask you to place your bet. Even though most people perceive coin B as fair as well, they prefer betting on coin A – which is commonly perceived as less ambiguous in terms of fairness (Camerer & Weber, 1992).

People's intuitive distinction between risk and ambiguity is reflected in dissociable physiological (FeldmanHall et al., 2016) and neural signatures (Hsu et al., 2005; Huettel et al., 2006). Functional explanations of ambiguity aversion have proposed, for example, that ambiguity amplifies pessimistic outcome expectancies in compensation for missing probability information and in preparation for undesired outcomes (Erev et al., 2017; Mandel, 2008; Pulford, 2009). However, more recent work has observed a weaker link between

ambiguity-aversion and pessimism, pointing instead towards general caution as an adaptive response amidst insufficient information (Ahrends et al., 2019). While more remains to be learned about the adaptive value of the dissociation between risk and ambiguity, its value as a more nuanced notion of uncertainty has become evident in facilitating the mechanistic understanding of real-world phenomena such as inefficiencies in financial (Epstein & Schneider, 2008; Mukerji & Tallon, 2001) and insurance (Hogarth & Kunreuther, 1989) markets, suboptimal health choices (Han et al., 2007, 2018; Ritov & Baron, 1990), or (lack of) prosocial behaviour (Vives & FeldmanHall, 2018).

Despite this precedency and occasional calls for a more nuanced notion of uncertainty (Dequech, 2004; Rowe, 1994; Teigen, 1994), otherwise only the pair of *epistemic* (knowable) versus *aleatory* (random) uncertainty (Fox & Ülkümen, 2011; Ülkümen et al., 2016) has gained notable prominence (Weber & Johnson, 2009). Epistemic uncertainty signifies missing information or knowledge that could, in principle, be obtained (as in "what is the average temperature in your town?"), whereas aleatory uncertainty relates to unpredictable stochastic events (as in "will it rain in your town at this hour in two months' time?"). People intuitively distinguish between the two variants, as mirrored in their use of confidence statements when referring to epistemic uncertainty (e.g., "I'm fairly confident the average temperature in my town is around 17°C") versus likelihood statements when referring to aleatory uncertainty (e.g., "I think there is a low probability that it will rain in my town at this hour in two months' time"; Ülkümen et al., 2016) and in dissociable neural responses (K. G. Volz et al., 2005). As with the risk-ambiguity distinction, dissociating epistemic and aleatory uncertainty has helped our understanding of real-world phenomena. For instance, in financial trading the salience of epistemic uncertainty promotes additional information-

search whereas the salience of aleatory uncertainty is associated with stronger diversification of portfolios (Walters et al., 2020).

The epistemic-aleatory distinction already surfaced in Kahneman and Tversky's (1982) proposal of distinguishing internal versus external uncertainty and can be traced back to the early days of probability theory (for an overview, see Hacking, 2006 [1975]). Notably, it roughly maps onto two fundamental conceptualisations of probability: the Bayesian tradition, which interprets probabilities as degrees of (subjective) beliefs (e.g., Savage, 1972), and the frequentist tradition, which is concerned with probabilities sampled in sequences of events (e.g., von Mises, 1939). The observation that people intuitively distinguish between epistemic and aleatory uncertainty affirms the validity of the theoretical bifurcation between Bayesian and frequentist approaches. That said, because – realistically – most assessments of stochastic events entail at least some residual uncertainty about one's observations, assumptions, and/or inferences that is epistemic by nature (Camerer & Weber, 1992; Fox & Ülkümen, 2011), human experience of uncertainty might be viewed as inherently subjective in line with the dominant approach in probability theory, that is the Bayesian view[1] (Machina & Schmeidler, 1992; Ramsey, 1926; Savage, 1972; Tversky & Kahneman, 1974, 1992).

It is all the more surprising that the substantive body of work on decision-making under uncertainty – including the contributions which acknowledge the above nuances – has largely ignored a dimension of uncertainty that naturally epitomizes subjectiveness: uncertainty about others' experiences. As much as uncertainty is an inevitable part of human life, our lives are fundamentally social

[1] Because my thesis is concerned with the psychology of uncertainty, it focuses on the subjective kind of uncertainty and assumes that the above and following distinctions are psychological rather than ontological.

(Adolphs, 2003; Baumeister & Leary, 1995; Heider & Simmel, 1944; Henrich & Muthukrishna, 2021; Heyes, 2018). The inherent opacity of others' minds amplifies the unknowns we encounter in solitary decisions – from perceptual features of the environment (is this a genuine smile, or not?), over the feasibility of choice options (should I signal my willingness to cooperate, or not?), to the odds of outcomes (will they reciprocate, or not?; FeldmanHall & Shenhav, 2019; Kappes et al., 2019). Does this amplified uncertainty in social settings affect human decision-making and behaviour and if so, how?

Varieties of uncertainty: The social dimension

Decisions bearing consequences for others but not (directly) for ourselves might be considered one extreme of a continuum, where the other extreme represents decisions bearing consequences only for ourselves but not for others (i.e., individual decisions). Decisions with consequences for others but not ourselves occur, for example, in surrogate medical choices (Batteux et al., 2017), financial trading (Batteux et al., 2020; Eriksen & Kvaløy, 2010), or politics (Camerer & Kunreuther, 1989; Patt & Weber, 2014). How we choose for others in such decisions might be informed by how we would choose for ourselves (Epley et al., 2004) – a conjecture that has inspired a growing body of research comparing the domains of individual and social decision-making under uncertainty.

While some work has found that personal preferences over uncertain outcomes are predictive of advice to others (Hadar & Fischer, 2008), this does not seem to translate into choices for others: Even when there is a positive correlation between personal preferences and predictions about others' preferences over uncertain prospects, choices for others have been observed to deviate from the predicted preferences of others (and personal preferences to

the degree by which the two are correlated; Batteux et al., 2020). Oftentimes, both predictions of others' choices and choices on others' behalf have been characterised as closer to risk-neutrality compared to when we choose for ourselves (Batteux et al., 2017; Ogawa et al., 2018; Sun et al., 2017; X. Zhang et al., 2017), or when others choose for themselves (Faro & Rottenstreich, 2006; Hsee & Weber, 1997). Other times, choices for others are – relative to risk-neutrality – overly cautious (e.g., Bolton & Ockenfels, 2010; Charness & Jackson, 2009; Eriksen & Kvaløy, 2010; Fernandez-Duque & Wifall, 2007; Jingyi Lu et al., 2018; Zaleska & Kogan, 1971) or exceedingly risky (e.g., Batteux et al., 2017; Beisswanger et al., 2003; Mengarelli et al., 2014; Pollmann et al., 2014; Polman, 2012; Rigoli et al., 2018).

Two recent meta-analyses (Atanasov, 2015; Polman & Wu, 2020), which revealingly observe opposite overall effects, suggest that moderating factors (e.g., reciprocity, decision-frame, or social distance) can partly explain the evidence's inconclusiveness. To reconcile the remaining inconsistencies, the authors affirm that it will be crucial to further characterise what makes social decisions different in kind, rather than merely different in degree, from individual decisions under uncertainty (Atanasov, 2015; Polman & Wu, 2020).

Interdependencies between different agents naturally distinguish social from individual decisions. They come in the form of opportunities for reciprocity, retaliation, cooperation, or corruption among others. Considering social interdependencies shifts the focus away from the two extremes (i.e., decisions with consequences only for self, as opposed to decisions with consequences only for others), towards decisions with inherently connected outcomes. Studies on how such interdependencies form social decisions writ large have produced a thriving body of literature, which has been further incited by the rise of game theoretical approaches to social preferences (e.g., Fehr & Camerer, 2007). These

game theoretical approaches oftentimes implicitly model how we respond to uncertainties arising from our limited ability to predict others' behaviour (Weber & Johnson, 2009). For instance, an ultimatum game (Güth et al., 1982) requires one player to evaluate the chances that another (anonymous) player will accept their allocation offer; if the second player does not accept it, both end up empty-handed. Such situations in which we estimate the intentions and behaviour of others to optimise our own choices are ubiquitous (Camerer, 2011) and their study is key to understanding how uncertainty shapes social preferences (FeldmanHall & Shenhav, 2019). However, in this thesis I take a step back by focusing on non-strategic social decisions that are not impeded by uncertainty arising from dependencies on others' decisions and behaviour.

Taking this step back helps realizing that whenever our decisions bear outcomes for others, we inevitably face uncertainty about the impact of those outcomes on others' well-being (Hammond, 1990; Harsanyi, 1955; Nagel, 1974). Imagine, for instance, you wake up feeling sick yet urging going into office to finish a project important for your career. If you decide going to work despite feeling sick, might you infect a coworker, or not? How badly might that coworker get sick – might they be someone young and healthy, or someone vulnerable to severe disease? Our recent work suggests that these two questions reflect dissociable variants of uncertainty: uncertainty about what outcomes your decision might entail for others (i.e., *outcome uncertainty*) and uncertainty about the impacts of possible outcomes on others' well-being (i.e., *impact uncertainty*; Kappes et al., 2018). People seem to intuitively distinguish between outcome uncertainty, which indexes uncertainty about possible states of the world, and impact uncertainty, which indexes uncertainty about the subjective utilities over those states (see Kappes et al., 2018, Supplementary Methods).

Importantly, this distinction has behavioural consequences with people tending to behave more selfish under outcome uncertainty (Budescu et al., 1990; Dana et al., 2006; Haisley & Weber, 2010) but less selfish under impact uncertainty (Kappes et al., 2018).

Notably, the *uncertainty* in impact uncertainty seems crucial in driving its beneficial effect on prosociality: merely mentioning the worst possible outcome (e.g., the possibility of coworkers being vulnerable) does not produce increases in prosocial behaviour (Study 4 in Kappes et al., 2018), whereas enhancing impact uncertainty's salience (e.g., by pointing out that one does not know whether they might infect a healthy coworker who would suffer little from the disease, or a vulnerable coworker who would suffer severely from the disease) has proven critical in increasing prosocial behaviour, potentially by nudging people to overcome a reluctance to consider harmful impacts for others (Study 7 in Kappes et al., 2018). It is remarkable that people seem to exploit outcome but not impact uncertainty for adhering to self-serving interpretations of the unknowns they face: while they seem to optimistically underestimate the odds of bad outcomes for others materializing from a self-serving choice (Dana et al., 2006; Exley, 2016; Haisley & Weber, 2010; Shalvi et al., 2015), they seem to attach greater weight to worst-case-scenarios about their choices' impact of others' well-being when uncertain (Crockett et al., 2014; Kameda et al., 2016; Kappes et al., 2018).

These findings suggest that impact uncertainty, which is unique to the social domain (whereas outcome uncertainty pervades individual and social decisions alike), might elicit *precautionary* responses whereby people endorse whatever minimises harm to others. Ultimately, such precautionary responses might have adaptive value for ensuring compliance with the do-no-harm-principle (Fehr & Fischbacher, 2004; Henrich et al., 2010). Supporting this

conjecture, Crockett and colleagues (2014) have described an experimental setting in which people value harm to anonymous others more than harm to themselves. This "hyperaltruistic" valuation of harm for others relative to harm for self could not be explained merely in terms of empathic concern for the welfare of others (as suggested, for instance, in Cushman et al., 2012; Pizarro, 2000), nor in terms of strategic motives such as reciprocity or reputational concerns. Impact uncertainty did, however, emerge as a proximate explanation for the observed hyperaltruistic preferences from both, participants' explicit reasoning about their behaviour and from behavioural indices of decisional uncertainty (Crockett et al., 2014; see Chapter 3). As the authors speculate, people who respond more cautiously when in doubt about their decision's impact on others might have selective value not only in terms of safeguarding others' welfare but also in signalling moral responsibility through actions that may remain observable even when outcomes are obscured or delayed (ibid.; Jordan et al., 2016).

Additional evidence on uncertainty's role in social decisions stems from our recent research on individual differences, where dispositional tolerance to risk and ambiguity is predictive of prosocial behaviours such as helping others (Gross et al., 2021), or trusting others and cooperating with them (Vives & FeldmanHall, 2018). Research on social norms furthermore suggests that individuals who make altruistic decisions amidst uncertainty are judged as more altruistic and trustworthy by others (Capraro & Kuilder, 2016; Jordan et al., 2016; Kappes et al., 2018; Pérez-Escudero et al., 2016; Xiong et al., 2020). Together, these findings indicate that people might respond to uncertainty about their decision's impact on others' welfare by systematically erring on the side of caution and that such a precautionary response has adaptive value in fostering

cooperation and trust among fellow humans. A related question that arises is how uncertainty about artificial agents, with whom we are increasingly interacting in traditional "social" settings such as gaming, education, or caregiving (Henschel et al., 2021), might shape our readiness to trust them and cooperate with them.

Uncertainty in a different variety of "social": Human-machine interactions

Our daily interactions increasingly involve intelligent artificial instead of human agents. Many among us start their days by receiving their feed of morning news from a personalised algorithm; asking their virtual assistant – which often goes by the name of Alexa – to turn up the heating and to switch on the coffee machine; admiring Google map's nifty ways of evading the morning's rush-hour traffic; and finally booting up their artificial co-worker, also known as a computer. What seems nothing out of the ordinary today, was unthinkable just a decade ago. Autonomous vehicles, superhuman chess-masters, or drone warfare might still appear extraordinary, but they are already reality in our present times, where machines increasingly mediate our interactions across private and public spheres. This begs the question how we might respond to uncertainty arising from the unknowns about machines' inner workings, paralleling the question how we respond to uncertainty arising from the unknowns about the minds of other human beings (Dequech, 2004).

Machines powered by artificial intelligence are increasingly tuned to detecting social signals and/or to behave in ways that humans commonly perceive as social signals (Rabinowitz et al., 2018), loosely defining the emergent class of *social machines* (e.g., Cross & Ramsey, 2021; Henschel et al., 2021). For

example, recent work has equipped a relatively simple learning algorithm with the ability to generate and respond to cheap-talk (i.e., costly, non-binding signals such as “We can both do better than this”) and shown this artificial agent to generate cooperative relationships with humans (and other algorithms) in repeated interactions at levels that rival cooperation in human-human interactions (Crandall et al., 2018). In fact, many human participants were unable to distinguish whether they were interacting with a human or a machine, and subjective ratings of their interaction partner’s intelligence, intentionality, and usefulness of signals did not differ between human and machine partners (ibid.).

Research characterising the mental processes underpinning our interactions with social machines more generally still is in its early infancy (Cross & Ramsey, 2021; Henschel et al., 2020). Early work in this area suggests that key phenomena observed in social cognition about human agents extend to cognition about social machines (Cross & Ramsey, 2021; Wykowska, 2020; Wykowska et al., 2016), including the automatic imitation of movements (Gazzola et al., 2007; Oztop et al., 2005; Press et al., 2005, 2007; Wykowska et al., 2014), joint attention (Chaminade & Okka, 2013; Wiese et al., 2012), or the perceptual processing and recognition of emotional expressions (Chaminade et al., 2010). Meanwhile, higher-order cognitive processes such as the ascription of mental states including beliefs, desires, and intentions (Frith & Frith, 2006) seem to be qualified in interactions with artificial social agents (Chaminade et al., 2012; Krach et al., 2008). In particular, while people may ascribe agency and intentionality to machines (Crandall et al., 2018; Gamez et al., 2020; Shank & DeSanti, 2018; Waytz et al., 2014; Weisman et al., 2017), they tend to perceive machines as devoid of experience (Bigman & Gray, 2018; Gray & Wegner, 2012)

and do not show empathic responses to machines' display of experiences such as pain (Cross et al., 2019).

The perceived lack of sentience in machines conceivably suspends any uncertainty pertaining to others' (i.e., machines') subjective experience, such as impact uncertainty. Meanwhile, uncertainty about machines' inner workings becomes a pertinent variety. For instance, equipping artificial agents with the ability to engage in cheap-talk during repeated interactions caters human counterparts' urging to understand the underlying algorithm's decision-making strategies (Crandall et al., 2018). But against humans' fundamental need to explain and understand other agents' behaviour (Keil, 2006; Lombrozo, 2006, 2011; Miller, 2019), many of the most potent machines available to date are powered by black-box algorithms that are inscrutable not only to end-users, such as citizens and policy-makers, but even to their designers (Adadi & Berrada, 2018; Gunning et al., 2019; Murdoch et al., 2019; Waldrop, 2019). Reducing uncertainty about machines' inner workings thus has become a grand challenge in the engineering and machine learning sciences (Doshi-Velez & Kim, 2017; Gunning et al., 2019; Murdoch et al., 2019).

Concern about machines' opacity is ubiquitous beyond academics and developers, as illustrated by the European Union's attempt to enshrine a right to obtain meaningful explanations about automated decisions in its General Data Protection Act (European Parliament & European Council, 2016), medical doctors rejecting diagnostic tools outperforming humans in terms of accuracy amidst lack of insight about their inner workings (Finlayson et al., 2019; Price, 2018), or business executives asking for more explainability in algorithmic decisions (Kahn, 2018). Social scientists recently have affirmed "the imperative of interpretable machines" (Stoyanovich et al., 2020) and proposed an inevitable

link between explaining machines' inner workings and enabling trust in machines (also see Logg et al., 2019). Much of this recent attention has centered around the technical feasibility of interpretable machines (Lipton, 2017; Mittelstadt et al., 2019; Rudin, 2019). Meanwhile, the fundamental question whether and when the general public would value interpretable machines remains neglected. As artificially intelligent machines become increasingly ubiquitous companions in settings traditionally considered "social" (e.g., chatting with others, advising others, or caring for others), it seems worthwhile to consider how uncertainty about their inner workings shapes our interactions with them.

Contribution to the literature

My thesis explores nuances of uncertainty in different varieties of social decision-making. As such, my doctoral work endeavours to complement existing research in two ways: First, previous work has considered uncertainty oftentimes as a unitary concept (Ülkümen et al., 2016), or at least no less than the presence or absence of knowledge about the probability distributions of outcomes (Huettel et al., 2006; Platt & Huettel, 2008). In contrast, few real-life situations feature defined sets of outcomes, leave alone probabilities over those outcomes. By considering different varieties of uncertainty and their interactions – from risky outcomes to the unknowns of others' experience and decision-making – I seek to broaden our understanding about the manifold manifestations and paths in which uncertainty shapes human decision-making.

Second, I hope to extend existing research by considering uncertainty in decisions that pertain to other agents. Surprisingly, to date decision-making under uncertainty has mostly been explored in the context of solitary choices

(Polman & Wu, 2020). But given that our human nature is deeply social (Henrich & Muthukrishna, 2021) and has evolved in an intricate web of interdependencies (Heyes, 2018), even seemingly solitary decisions bear outcomes for others. Think of someone choosing to live the isolated life of a self-sufficient hermit; even they will, by this very choice, have an impact on others – from the close ones who might have preferred to keep them in their midst, to the distant unknown ones who might appreciate the incremental (or infinitesimal) redemption of resources. Ultimately, we can never be sure whether and how the choices we make – even the ones that seem to concern only ourselves at a first glance – affect others. This is, first, because we often lack knowledge about the set of possible outcomes and whether it includes outcomes for others; and, second, because we can never truly know what it is like to be in another's shoes (Frith & Frith, 2006; Harsanyi, 1955; Nagel, 1974). To understand social decisions at large, it seems paramount to understand how uncertainty about others' internal states factors into people's decisions and I hope that my thesis can contribute to this challenge.

Overview of empirical chapters

My thesis has two main goals: first, I want to contribute to a more nuanced empirical understanding of uncertainty and how it shapes social decisions in particular (Chapters 2 and 3). Second, I seek to further expand this understanding by considering uncertainty in the domain of social decisions in a broader sense, including interactions between humans and machines (Chapter 4). In the following I give a brief overview on how my empirical chapters seek to address these goals.

Social versus individual decision-making under risk (Chapter 2)

In Chapter 2, I explore when social decisions (i.e., decisions with outcomes for others) under uncertainty diverge from individual decisions (i.e., decisions with outcomes for self) under uncertainty. I propose that social decisions are different in kind because they inherently come with uncertainty about others' experience of potential outcomes. This enhanced uncertainty, in turn, may produce precautionary preferences whereby decisions for others are less risky than decisions for self. To test this hypothesis, I compared social and individual decisions under risk. I focused on decisions under risk, where outcome probabilities are known to decision-makers, because this creates a natural link to past research on (individual) decision-making under uncertainty, which has most commonly modelled uncertainty in terms of risk (e.g., Kahneman & Tversky, 1979).

Across three studies, I parametrically varied outcome probability and magnitude in choices about (hypothetical) physical outcomes (Study 2.1) and real monetary outcomes (Studies 2.2 and 2.3) for others versus self. Indicating a precautionary pattern, choices for others were markedly more risk-averse than choices for self when the odds of a desirable outcome were low, while they converged for higher probabilities (Studies 2.1 - 2.3). In line with the hypothesis that enhanced uncertainty in social choices contributes to this pattern, I observed greater indecision (Study 2.1) and less confidence (Study 2.3) in social relative to individual decisions, particularly for low probabilities and high magnitudes of a desirable outcome. Meanwhile, motivational differences in risky choices for others versus self could not explain the observed differences (Study 2.3).

Uncertainty's role in harm aversion for others versus self (Chapter 3)

In Chapter 3, I consider uncertainty in social decisions with deterministic outcomes. Looking at deterministic outcomes allows me to focus on uncertainty about the impact of our choices on others' welfare, i.e., impact uncertainty. In particular, I explore whether impact uncertainty can explain a previously observed precautionary pattern whereby people are more averse to harming anonymous others than themselves. To this end, I manipulated impact uncertainty in the context of the harm aversion paradigm, where hyperaltruistic valuation of harm to others has been observed previously (Study 3).

In brief, I invited pairs of participants to a laboratory study where they made choices between an option yielding a higher amount of money for a higher number of mildly painful electric shocks, or a lower amount of money for a lower number of electric shocks. While a given participant would always receive the monetary payoffs according to their choices, the electric shocks would be borne by themselves on some rounds and by the other participant on other rounds, allowing me to compute behavioural and computational indices of participants' valuation of harm for self and for other. While some participants were instructed that people vary greatly in how they experience the electric shocks and that they could hence not assume their counterpart to experience them similarly to themselves (high impact uncertainty condition), another group was instructed that people vary little and they could hence experience the other's shock experience to be similar as their own (low impact uncertainty condition), and yet another group received no instructions relating to uncertainty (control condition). Larger self-other differences in reaction times and choice noisiness under high impact uncertainty pointed towards a

modulating effect of impact uncertainty in choices for others, but not to a degree that modulated hyperaltruism.

Uncertainty in human-machine interactions (Chapter 4)

In Chapter 4, I seek to explore uncertainty in social interactions in a broader sense, specifically in human-machine interactions. As empirical research into the psychology of human-machine interactions is only just emerging (Rahwan et al., 2019), I start by characterising public demand for interpretability (in other words, lack of uncertainty) regarding the inner workings of machines. Across five studies (4.1A/B – 4.4), I observed a robust and positive demand for interpretable machines among non-experts that generalizes across a variety of real-world applications and follows predictable patterns. In particular, demand for interpretability was strongest for machines that make decisions compared to recommendations (Study 4.1A) and for decisions involving high stakes and scarce resources (Studies 4.1A/B – 4.2). At the same time, participants overwhelmingly prioritized machine accuracy over machine interpretability when the two traded off against one another (Studies 4.3-4.4), regardless of whether they took the perspective of a responsible agent overseeing the machine's decision or a patient directly affected by the machine's decision (Study 4.4). Crucially, participants' preferences for accuracy over interpretability were amplified under the very same conditions that drive preferences for interpretability in the first place (Studies 4.3-4.4): for AI decisions involving high stakes and scarce resources, accuracy dominated interpretability. Overall, my results highlight that while non-experts seem to generally value interpretability in AI, they may prioritize AI performance at the expense of understanding especially in those domains where understanding is most important.

Nomenclature

In the present thesis, I refer to choices in which the decision-maker bears outcomes exclusively (or mostly) themselves (sometimes also referred to as "nonsocial" decisions in the literature, e.g., Ruff & Fehr, 2014) as *individual decisions*. Choices with outcomes for others are referred to as *social decisions*, including choices made on behalf of others as well as choices with outcomes for both, the decision-maker and others.

I will use the term *uncertainty* as referring to the precision with which a prediction can be generated based on the available information (Hirsh et al., 2002). As such, uncertainty may refer situations involving risk, where probability distributions are known, or to situations involving ambiguity, where probability distributions are unknown (Ellsberg, 1961; F. H. Knight, 1921). Whenever I consider decisions over outcomes with known probability distributions, as in Chapter 2, I will specifically refer to decisions under risk.

The terms *artificial agents*, *Artificial Intelligence (AI) agents/tools/systems*, and *machines* will liberally to refer to complex and simple algorithms used to make decisions (Rahwan et al., 2019). I will use the terms *AI* and *machines* interchangeably, with the latter referring to machines powered by AI. I will speak of *social machines* when referring to artificial agents who are specifically tuned to detecting social signals and/or to behave in ways that humans commonly perceive as social signals (Cross & Ramsey, 2021; Henschel et al., 2021).

2 Social versus individual decision-making under risk

Social decisions under risk are diverse and ubiquitous. Examples range from parents contemplating the uncertainties of choosing "the best" school for their children (Ziegelmeyer & Ziegelmeyer, 2016), over surrogate medical choices on risky treatments for a loved one (Moorman et al., 2009), to policy decisions about low-probability-high-impact risks such as AI that can affect entire populations or even humanity at large (Camerer & Kunreuther, 1989; Fischhoff, 1995). And yet, we know comparatively little about *how* people make social decisions under risk (e.g., Polman & Wu, 2020). Previous work has often pitched how people respond to risk in decisions about others' outcomes against how people respond to risk in decisions about their own outcomes. Using individual choices under risk as a benchmark is useful from a scientific perspective, because it has been extensively studied and is thus relatively well characterised (Platt & Huettel, 2008; Tversky & Kahneman, 1974). Conceivably, it is also appealing from a decision-maker's perspective: what we would choose for ourselves under risk, might well inform what we choose for others under risk (Epley et al., 2004).

While the relationship between individual and social decisions under risk has recently received more attention, a puzzling picture has emerged with some studies finding that we make less risky choices on behalf of others (e.g., Bolton & Ockenfels, 2010; Charness & Jackson, 2009; Eriksen & Kvaløy, 2010; Fernandez-Duque & Wifall, 2007; Jingyi Lu et al., 2018; Zaleska & Kogan, 1971), and others observing the exact opposite (e.g., Batteux et al., 2017; Beisswanger et al., 2003; Mengarelli et al., 2014; Pollmann et al., 2014; Polman, 2012; Rigoli et al., 2018).

These inconsistencies are reflected in two recent meta-analyses, where Atanasov (2015) observed more risk-averse choices for others than self whereas Polman and Wu (2020) observed less risk-averse choices for others than self. Both meta-analyses concur in observing moderating factors, particularly decision reciprocity and decision frame (also see Mengarelli et al., 2014; Vlaev et al., 2017; X. Zhang et al., 2019). Even if such moderating factors can explain some inconsistencies between contributions, they cannot account for inconsistencies persisting in studies where moderating factors were held constant or across studies using the same experimental task (e.g., Jung et al., 2013; Jingyi Lu et al., 2018; Sun et al., 2017; X. Zhang et al., 2017). Reflecting and reinforcing the fragmentation of empirical results, there are only few attempts to apply and extend existing theoretical accounts of individual decisions under risk – such as prospect theory (Tversky & Kahneman, 1992) – to social decisions under risk (for an exception see X. Zhang et al., 2019). Consequently, there is no robust theoretical account of risky choices for others (Atanasov, 2015; Polman & Wu, 2020).

I propose that addressing this gap requires considering a qualitative difference between social and individual decision-making under risk. In short, choosing for others inherently involves enhanced uncertainty (FeldmanHall & Shenhav, 2019; Kappes et al., 2018; Vlaev et al., 2017) because we cannot possibly be certain about others' experience of potential outcomes and thus their (risk) preferences (Harsanyi, 1955; Nagel, 1974). How, then, do people bridge this knowledge-gap they face in social decisions? Tunney & Ziegler (2015) propose three possible avenues: people might project what they would choose in the other person's position, or they might simulate what the other person would prefer, or they might endorse a benevolent perspective from which they would

choose what they consider best for the other person – independent from their own projected or the other's simulated preferences. The very fact that previous research has repeatedly observed differences between risky choices for others versus oneself (e.g., Polman & Wu, 2020), even if effects go in different directions, challenges the idea that we make such decisions by merely projecting our personal preferences. Existing work also has documented that choices for others often do not match beliefs about those others' preferences (Garcia-Retamero & Galesic, 2012; Pahlke et al., 2015; Stone et al., 2013; Stone & Allgaier, 2008).

Recent work by Batteux, Ferguson, and Tunney (2020) directly tested the relationship between people's own preferences, their simulation of others' preferences, and choices on behalf of those others in risky decisions. They observed that, even though simulations of others' preferences did inform choices for those others to some extent, simulations and choices for others overall diverged, as did choices for oneself and choices for others. Notably, in most cases, these divergences were driven by choices for others being more risk-averse than predicted others' preferences or participants' own choices, in particular for high-risk scenarios. These findings align with people taking a benevolent perspective when making social decisions and, more precisely, a precautionary perspective whereby people err on the side of caution in choices for others when the odds of an undesirable outcome are high.

Adopting precautionary preferences when choosing for others is arguably a best response from the decision-maker's perspective, because it minimises possible harm to others (Chakroff et al., 2013; Crockett et al., 2014; Cushman et al., 2006, 2012) and thus the responsibility one might come to bear for others' outcomes (Bolton et al., 2015; Bolton & Ockenfels, 2010; Charness & Jackson, 2009; Garcia-Retamero & Galesic, 2012). Reflecting its value as best response from the

perspective of affected others (or even a societal perspective), the “do no harm” principle is enshrined in a wide-spread social norm (Henrich et al., 2010) that is enforced at positive costs (Fehr & Fischbacher, 2004). It is also reflected at the level of policymaking, where the precautionary principle prohibits actions that carry any risk of causing harm to others (Arrow & Fisher, 1974). In line with these considerations, work using computational models has interpreted increased randomness in social relative to individual decisions as enhanced uncertainty in social decisions (Rigoli et al., 2018) and proposed that the degree of randomness explains the degree to which decisions for others are less risky than individual ones (Vlaev et al., 2017). Furthermore, when others’ welfare is affected, people tend to demand more certainty about what constitutes the best choice relative to a situation where their own welfare is at stake (Edelson et al., 2018). This increased demand for certainty at the cognitive level translates into responsibility-aversion at the behavioural level: when people can defer choosing between a safe and a risky option, they are more likely to defer when deciding on behalf of others rather than themselves (Edelson et al., 2018; Steffel et al., 2016).

In cases when there is no way to reduce uncertainty in decisions for others or to defer them, people seem to attach greater weight to the worst possible outcome when choosing for others compared to choosing for themselves (Kameda et al., 2016; Kappes et al., 2018). For example, people readily exploit *outcome uncertainty* about what outcomes their decisions might produce for others (e.g., whether or not they would infect someone going to work when sick). In contrast, *impact uncertainty* about how possible outcomes might affect others’ well-being (e.g., whether a co-worker might suffer from mild or severe symptoms) produces precautionary preferences whereby people choose whatever option minimises potential harm for others, even at a cost to

themselves (Kappes et al., 2018). In a similar vein, work showing that people are more averse to harming others than themselves (e.g., Crockett et al., 2014; L. J. Volz et al., 2017) has proposed that enhanced uncertainty in social decisions could drive this pattern, leading people to add a risk premium to potentially harmful choices for others (Crockett et al., 2014).

Based on these considerations, I hypothesise that people respond to enhanced uncertainty in risky social decisions by adopting precautionary preferences. More specifically, in keeping with observations that people require more certainty in risky social choices (Edelson et al., 2018; Steffel et al., 2016), I predict that people will be more risk-averse (i.e., precautionary) in social relative to individual decisions. Notably, past work would predict the opposite for individual decisions under risk, i.e., enhanced risk-seeking amidst low probabilities for a desirable outcome (Tversky & Kahneman, 1974, 1992). I expect magnitude to play a more uniform role across social and individual decisions given that precautionary preferences have been observed in social decisions involving various types and levels of outcome magnitude and in light of recent evidence suggesting that value functions do not differ across risky choices for self versus others (Sun et al., 2020; Vieider et al., 2016).

In the following, I will present a series of studies which aimed to test the proposed precautionary pattern by parametrically varying outcome probability and magnitude while minimising potential moderating factors identified in previous work (Atanasov, 2015; Polman & Wu, 2020). In particular, across all three studies, I used a uniform decision frame, the same “baseline” social distance (i.e., strangers), and excluded reciprocity concerns. Since my predictions derive from considering social decisions as different in kind from individual decisions, I expected them to hold independently of such boundary conditions. Diverging

from the "risk as value" hypothesis (Beisswanger et al., 2003; Stone et al., 2013; Stone & Allgaier, 2008), which posits that decisions for others are more risk-seeking than decisions for self when risk is socially valued (e.g., in romantic partner search), I expected to observe differences in risk-taking within the same domain of social valuation and in fact in a domain where the social value of risk arguably is ambiguous (i.e., monetary choices). Similarly, I expected differences in risk-taking between social and individual decisions to occur within the same decision frame, contrary to "risk as feeling" accounts (Loewenstein et al., 2001; Fernandez-Duque & Wifall, 2007; Sun et al., 2020; X. Zhang et al., 2019) that would predict more (less) risk-taking in social decisions for gain (loss) frames. Again, this is because my proposal's crucial motivation is to consider social decisions as qualitatively different from individual decisions under risk, rather than different in degree of affective responses or other aspects. Finally, by focusing on anonymous one-shot decisions in non-reciprocal settings, I aimed to minimise responsibility concerns as potential driving forces (Bolton et al., 2015; Bolton & Ockenfels, 2010; Charness & Jackson, 2009).

All three studies followed relevant ethical guidelines and were approved by the University of Oxford's Central University Research Ethics Committee (Study 2.1: MS-IDREC-C1-2015-098; Studies 2.2 and 2.3: MS-IDREC-R56657/RE001). Participants were recruited online via Amazon's Mechanical Turk (MTurk; Studies 2.1 and 2.3) and Prolific Academic (Prolific; Study 2.2), which provide ethnically and socioeconomically more diverse participant pools than university students (e.g., Buhrmester et al., 2011; Horton et al., 2011; Paolacci & Chandler, 2014), and which allow – in combination with the use of response-validity indicators – to efficiently collect significant amounts of data comparable in quality to laboratory studies (Chmielewski & Kucker, 2020; Kees et al., 2017). Participants in each study

gave their informed consent beforehand and were compensated at an hourly rate of $7.25 in line with the US minimum wage (with choice-dependent bonus payments paid on top in Studies 2.2 and 2.3). Pre-registrations for Studies 2.1-2.3 with the Open Science Framework are accessible here. Together with my supervisors and collaborators Jim A. C. Everett, Julian Savulescu, Molly J. Crockett, and Nadira S. Faber I am preparing a manuscript for publication based on this work.

Study 2.1: Self-other differences in hypothetical choices under risk

To investigate whether people respond to enhanced uncertainty in risky social decisions by adopting precautionary preferences, my first study used a hypothetical harm aversion paradigm (Crockett et al., 2014, 2015) that I could run online. In brief, the harm aversion paradigm asks participants to choose between an option yielding a lower amount of money for a lower number of mildly painful (but safe) electric shocks or an option yielding a higher amount of money for a higher number of shocks. Building on this basic paradigm, I had one group of participants making *individual* choices between a small but safe or a higher but risky reduction of (hypothetical) shocks for themselves; the other group made *social* choices between the same options, but a different (hypothetical) person received the (hypothetical) shocks. Given that the task involved physical harm, even if hypothetical, I also asked participants to indicate how wrong or right it would be to choose the risky option.

Design

I used a mixed design with two within-subjects factors of five levels (*probability*: 5%, 25%, 50%, 75%, 95% probability of shock reduction and *magnitude*: 2, 5, 10, 15, 20 shocks) and one between-subjects factor (*domain*:

individual versus social). Participants were randomly allocated to either the *individual* or *social* condition, and the order in which participants were presented with the 25 decision-rounds of varying probability and magnitude was randomised.

Participants

I used the *smpsize_llm* function from the *sjstats* package in *R* for mixed-effects model sample-size calculation (Hsieh et al., 2003). A required sample of *n* = 195 was estimated to detect a medium effect size (Cohen's d = .5, estimate based on Shenhav & Greene, 2010) at power .85 and over 25 rounds. I recruited a total of *N* = 420 (*n* = 210 for the *individual* and *social* choices conditions, respectively) participants from the US on MTurk. After excluding duplicates and incomplete responses, I obtained a final sample of *N* = 417 that was 57% male with an average age of 36.46 (*SD* = 11.27).

Procedure

After providing informed consent, participants were instructed on the main task depicted in *Figure 1*. In the *individual* condition, subjects imagined participating in an experiment together with another person. They were told that the other participant had randomly been assigned the 'Decider' role and that they were the 'Receiver'. The Decider would make a series of decisions about monetary payoffs coming at the cost of mildly painful electric shocks. As the Receiver, they would get both the money and the shocks, but they would have a chance to reduce the number of shocks resulting from the Decider's choices. Similarly, subjects in the *social* condition imagined participating in an experiment together with two other people, who had randomly been assigned the roles of the Decider and Receiver, and that they got to be the 'Helper'. Again,

they learned that the Decider would make a series of decisions involving money at the cost of electric shocks given to the Receiver. As the Helper, they would have a chance to reduce the number of shocks the Receiver would get due to the Decider's choices.

Across both conditions, participants learned that in each of 25 decision-

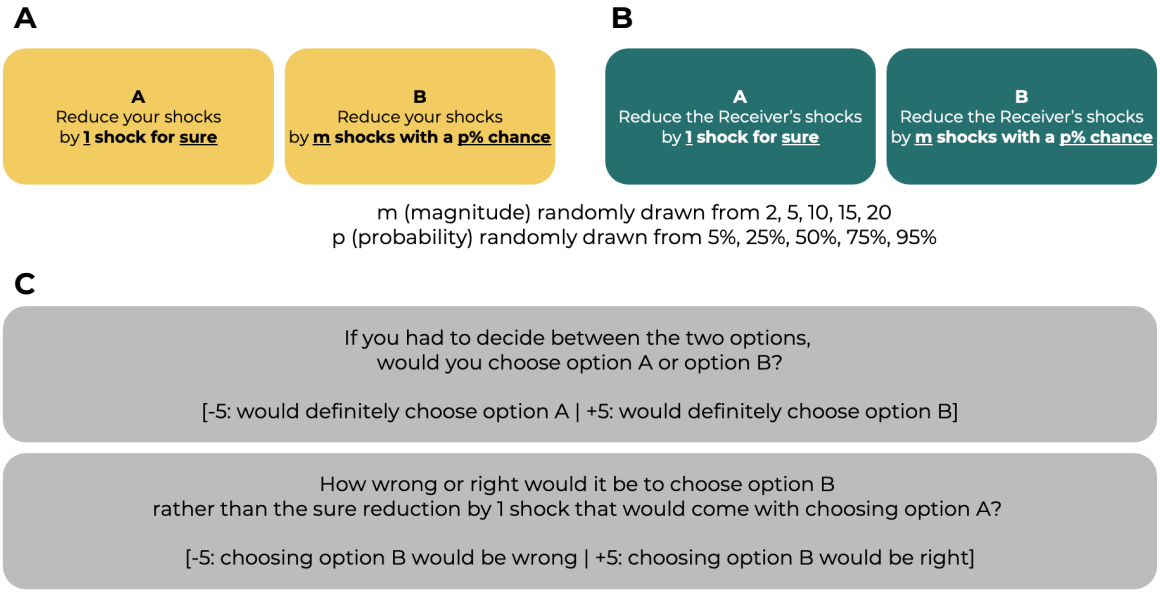


Figure 1. Schematic representation of the main task in Study 2.1. (**A**) Participants making individual decisions chose between reducing their own shocks by one shock for sure or by m shocks at a chance p. As depicted in (**B**), participants making social decisions faced the equivalent chance but decided over outcomes for the Receiver who was a randomly paired other participant. (**C**) Participants indicated behavioural intentions and judgments, each on 11-point rating scales. The order of the two dependent variables was counter-balanced across participants.

rounds, the Decider made a choice whereby the Receiver would get some money at the cost of 20 electric shocks. They were also told that the shocks were quite unpleasant and that the Receiver (i.e., themselves in the *individual* condition) would be grateful for receiving even one less shock on any decision-round. For each round, participants – either acting as Receiver or Helper – were given a choice between two options: a sure reduction by one electric shock (i.e., the safe option), or a probabilistic reduction by multiple electric shocks at a certain chance (i.e., the risky option). The risky option's number of shocks

(*magnitude*) and the chance that they would indeed be reduced (*probability*) varied for every decision-round (see *Figure 1*). After reading through two examples and conditional on successfully completing a simple comprehension check, participants proceeded to make choices on the 25 decision-rounds. They answered two questions for each round: how they would choose if they had to decide (intention) and how wrong or right would it be to choose the risky rather than the safe option (judgment). The order of these two questions was counterbalanced across participants but consistent for each individual. Intentions were reported on an 11-point rating scale ranging from "-5 – would definitely choose the safe option" to "+5 – would definitely choose the risky option", and the judgment-question was answered on an 11-point rating scale ranging from "-5 – choosing the risky option would be wrong" to "+5 – choosing the risky option would be right".

After completing all rounds, participants indicated how much their decisions were based on five different reasons: (i) on the number of shocks associated with the risky option (magnitude), (ii) the chance of shock-reduction associated with the risky option (probability), (iii) the balance between magnitude and probability, (iv) a belief that it is wrong to sacrifice a safe for a risky option, and (v) a belief that it is right trying to minimise the number of shocks by choosing the risky option, even at the sacrifice of a safe (but small) reduction. They then answered standard measures of baseline risk preferences (Falk et al., 2016) and empathy (Questionnaire of Cognitive and Affective Empathy (QCAE), Reniers, Corcoran, Drake, Shryane, & Völlm, 2011) as well as standard demographic questions.

Results

First, as a sanity check, I confirmed whether participants were sensitive to variation in the two parameters, probability and magnitude. This was the case for both individual and social decisions in terms of both intentions and judgments (*Figure 2A*), with endorsements transitioning from the safe to the risky option for higher levels of probability and magnitude. Neither dependent variable was normally distributed, with data points accumulating towards the

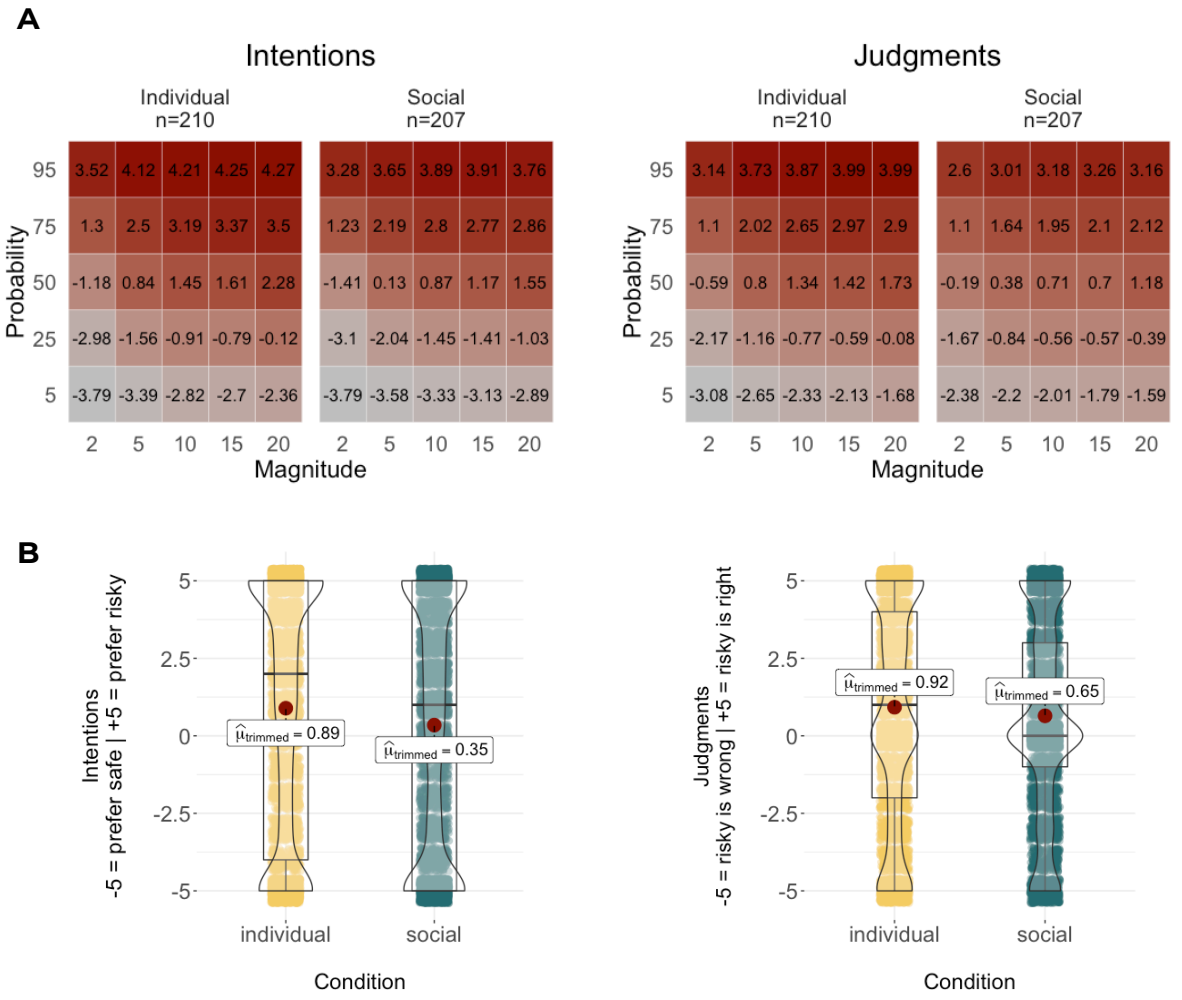


Figure 2. (**A**) Heatmaps for intentions and judgments across rounds and conditions. Each tile represents the parameters for the risky option on a given trial as combination of magnitude (number of shocks; x-axis) and probability (chance that shocks will indeed be reduced; y-axis). Numbers in the tiles indicate the mean response value for a given trial. Grey shades indicate stronger endorsement of the safe option, red shades indicate stronger endorsement of the risky option. (**B**) Distributions of responses for intentions and judgments for individual and social choices aggregated across all rounds. Trimmed means indicated by annotated red dots.

scale ends and midpoints (*Figure 2B*). Notably, participants making social decisions were significantly more likely to choose the scale midpoint compared to participants making individual decisions as indicated by Fisher's exact tests, *intentions*: p = .004, OR = 1.29, 95% CI_{OR} [1.08, 1.55]; *judgments*: p < .001, OR = 1.63, 95% CI_{OR} [1.48, 1.79]. I thus used nonparametric tests and robust regression throughout the following analyses. To account for the nested structure of the data (i.e., 25 decisions rounds per participant), I used robust mixed effects models with random intercepts for analyses on the two main dependent variables (i.e., intentions and judgments) implemented with the *robustlmm* package (Koller, 2016) and the *WRS2* package (Mair & Wilcox, 2020) in *R* (R Core Team, 2020).

Intentions and judgments. Intentions and judgments were substantially correlated in both individual decisions (r_τ = .75, p < .001, 95% CI [0.74, 0.76]) and in social decisions, r_τ = .67, p < .001, 95% CI [0.66, 0. 68]. Notably, this correlation was significantly smaller for social relative to individual decisions, z = 2.91, p = .004, 95% CI [0.02, 0.10]. In line with this observation, examining intentions and judgments pairwise across decision-rounds using Wilcoxon Signed-Rank tests indicated that they diverged significantly for social decisions (Z = 5.99, p < .001, 95% CI [0.01, 0.50]) but not for individual ones, Z = 0.12, p = .904, 95% CI [-0.01, 0.01].

To follow up on those differences in social decisions, I ran a robust random intercept model that predicted a difference-score between judgments and intentions by fixed effects for 'probability', 'magnitude', and their interactions as well as a random intercept effect for 'subject'. Significant main effects for probability (b = -0.64, SE = 0.02, 95% CI [-0.68, -0.59], d = 0.39, 95% CI_d [0.36, 0.43]) and magnitude (b = -0.13, SE = 0.02, 95% CI [-0.18, -0.09], d = 0.08, 95% CI_d [0.05, 0.12]), which were not qualified by an interaction (95% CI [-0.02, 0.07]), indicated that – focusing on social decisions – the risky option was less endorsed in terms

of intentions than judgments for lower levels of probability and magnitude, while the opposite was true for higher levels of probability and magnitude. Put differently, for low levels of probability and magnitude, participants making social decisions seemed hesitant to submit the risky option as (intended) choice even when they thought it might be the right option to choose. Meanwhile, for participants making individual decisions, intentions and judgments were more aligned.

Intentions. Since the previous analyses suggested that intentions and judgments influenced each other and did so differently across the two conditions, the following analyses on intentions controlled for 'judgments' and their interaction with 'condition'. Using a robust random intercept model with 'condition', 'judgment', and their interaction entered as fixed effects and 'subject' as random intercept, I observed weaker endorsement of the risky option for social relative to individual decisions, b = -0.17, SE = 0.06, 95% CI [-0.29, -0.04], d = 0.03, 95% CI_d [-0.02, 0.07].

Next, I sought to explore whether and to what degree this overall difference in social versus individual decisions was reflected in differences for how probability and magnitude factor into intentions across the two domains. To this end, I added fixed effects for 'probability', 'magnitude', and their two-way interactions with another and with 'condition' to the random intercept model. As summarised in *Table 1*, main effects were qualified by two-way *probability x condition* and *probability x magnitude* interactions. In particular, participants indicated weaker endorsement of the risky option in social relative to individual decisions for lower probability levels (see *Figure 3*). Magnitude was more decisive on decision-rounds that came with a 50% probability, whereas the relationship between intentions and magnitude was weaker for decision-rounds with

probability values below and above the 50% level. Hence, the observed interactions indicated that magnitude played a lesser role for participants' decisions and for the observed overall difference of more risk-aversion in social compared to individual intentions. Meanwhile, the interaction between probability and condition indicated that an asymmetric effect of probability contributed to this difference with social decisions being more risk-averse than individual ones particularly for low levels of probability (i.e., 5%, 25%, 50%), while intentions across the two domains converged for higher levels of probability (i.e., > 50%).

Table 1. Robust random intercept model predicting intentions (Study 2.1)

Variables	*b* (SE)	95% CI	*d*	95% CI_d
(Intercept)	0.09 (0.05)			
Judgments	0.89 (0.01)	[0.87, 0.91]	1.01	[0.99, 1.03]
Condition	-0.14 (0.08)	[-0.29, 0.00]	0.02	[-0.03, 0.07]
Probability	0.53 (0.03)	[0.48, 0.58]	0.19	[0.16, 0.22]
Magnitude	0.16 (0.02)	[0.12, 0.19]	0.08	[0.05, 0.10]
Judgments x condition	-0.07 (0.01)	[-0.10, -0.05]	0.06	[0.04, 0.08]
Probability x condition	0.47 (0.04)	[0.40, 0.55]	0.13	[0.09, 0.16]
Magnitude x condition	0.04 (0.03)	[-0.01, 0.10]	0.02	[-0.02, 0.05]
Probability x magnitude	-0.05 (0.01)	[-0.07, -0.02]	0.03	[0.01, 0.06]

Note. Estimates from robust multilevel model. Confidence intervals for the estimates are reported because p-values are not available for robust multilevel models.

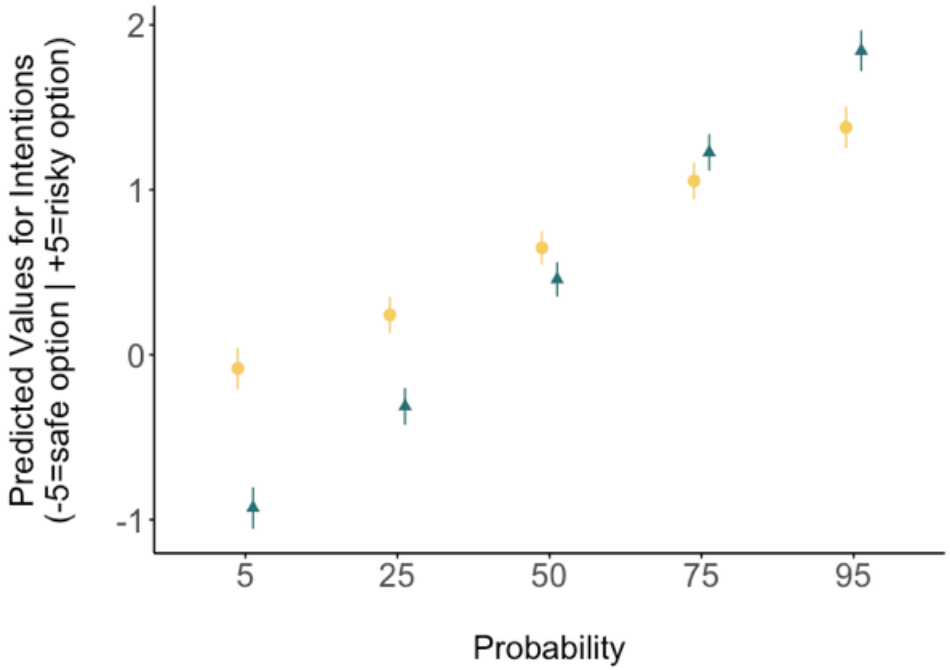


Figure 3. Predicted values of intention to choose the risky option for probability x condition interaction of random intercept model from Study 2.1 (see also *Table 1*).

These results did not change when I controlled for gender, age, baseline risk-preferences, and empathy (probability x condition: b = 0.47, SE = 0.04, 95% CI [0.40, 0.54], d = 0.13, 95% CI_d [0.09, 0.16]; probability x magnitude: b = -0.05, SE = 0.01, 95% CI [-0.07, -0.02], d = 0.03, 95% CI_d [0.01, 0.06]). While gender, age, and empathy did not affect participants' choices, participants with stronger baseline preferences for risk-seeking were more inclined towards the risky option, b = 0.03, SE = 0.01, 95% CI [0.02, 0.05], d = 0.05, 95% CI_d [0.03, 0.06].

Importantly, the observed asymmetric effect of probability on social compared to individual intentions depended on controlling for judgments and was not significant (95% CI [-0.09, 0.10]) in a model omitting judgments as control variable. In this case, participants making social decisions showed less endorsement of the risky option consistently at all levels of probability, b = -0.44, SE = 0.19, 95% CI [-0.80, -0.07], d = 0.02, 95% CI_d [-0.06, 0.11].

Judgments. Since intentions were the primary measure of interest, I will give a cursory summary on the results for judgments. A random intercept model with 'intentions', 'condition', 'probability', 'magnitude', and their two-way interaction entered as fixed effects and a random intercept effect for 'subject' indicated that there were no interactions between outcome probability or magnitude with condition. In line with the previously reported findings, there was an interaction between intentions and condition (b = -0.14, SE = 0.01, 95% CI [-0.16, -0.12], d = 0.13, 95% CI_d [0.11, 0.15]) indicating a weaker association between judgments and intentions for participants making social decisions. Otherwise, I observed significant main effects for intentions (b = 0.74, SE = 0.01, 95% CI [0.72, 0.75], d = 0.96, 95% CI_d [0.94, 0.97]), probability (b = 0.43, SE = 0.03, 95% CI [0.38, 0.38], d = 0.16, 95% CI_d [0.12, 0.19]), and magnitude, b = 0.08, SE = 0.02, 95% CI [0.04, 0.12], d = 0.04, 95% CI_d [0.01, 0.07]).

Reasons. When I asked participants to explicitly rate how much they based their decisions on a given trial's probability, magnitude, or their balance for the risky option on a scale from 1 "not at all" to 7 "very much", answers did not differ between those making individual versus social decisions according to Welch's t-test (ps > .11; note that I did not differentiate between intentions and judgments for these questions). However, according to Welch's t-test, participants who made social decisions assigned greater importance to the deontological "belief that it is wrong to sacrifice a sure reduction by 1 shock that would come with choosing [the safe option], when the reduction that might come with choosing [the risky option] is uncertain", $M_{individual}$ = 3.05, M_{social} = 3.81, t(248.61) = 3.06, p = .003, d = -0.40, 95% CI_d [-0.64, -0.23]. There was a similar but not significant trend for the consequentialist "belief that it is right to try and minimise the number of shocks by choosing [the risky option], even at the

sacrifice of a sure reduction by one shock that would come with [choosing the safe option]", $M_{individual}$ = 4.45, M_{social} = 4.88, t(232.16) = 1.88, p = .062.

Discussion

Study 2.1 supported the hypothesis that people experience enhanced uncertainty in social decisions under risk and respond to it by adopting precautionary preferences. Pointing towards enhanced decisional uncertainty in social decisions, I observed that judgments and intentions diverged more strongly for social compared to individual decisions. This was especially the case for low (and high) probabilities and magnitudes, where social decisions indicated a relatively stronger (respectively weaker) endorsement of the risky option in terms of judgments than intentions. As a further indication of enhanced uncertainty in social decisions, participants making social decisions chose the scale midpoint more often relative to those making individual decisions. In result, I observed that social decisions were consistently more risk-averse than individual ones in terms of intentions considered in isolation. However, when controlling for their relationship with judgments, I observed an asymmetric effect of probability on social intentions. In particular, when controlling for judgments, social decisions were markedly more risk-averse than individual ones for low to moderate probability levels (≤ 50%) but converged for higher probability levels (≥ 75%). In terms of judgments, individual and social decisions did not differ across variations of probability.

These findings align with the *uncertainty hypothesis* that participants choosing on behalf of others experienced greater decisional uncertainty than those choosing for themselves (Rigoli et al., 2018). In particular, decisional uncertainty might explain participants reverting to the neutral scale midpoint more often in social relative to individual choices. Furthermore, participants

making social decisions to a larger extent indicated basing their choices on deontological and consequential reasoning. Taking a more principled approach might provide a way for them to cope with decisional uncertainty and explain the stronger divergences between intentions and judgments. Meanwhile, the finding that both deontological *and* consequential beliefs received stronger endorsement among participants making social decisions relative to those making individual decisions, could be seen as further evidence for decisional conflict.

Alternatively, participants making social decisions might have been less motivated to commit to a directional choice and/or to maximise shock reduction compared to those making individual decisions (Rigoli et al., 2018). This *motivational hypothesis* could also explain why participants chose the scale midpoint more often in social compared to individual decisions and why the relationship between intentions and judgments was less systematic.

On that note, it seems important to remark that the interdependence between the two dependent variables, intentions and judgments, was imposed by design of the present study. While there is emerging evidence that judgments, which might be viewed as explicit expressions of moral norms, play a formative role in social preferences and as such deserve more attention (Capraro & Rand, 2018; Capraro & Perc, 2021), it also seems important to explore social versus individual *preferences* under risk in their own right. For the present study, the conclusions that can be drawn in this regard are not only limited by the interdependence between the two dependent variables (*viz.* judgments and intentions) but also by the format of the intention-measure, which consisted of an 11-point rating scale. While the accumulation of data-points towards the extreme ends of the intentions-measure might suggest that participants viewed

the risky and safe option as orthogonal behavioural choices, the interpretation remains difficult based on the present data.

In my next study, I sought to address these caveats by having participants make real incentivised (social or individual) choices between a clearly defined pair of a safe and risky option.

Study 2.2: Self-other differences in incentivised choices under risk

Study 2.2 aimed to further map out individual and social decisions under risk by using a task with actual incentivised rather than hypothetical outcomes, as past work has shown these contextual factors to influence social decisions (FeldmanHall et al., 2012). Hence, participants in the present study chose between a small but safe or a larger but probabilistic reduction of a (real) loss in bonus money. For participants making individual choices, the monetary loss affected only their own payoff. For participants making social choices, the monetary loss affected the payoff of an anonymous other participant, their "partner". Reflecting Study 2.2's aim of revealing preferences over safe and risky outcomes in individual and social decisions, I used a binary dependent variable that asked participants to endorse either the safe or the risky outcome. To minimise potential spillover effects between behavioural preferences and moral judgments, I focused on behavioural preferences.

Design

I used a mixed design with two within-subjects factors of seven levels (*probability*: 5%, 25%, 45%, 50%, 55%, 75%, 95% probability of loss-reduction for the risky option and *magnitude*: 15, 20, 25, 30, 40, 75, 115 points of loss-reduction for the risky option) and one between-subjects factor of two levels (*domain*: individual versus social). Participants were randomly allocated to one of the two

between-subjects conditions. The order in which participants were presented with the 49 rounds of varying probability and magnitude was randomised.

Participants

I pre-registered to collect 210 participants per condition or 420 participants in total. However, because I noticed substantial failure rates for comprehension and attention checks in the process of data collection, I increased the target sample to 350 participants per condition or 700 participants in total. Out of a total of 760 participants from the US, who were recruited via Prolific, I excluded 100 participants who failed attention or comprehension checks in line with the pre-registration, leaving a final sample of N = 660 that was 49% male with an average age of 36.84 (SD = 12.44).

Procedure

Participants learned they would make a series of choices where, on each round, they could reduce a 150 points loss in bonus payment (points were converted into cents by multiplying with a factor of 0.2). With the between-subjects manipulation, half of the participants made *individual* choices affecting their own bonus payment, whereas the other half made *social* choices affecting their partner's bonus payment. Across both groups, on each round, participants faced a choice between a safe option, reducing the loss in bonus payment by 10 points for sure, or a risky option with varying levels of probability and magnitude as illustrated in *Figure 4*.

Before starting the choice task, participants worked through an example round and completed comprehension-checks that verified they understood that choosing the safe option always meant reducing their [partner's] bonus-loss by 10 points for sure, whereas choosing the risky option meant reducing their [partner's] bonus-loss by more than 10 points at a given probability. Upon completion of the main task, participants across all conditions indicated how much they felt facing a dilemma during the task and their (hypothetical) willingness to pay for finding out the outcome of a randomly selected round where they chose the risky option. Additionally, I asked participants in the social condition how much they based their decisions on what they thought their partner would choose and how much they based their decisions on what they considered best for their partner. The study concluded with standard measures of baseline risk-preferences (Falk et al., 2016) and empathy (QCAE, Reniers et al.,

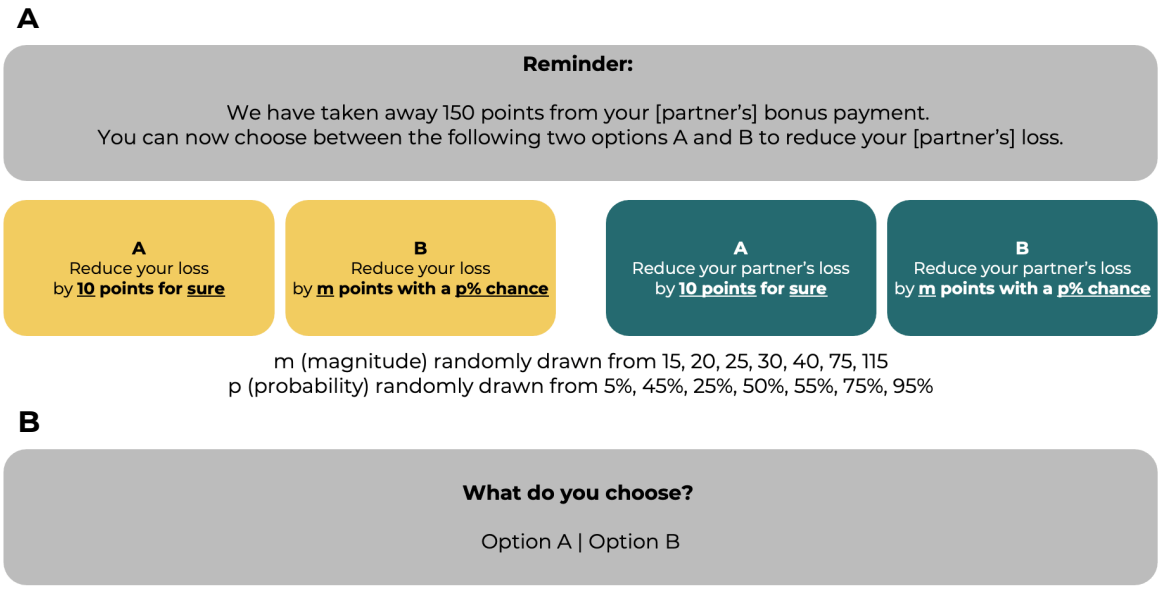


Figure 4. Schematic representation of the main task in Study 2.2. (**A**) Participants were reminded that we reduced their [partner's] bonus payment by 150 points (equivalent to 30 cents) and that they could win back some of this loss by choosing between the safe option A and the risky option B. The options for participants making individual choices are depicted in yellow and the options for participants making social choices in turquoise. The number of points and the chance they would be deducted from the pre-allocated loss varied across rounds as detailed. (**B**) Participants made a simple binary choice between the two options.

2011), as well as demographics questions. After data collection was complete, participants received a bonus payment based on their (or, in the social condition, a randomly matched other participant's) choice on a randomly selected decision-round.

Results

Participants were sensitive to variation in probability and magnitude across both individual and social choices, with proportions of risky choices increasing for higher levels of probability and magnitude (see *Figure 5*). Participants making s ocial decisions indicated they based their decisions to some degree on what they thought their partner would choose, M = 2.83, SD = 2.01, but to a larger extent on what they thought would be best for their partner (M = 6.28, SD = 1.23) according to Yuen's trimmed t-test, , $t_{yuen}(191) = -26.69$, $p < .001$, $\xi = 0.81$.

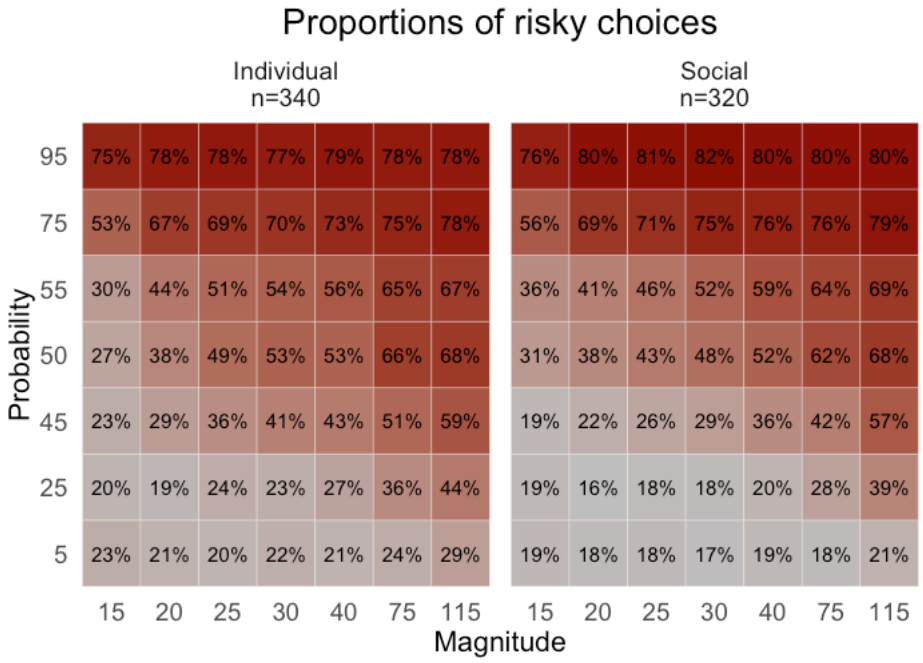


Figure 5. Heatmaps for risky choices across conditions from Study 2.2. Each tile represents the parameters for the risky option on a given trial as combination of magnitude (points; x-axis) and probability (chance that loss in points will indeed be reduced; y-axis). Numbers in the tiles indicate percentage of participants who endorsed the risky option on a given trial. Grey shades indicate higher endorsement-rates for the safe option, red shades indicate higher endorsement-rates for the risky option.

Aggregating across all decision rounds, logistic regression indicated a significant main effect of condition ($\chi^2(1) = 11.99$, $p = .001$) whereby participants making social choices were more risk-averse than those making individual ones, $b = -0.08$, $SE = 0.02$, $p = .001$, OR = 0.93, 95% CI_{OR} [0.89, 0.97]. However, this main effect was not significant ($\chi^2(1) = 1.52$, $p = .217$) in a generalized random intercept model that accounted for the nested data structure, $b = -0.11$, $SE = 0.09$, $p = .217$, OR = 0.87, 95% CI_{OR} [0.75, 1.07].

To further explore the hypothesis that probability and magnitude might differentially factor into social versus individual choices as indicated by the findings from Study 2.1, I added fixed effects for 'probability', 'magnitude', and their two-way interactions with each other and 'condition' to the random intercept model. Type III Wald chi-square tests indicated significant probability x magnitude ($\chi^2(1) = 7.12$, $p = .008$) and probability x condition ($\chi^2(1) = 26.48$, $p < .001$) interactions, which qualified main effects of probability ($\chi^2(1) = 7.12$, $p = .008$) and magnitude, $\chi^2(1) = 7.12$, $p = .008$.

Paralleling the results of Study 2.1, the interaction between probability and condition ($b = 0.18$, $SE = 0.03$, $p < .001$, OR = 1.19, 95% CI_{OR} [1.12, 1.28]) was driven by an asymmetric effect of probability on social versus individual choices: when the risky option reduced the monetary loss at low to moderate probabilities, social decisions were markedly more risk-averse than individual decisions. However, for probabilities around and above the 50% level, this pattern converged and ultimately crossed with participants making social choices endorsing the risky option more often than participants making individual choices. The interaction between probability and magnitude ($b = -0.04$, $SE = 0.02$, $p = .008$, OR = 0.96, 95% CI_{OR} [0.93, 0.99]) indicated that, again, magnitude was more decisive on decision-rounds that came with a 50% probability, whereas the relationship between

magnitude and choices was weaker for decision-rounds with more extreme probability values. Notably, roughly 20% of participants chose the risky option even when the probabilistic outcome was not likely at all (5%) and when its magnitude was only minimally larger compared to the safe option (15 points compared to 10 points); conversely, roughly 20% of participants chose the safe option even when the probabilistic outcome would have been highly likely (95%) and comparatively large (115 points). Individual differences, for instance in terms of baseline risk preferences, might provide a possible explanation for these floor and ceiling effects.

The reported main results did not change when I controlled for gender, baseline risk-preferences, and empathy (probability x condition: b = 0.18, SE = 0.03, p < .001, OR = 1.19, 95% CI_{OR} [1.11, 1.28]; probability x magnitude: b = -0.04, SE = 0.02, p = .008, OR = 0.96, 95% CI_{OR} [0.93, 0.99]). While age and empathy did not affect participants' choices (ps > .713), females were less risk-seeking than males (b = -0.24, SE = 0.12, p = .048, OR = 0.79, 95% CI_{OR} [0.62, 1.00]) and participants with stronger baseline preferences for risk-seeking were more inclined towards the risky option, b = 0.06, SE = 0.01, p < .001, OR = 1.06, 95% CI_{OR} [1.04, 1.08]. This latter finding provides preliminary support for an interpretation of the floor and ceiling effects reported in the previous paragraph whereby roughly a fifth of all participants have relatively stable baseline risk preferences leading them to invariably choose the safe or risky option according to their baseline risk preference and to neglect situational information on probability and magnitude in doing so.

To further characterise how social choices deviated from individual choices, I compared observed choices against optimal choices. I defined optimal choices by comparing the safe option's value and the risky option's expected

value, that is the product of the risky option's outcome probability and magnitude. If the safe option's value (i.e., a 10 points loss-reduction for sure) exceeded the expected value of the risky option for a given round (e.g., an expected value of 7.50 for a round with a 25% chance of reducing the loss by 30 points), the safe option was defined as optimal and vice versa. Then, for each decision round, I determined whether a given participant deviated from optimal choice by computing a simple difference score between observed and optimal choice, and if so, in which direction they deviated with negative scores indicating more risk-averse choices than optimal and positive scores indicating more risk-seeking choices than optimal. Both individual (M = -0.19, SD = 0.56) and social choices (M = -0.21, SD = 0.55) were risk-averse relative to optimal choice (ps < .001 for one-sample t-tests). A linear regression model predicting deviation scores by 'condition' indicated a significant main effect ($F(1, 32338) = 9.03$, $p = .003$) whereby social decisions deviated stronger towards risk-aversion than individual decisions (b = -0.02, SE = 0.01, $p = .003$), but the effect size was infinitesimal small, and the effect was indeed not significant in a random-intercept model controlling for the nested data structure ($p = .284$). To obtain more power, I further simplified the analysis by computing a score that merely indicated whether or not observed choices aligned with optimal choices. A generalized random intercept predicting this binary alignment score based on 'condition' and 'optimal choice' (safe versus risky) indicated that participants making social decisions were overall more aligned with optimal choice relative to those making individual decisions, $b = 0.22$, $SE = 0.11$, $p = .049$, OR = 1.25, 95% CI_{OR} [1.00, 1.56]. This effect was driven by social and optimal choices aligning significantly more often than individual and optimal choices for decisions rounds on which choosing the safe option was optimal, signified by an interaction between

condition and optimal choice, b = -0.33, SE = 0.06, p < .001, OR = 0.72, 95% CI_{OR} [0.64, 0.82].

Participants making social choices were as likely as those making individual choices to indicate a positive willingness to pay from their own bonus payment for learning the outcome of a randomly selected round on which the risky option was chosen, b = -0.03, SE = 0.20, p = .882, OR = 0.97, 95% CI_{OR} [0.65, 1.45]. Of those who did indicate a willingness to pay, participants who made social decisions trended towards a higher willingness to pay (M = 20.32, SD = 17.76) than participants who made individual decisions (M = 16.66, SD = 17.82) but the difference was not statistically significant according to Welch's t-test, $t(67.64) = 1.52$, p = .133, d = -0.30, 95% CI_d [-0.61, 0.10]. Finally, participants making social decisions felt less like facing a decisional dilemma during the task, M_{social} = 3.65, SD_{social} = 1.60; $M_{individual}$= 4.04, $SD_{individual}$ = 1.65; $t(392.34) = 2.53$, p = .012, d = 0.25, 95% CI_d [0.06, 0.42].

Discussion

Using incentivised choices, Study 2.2 replicated the findings from Study 2.1, whereby an asymmetric effect of probability produces overall more risk-averse choices in social relative to individual decisions. In line with the hypothesised precautionary pattern, this was particularly the case for lower probabilities. Additionally, participants making social decisions indicated that they based their choices more on what they considered best for their partner than what they thought their partner would prefer. This paternalistic rationale might also be interpreted in line with a precautionary motive in decisions for others: for risky choices where participants might take a (unfavourable) chance when they bear the outcomes themselves, they think it is best for others to err on the safe side. This also became apparent in social decisions being more

aligned with optimal choices on rounds where choosing the safe option was best. Furthermore, participants' concern for their partners' outcomes in risky choices was reflected in a trend-wise higher willingness to pay for learning the outcome of such a decision.

By choosing what they considered best for their partner, participants making social decisions were more aligned with optimal choices based on expected value than participants making individual decisions. This was particularly the case for decision rounds when choosing the safe option was optimal in terms of expected value, indicating participants making individual choices opted for the risky option more often on these rounds. This observation connects two findings that have emerged in previous research, namely that the worst-case outcome anchors decisions for others (Kameda et al., 2016) and that decisions for others are sometimes attenuated in a more risk-neutral direction relative to decisions for self (Batteux et al., 2017; Ogawa et al., 2018; Sun et al., 2017; X. Zhang et al., 2017). The present findings indicate that an attenuation towards risk-neutrality in choices for others might occur precisely when choosing a risky option may be tempting but not optimal in terms of expected value and that this tendency is motivated by a benevolent consideration of the worst-case outcome (here, the risky loss-reduction not materialising).

Considering that – despite distorted weighting and value functions producing suboptimal choices under risk (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) – choices for self are ultimately motivated by outcome maximization (e.g., Platt & Huettel, 2008), an alternative explanation for the observed differences in individual and social decisions could be that participants choosing for others are foregoing risky prospects not because of precautionary motives, but because they are less motivated to maximise (risky) prospects in

choices for others. In Study 2.3, I sought to put this alternative explanation via a reduced motivation to maximise outcomes for others at test.

Study 2.3: Dissociating the roles of uncertainty and motivation for self-other differences in choices under risk

Paralleling the task deployed in the previous study, participants in Study 2.3 made incentivised choices between a small but safe reduction of a monetary loss or a risky but larger monetary loss reduction. Thus, the design included an *individual* and a '*social baseline'* condition analogous to the design of Study 2.2: individual choices affected participants' own bonus payment; social choices affected a randomly paired other participant's (i.e., their "partner's") bonus payment. I added two additional *social* conditions to triangulate the role of motivation for driving differences between individual and social decisions under risk. In particular, to enhance motivation in social decisions, one group of participants could earn a considerable bonus payment if their decisions matched what their partner would have chosen for themselves. Because social decisions in this condition were incentivised and geared towards matching the partner's preferences, I refer to it as *social + incentivised match-motivation* condition.

In the *social baseline* condition, participants might have followed a different motivational goal than matching their partner's preferences. Indeed, the findings from Study 2.2 indicated that participants making social decisions chose what they considered best for their partner rather than what they thought their partner would prefer. As such, incentivising participants to match their partner's choice might produce a motivational shift in social decisions. To control for this potential confound in the *social + incentivised match-motivation*

condition, I added another social condition in which participants were just told that I would compare their choices on behalf of the partner with what the partner would have chosen for themselves, but I did not incentivise participants' choices. Thus, I refer to this group as the *social + match-motivation* condition.

In line with Studies 2.1 and 2.2., I expected to replicate enhanced risk-aversion in social relative to individual choices when comparing the *social baseline* and *individual* conditions. If this difference were indeed (partly) driven by decreased motivation in social relative to individual decisions, it should be diminished when comparing the *social + incentivised match-motivation* condition with the *individual* condition. Meanwhile, if the previously observed differences persisted across all three social conditions, this would speak against motivational differences driving the divergence of risky choices for self and for others. To further explore the alternative hypothesis that enhanced uncertainty contributes to more risk-aversion in social relative to individual decisions, I included a second dependent variable that asked participants how uncertain or certain they felt in their decision on any given round.

Design

I used a mixed design with two within-subjects factors of seven levels (*probability*: 5%, 25%, 45%, 50%, 55%, 75%, 95% probability of loss-reduction for the risky option and *magnitude*: 15, 20, 25, 30, 40, 75, 115 points of loss-reduction for the risky option) and one between-subjects factor of four levels (*domain*: 'individual' versus 'social baseline' versus 'social + incentivised match-motivation' versus 'social + match-motivation'). Participants were randomly allocated to one of the four between-subjects conditions. The order in which participants were presented with the 49 rounds of varying probability and magnitude was randomised.

Participants

My sampling goal was n = 210 participants per condition (N = 840) after excluding and replacing participants who failed the attention or comprehension-check, in line with the pre-registration. To this end, I recruited US participants from MTurk in three subsequent batches, in-between which I ran no analyses on the data but merely excluded cases that failed either check in order to estimate the required number of participants to replace them. This process yielded 1,430 recorded responses, including 210 incomplete sessions, ten duplicate cases, 83 participants who failed comprehension, and 108 participants who failed attention checks. Cases falling in the latter two categories were excluded and replaced as pre-registered. The resulting final sample of N = 1,019 was 47% male with an average age of 36.63 (SD = 11.52).

Procedure

I used the same task and instructions as in Study 2.2. However, this time participants were randomly allocated to one of four experimental conditions, summarised here:

- *Individual choices*. Participants in this condition could win back some of their own bonus payment by choosing between a small but certain reduction of a bonus-payment loss, or a larger but probabilistic reduction of a bonus-payment loss. Losses were deducted from these participants' own bonus payments after completion of data collection.

Social choices +

- *Incentivised match-motivation*. Participants in this condition were randomly paired with another participant in the study, their "partner".

They could win back some of their partner's bonus payment by choosing between a small but certain reduction of a bonus-payment loss for their partner or a larger but probabilistic reduction of a bonus-payment loss for their partner. These participants were also told that I asked their partner what they would choose if the decisions were up to them, and that they could earn a $1 bonus payment if their choice matched their partner's choice on a randomly selected decision-round. The reminder displayed on each decision-round (see *Figure 4*) also told participants to "remember that we will compare your decision and your partner's decision for the round that is randomly picked at the end of the study. You get a $1 bonus if your choice if that round matches what your partner said they would choose if it was up to them". Losses were deducted from randomly paired participants' bonus payments after completion of data collection.

- *Match-motivation*. Participants performed the same task as participants in the *incentivised match-motivation* condition, but they did not receive a bonus payment for matching their partner's choices. Accordingly, the reminder displayed on each decision-round (see *Figure 4*) told participants to "remember that we will compare your decision and your partner's decision for the round that is randomly picked at the end of the study".
- *Baseline*. Participants in this condition performed the same task as participants in the other two *social* conditions, but they were not told that I would collect their partner's preferences and compare those against their choices.

In addition to the comprehension checks used in Study 2.2, which verified that participants understood the safe and risky option, participants in the *social + match-motivation* and *social + incentivised match-motivation* conditions answered questions that verified they understood I would compare their and their partner's response for a randomly selected round and that they could receive a $1 bonus for matching responses in the incentivised condition. Participants then started the main task, choosing between the safe and the risky option for 49 decision-rounds. Additionally, on each round, they indicated how uncertain or certain they felt in their decision on a 6-point rating scale from "1 – very uncertain" to "6 – very certain". On each round, participants were reminded of their task and in the social conditions of how their choices would affect their partner (and that they could earn a $1 bonus in the incentivised condition). Upon completing the main task, participants across all conditions indicated how much they felt facing a dilemma during the task. Additionally, I asked participants in the social conditions how much they based their decisions on what they thought their partner would choose and how much they based their decisions on what they thought would be best for their partner. The survey concluded with standard measures of baseline risk-preferences (Falk et al., 2016) and empathy (QCAE; Reniers et al., 2011), as well as demographics questions.

Results

First, as a manipulation check, I tested whether participants in the *social* conditions differed in how much they said they based their choices on what they thought their partner would choose if it were up to them. A Kruskal-Wallis test indicated significant differences between the three *social* conditions, $\chi^2(2) = 33.32$, $p < .001$, $\eta_H^2 = 0.04$, 95% CI [0.02, 0.09]. Confirming the effectiveness of the motivation-manipulation, posthoc comparisons using the Dwass-Steel-

Critchlow-Fligner method specified that participants in the *social + incentivised match-motivation* condition reported basing their choices on the partner's preference to a larger extent (M = 4.15, SD = 1.94) relative to the *social + match-motivation* (M = 3.43, SD = 1.98, p < .001) and the *social baseline* (M = 3.13, SD = 2.08, p < .001) conditions, which did not differ from another (p > .05). There were no significant differences across the three *social* conditions in regard to how much participants reported to base their choices on what would be best for their partner, $\chi^2(2) = 4.37$, $p = .112$. Overall, participants making social decisions indicated to base their choices more on what they thought would be best for their partner ($M_{\text{all social}} = 6.28$, $SD_{\text{all social}} = 1.33$) than on what they thought their partner would choose, $M_{\text{all social}} = 3.57$, $SD_{\text{all social}} = 2.04$, $t_{\text{yuen}}(450) = 24.79$, $p < .001$, $\xi = 0.75$.

Choices. Participants were sensitive to variation in probability and magnitude across all conditions, with proportions of risky choices increasing at higher levels (*Figure 6*). Aggregating across all decision rounds, logistic

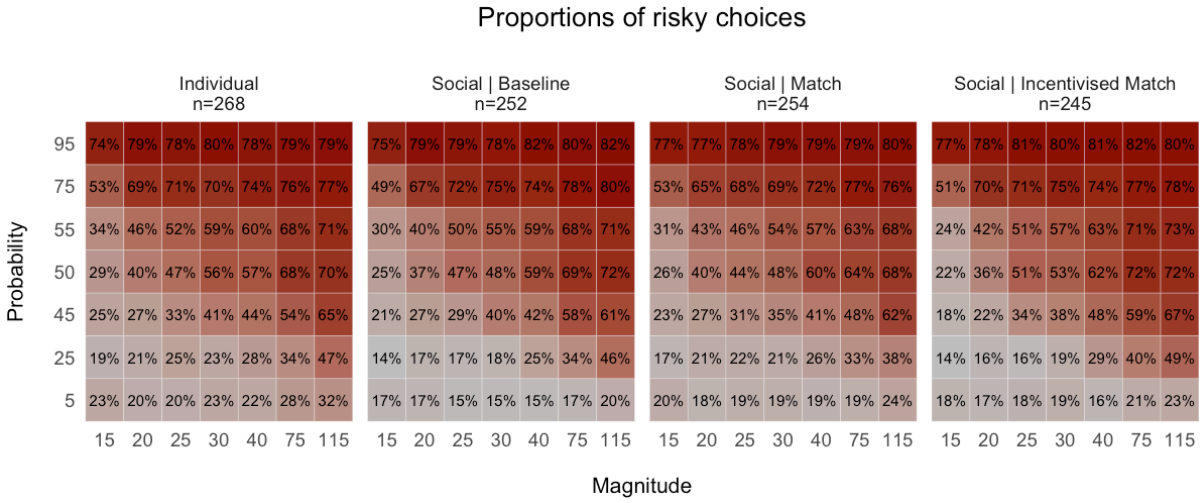


Figure 6. Heatmaps for risky choices across conditions from Study 2.3. Each tile represents the parameters for the risky option on a given trial as combination of magnitude (points; x-axis) and probability (chance that loss in points will indeed be reduced; y-axis). Numbers in the tiles indicate percentage of participants who endorsed the risky option on a given trial. Grey shades indicate higher endorsement-rates for the safe option, red shades indicate higher endorsement-rates for the risky option.

regression indicated significant differences across conditions ($\chi^2(3) = 21.38$, $p < .001$) that were driven by social choices being more risk-averse than individual ones in the *social baseline* ($b = -0.09$, $SE = 0.03$, $p < .001$, OR = 0.92, 95% CI_{OR} [0.87, 0.96]) and *social + match* conditions ($b = -0.10$, $SE = 0.03$, $p < .001$, OR = 0.90, 95% CI_{OR} [0.86, 0.95]) but not in the *social + incentivised* condition, $b = -0.04$, $SE = 0.03$, $p = .148$, OR = 0.96, 95% CI_{OR} [0.92, 1.01]. However, similar to Study 2.2, this main effect was not significant in a generalized random intercept model accounting for the nested data structure, $\chi^2(3) = 2.45$, $p = .449$.

Next, I explored differences in risky choices across conditions and variations of the probability and magnitude parameters. Again, I used a mixed-effects model with 'probability', 'magnitude', 'condition', and their two-way interactions entered as fixed effects and 'subject' as random intercept effect. Type III Wald chi-square tests replicated the significant interactions observed in previous study, namely between probability and condition ($\chi^2(3) = 54.40$, $p < .001$), probability and magnitude ($\chi^2(1) = 7.95$, $p = .005$), and additionally magnitude and condition, $\chi^2(3) = 25.35$, $p < .001$).

The interaction between probability and condition (see

Table 2 for statistical details) was again driven by an asymmetric effect of probability on social versus individual decisions, which persisted across the manipulations of motivation (see *Figure 7*): when the risky option reduced the monetary loss at low to moderate probabilities, social decisions were markedly more risk-averse than individual decisions. However, for probabilities around and above the 50% level, this pattern converged and participants making social decisions endorsed the risky option as much as participants making individual decisions. This pattern also emerged for the comparison between *individual* and

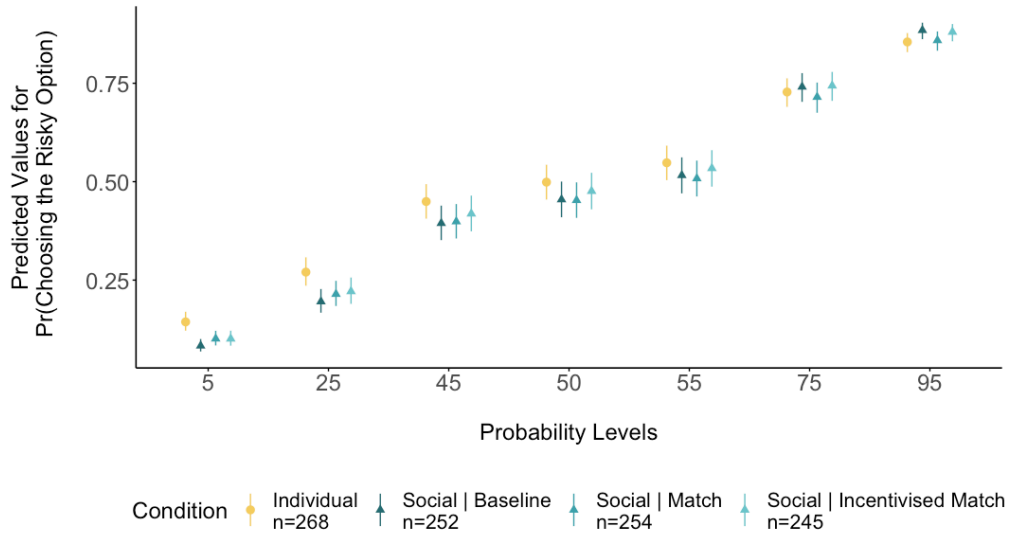


Figure 7. Predicted values for of risky choices for probability x condition interaction in Study 2.3 (also see *Table 2*). Error bars represent 95% confidence intervals.

social + incentivised match-motivation decisions.

Social decisions in the *baseline* and *incentivised match-motivation* were more sensitive to magnitude relative to individual decisions, whereas there was no difference between individual and social *match-motivation* choices (

Table 2). Similar to the previous studies, a significant interaction between probability and magnitude indicated that magnitude was more decisive on decision-rounds that came with a 50% probability, whereas the relationship between magnitude and choices was weaker for decision-rounds with more extreme probability values.

Table 2. Generalized random intercept model predicting choices (Study 2.3)

Variable	*b* (SE)	*p*	OR	95% CI_{OR}
(Intercept)	0.00 (0.09)			
Probability	1.09 (0.03)	< .001	2.98	[2.84, 3.13]
Magnitude	0.46 (0.02)	< .001	1.58	[1.51, 1.65]

Social baseline	-0.18 (0.13)		.173	0.84	[0.65, 1.08]
Social + match-motivation	-0.18 (0.13)		.155	0.83	[0.64, 1.07]
Social + incentivised match motivation	-0.09 (0.13)		.485	0.91	[0.71, 1.18]
Probability x magnitude	-0.04 (0.01)		.005	0.96	[0.94, 0.99]
Probability x social baseline	0.27 (0.04)	<	.001	1.31	[1.22, 1.41]
Probability x social + match motivation	.13 (0.04)	<	.001	1.14	[1.06, 1.23]
Probability x social + incentivised match-motivation	0.19 (0.04)	<	.001	1.21	[1.12, 1.31]
Magnitude x social baseline	0.1 (0.03)		.002	1.11	[1.04, 1.18]
Magnitude x social + match-motivation	-0.01 (0.03)		.712	0.99	[0.93, 1.05]
Magnitude x social + incentivised match-motivation	0.12 (0.03)	<	.001	1.13	[1.06, 1.21]

Again, these results did not change when I controlled for gender, age, baseline risk-preferences, and empathy (*p*s ≤ .005) and while gender, age, and empathy did not affect participants' choices, participants with stronger baseline preferences for risk-seeking were more inclined towards the risky option, b = 0.07, SE = 0.02, p < .001, OR = 1.07, 95% CI_{OR} [1.04, 1.11]. Notably, the relationship between baseline risk preferences and choices in this present study thus was comparable to the one observed in Study 2.2, as was the proportion of about 20% participants choosing the risky option even when it almost certainly yielded lower expected value than the safe option and a similar proportion choosing the safe option even when the risky option almost certainly yielded substantially higher expected value. Together, these findings support and interpretation of these floor and ceiling effects in terms of baseline risk preferences that might be

distributed relatively equally in the US population from which I recruited the separate samples for the reported studies.

Following the procedure that showed more power in Study 2.2, I evaluated the alignment of observed choices with optimal choices using a simple binary alignment-score. A generalized random intercept model that predicted alignment based on 'condition' and 'optimal choice' (safe versus risky), besides a random intercept effect for 'subject', indicated that participants in the *social baseline* and *social + incentivised match-motivation* conditions overall aligned more often with optimal choice relative to those making *individual* decisions (*social baseline: b* = 0.43, *SE* = 0.14, *p* = .002, OR = 1.54, 95% CI_{OR} [1.18, 2.01]; *social + incentivised match-motivation: b* = 0.52, *SE* = 0.14, *p* < .001, OR = 1.68, 95% CI_{OR} [1.28, 2.20]). There was no difference for the comparison of the *individual* and *social + match-motivation* conditions, *p* = .109. Paralleling the results from Study 2.2, relative to individual decisions, social decisions (in the *baseline* and *incentivised match-motivation* conditions) were more aligned with optimal choice for rounds on which choosing the safe option was optimal. This was signified by an interaction between condition and optimal choice (*social baseline: b* = -0.44, *SE* = 0.07, *p* < .001, OR = 0.64, 95% CI_{OR} [0.56, 0.74]; *social + incentivised match-motivation: b* = -0.41, *SE* = 0.08, *p* < .001, OR = 0.66, 95% CI_{OR} [0.57, 0.77]).

Uncertainty ratings. To validate that the uncertainty ratings were meaningful, I first tested whether participants overall felt more certain on rounds when they opted for the safe option and on rounds with low (5%, 25%) and high (75%, 95%) probabilities relative to rounds with probabilities around chance-level (45%, 50%, 55%). A random intercept model with 'choice', 'probability' (clustered into three levels: low, medium, high), and their two-way

interaction entered as fixed effects besides a random intercept effect for 'subject', indicated that significant main effects for choice ($\chi^2(1) = 1341.25$, $p < .001$) and probability ($\chi^2(2) = 812.65$, $p < .001$) were qualified by a significant interaction, $\chi^2(2) = 2142.19$, $p < .001$ (see

Table 3): When participants chose the safe option, their uncertainty ratings described a U-shaped function signifying more confidence in low- and high-probability decision rounds relative to medium-probability decision rounds. Meanwhile, when they chose the risky option, participants' confidence varied in par with the probability level on a given round: they felt the least confident on low probability rounds and the most confident on high probability rounds.

Table 3. Random intercept model predicting uncertainty ratings (Study 2.3)

Variables	*b* (SE)	*p*	*d*	95% CI_d
(Intercept)	5.13 (0.02)			
Choice (risky)	-0.98 (0.02)	< .001	0.23	[0.21, 0.24]
Medium probability	-0.36 (0.01)	< .001	0.13	[0.11, 0.14]
High probability	-0.14 (0.02)	< .001	0.03	[0.02, 0.05]
Choice x medium probability	0.47 (0.02)	< .001	0.09	[0.07, 0.11]
Choice x high probability	1.29 (0.03)	< .001	0.20	[0.18, 0.22]

Note. Estimates from random intercept model. Probability was clustered into three levels: low (5%, 25%), medium (45%, 50%, 55%), and high (75%, 95%).

Next, I explored the role of probability and magnitude for participants' reported uncertainty across individual versus social decisions, analogously to the parameters' role in choices between the safe and risky option. To this end, I added a fixed effect for 'magnitude' to the above random intercept model, as well as three-way interactions between 'choice', 'probability'/'magnitude', and 'condition'. Type III Wald tests indicated that three-way interactions between probability, condition, and choice ($\chi^2(9) = 1722.17$, $p < .001$) as well as between magnitude, condition, and choice ($\chi^2(3) = 45.38$, $p < .001$) qualified lower-order effects (see *Table A1* in *Appendix D*: Additional results). More specifically, participants making social decisions reported greater uncertainty relative to those making individual decisions – a pattern particularly pronounced when participants chose the risky option and for decisions rounds with probabilities around the 50% level (*Figure 8*). Furthermore, while participants generally responded to variations in magnitude by feeling less certain for higher magnitudes when they endorsed the risky option, those in the *social +*

incentivised match-motivation condition did not feel markedly less certain for high magnitudes when they endorsed the risky option but when they endorsed the safe option.

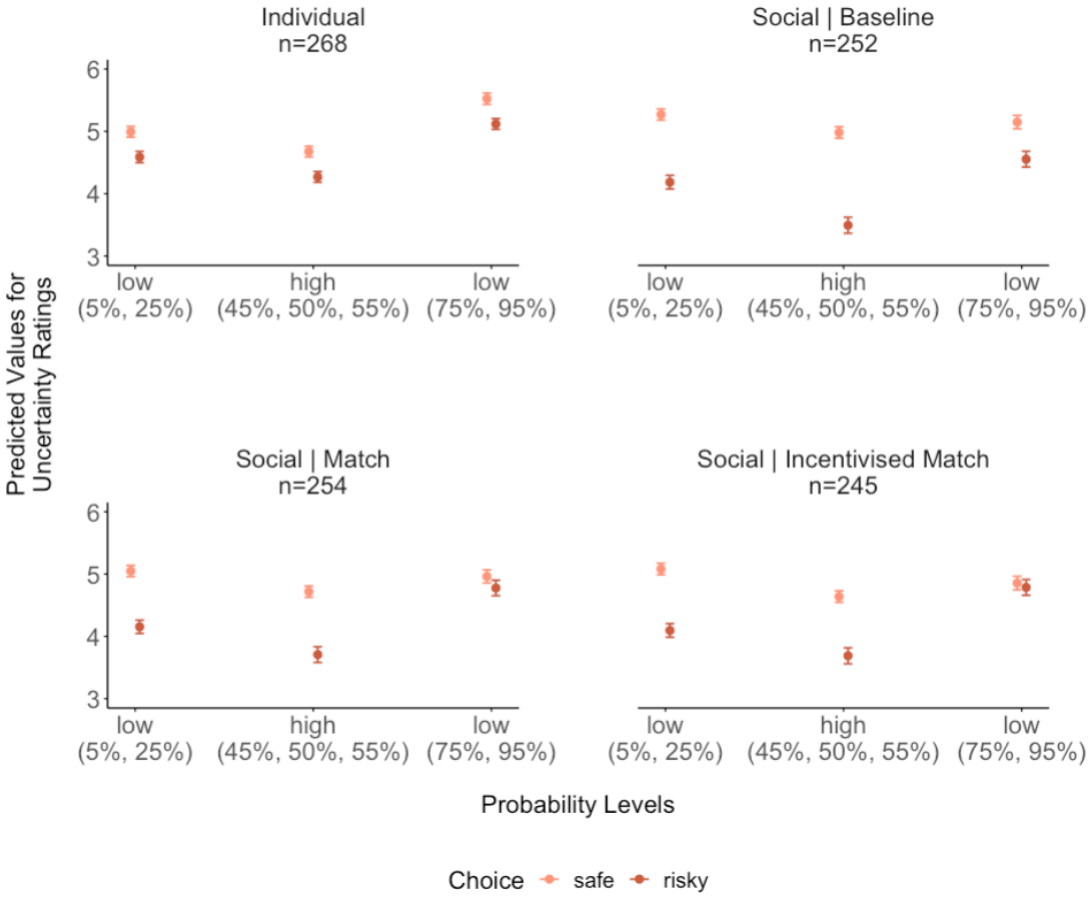


Figure 8. Predicted values of uncertainty ratings for probability x choice x condition interaction in random intercept model for Study 2.3. Error bars represent 95% confidence intervals.

Discussion

Study 2.3 replicated the previously observed pattern whereby social decisions are overall more risk-averse than individual ones, particularly for low probabilities of a desirable outcome. This was still the case when I experimentally enhanced participants' motivation to make social decisions in the interest of others, although differences between social and individual decisions were less pronounced in this case. Participants reported greater decisional uncertainty

when making social decisions, especially when endorsing the risky option and when decision outcomes were uncertain (i.e., (not) materialising with a 50% chance). Again, replicating observations from Study 2.2, participants reported making social decisions from a benevolent rather than a simulated perspective and, in doing so, made decisions that were more aligned with optimal choice based on expected value relative to participants choosing for themselves.

In summary, the findings from Study 2.3 support the hypothesis that social decisions elicit a precautionary response pattern with deciders erring on the safe side when the odds of a desirable outcome are low and when the expected value of the safe option is higher or equal than the risky option. The experimental manipulation of participants' motivation in social decisions did attenuate the asymmetric effect of probability on social compared to individual choices, but not to a significant degree when accounting for the asymmetric effect of probability. Together with the observation that not only choices in the *social + incentivised match-motivation* condition but also choices in the *social baseline* condition were more sensitive to variations in magnitude, these results highlight that decisions under risk for others cannot be explained in terms of people simulating others' preferences, nor by decreased motivation to integrate outcome-information. Meanwhile, the finding that participants making social decisions felt particularly less confident about their choices when they endorsed the risky option suggests that risk-taking for others might be associated with a more aversive experience that could, in turn, explain why participants erred on the safe side more often in social relative to individual decisions.

Overall discussion

Across three studies, I showed that choices for others are markedly more risk-averse than choices for self when the odds of a positive outcome are low, while they converge for higher odds of a positive outcome (Studies 2.1-2.3). In line with the hypothesis that enhanced uncertainty in social choices might produce such a precautionary pattern, I observed greater indecision (Study 2.1) and less confidence (Study 2.3) in social relative to individual decisions, particularly for low outcome probabilities and high magnitudes of a desirable outcome. Meanwhile, I found little evidence for motivational differences in risky choices for others versus self that could explain the observed pattern (Study 2.3).

Previous work on social versus individual decisions under risk has produced inconsistent results that cannot be satisfactorily reconciled by considering moderating factors (Atanasov, 2015; Polman & Wu, 2020). My findings offer an alternative explanation whereby enhanced uncertainty in social decisions produces a precautionary pattern with risky choices for others diverging from risky choices for self, especially when the odds of a desirable outcome are low. Meanwhile, differences between social and individual decisions under risk converge when precautionary preferences are less important, that is, when a desirable outcome is more likely. I tested this proposition while holding constant constraints such as decision frame, social distance, or reciprocity summoned in previous accounts for differences in risky choices for others versus self. While this approach allowed me to identify baseline differences between social and individual decisions under risk, and to run a feasible set of studies, an important next step will be to explore potential qualifying factors. For instance, to sidestep framing effects, I used a decision frame that combined features of loss and gain frames with participants making

choices about reducing a loss. It will be interesting to test whether the asymmetric effect of probability on social relative to individual choices is enhanced in simple gain-frames but diminished in simple loss-frames.

Considering existing work on differences in risky decisions for others versus self more broadly, my experiments provided a conservative test for the hypothesised precautionary pattern. For instance, existing work suggests that studies examining social decisions affecting an anonymous rather than a familiar other person – as in my studies – tend to produce a risky shift in choices for others (Montinari & Rancan, 2013; Sun et al., 2017); that the absence of reciprocity entails a larger risky shift in choices for others (Polman & Wu, 2020); and that gain-framed, but to a smaller extent also loss-framed, decisions exhibit a risky shift in choices for others (Polman, 2012; X. Zhang et al., 2017, 2019). Thus, by endorsing the very conditions under which previous work has, on average, observed a risky shift in decisions for others, I subjected my hypotheses to a strong test. That said, it will be important to extend the present work by explicitly looking at variations in boundary conditions and by further developing the proposition of precautionary preferences in social decisions under risk.

In particular, the preliminary evidence I provide for the uncertainty hypothesis, that is, enhanced uncertainty in choices for others driving precautionary preferences amidst low odds of a desirable outcome, requires further attention. Instead of using posthoc uncertainty ratings about decisional uncertainty, behavioural measures such as providing participants with a choice between a safe, a risky, and a coinflip option (similar to Edelson et al., 2018) could offer a more insightful approximation for measuring aversely experienced uncertainty. I relied on online studies that limited the number of rounds with which I could present participants for feasibility reasons. Besides allowing for a

larger number of rounds, including trial repetitions as well as more rounds where the expected value of the probabilistic option and safe option's value are equal, lab studies could elucidate how outcome information is processed differently in social relative to individual decisions by deploying techniques such as eye-tracking and computational modelling. Such approaches would also allow to characterise and compare weighting- and value-functions across domains, which could be a stepping stone for extending accounts of individual decision-making under risk, such as Prospect Theory (Tversky & Kahneman, 1992), to social decision-making under risk.

In the present work, I explored integral differences in social and individual decision-making under risk, focusing on the role of probability and magnitude while holding constant potential moderating factors (e.g., decision frame, social distance, reciprocity). In line with my proposition that enhanced uncertainty in social decisions leads people to adopt precautionary preferences, I observed an asymmetric effect of probability whereby choices for others were more risk-averse than choices for self when the odds of a desirable outcome were low. Notably, in exhibiting this precautionary pattern, social decisions matched more closely with optimal choice based on expected value compared to individual decisions. My results further affirmed that participants experienced enhanced decisional uncertainty in choices for others, while motivational differences related to outcome-maximization between social and individual choices could not explain the divergences. As such, the present work contributes to our understanding of what is unique in social decision-making under risk as opposed to individual decision-making under risk, and how such differences translate into systematic divergences. While I focused on minimal designs that sidestepped potential moderating factors such as decision frames or social

distance, exploring how those interact with the baseline differences established in this work provides a promising avenue for future research and for building a theoretical framework for social decision-making under risk that can bridge the inconsistencies in existing accounts.

I started by exploring social decisions under uncertainty in the context of risky choices, because this context ties in naturally with how individual decisions under uncertainty have most commonly been operationalized (e.g., Christopoulos et al., 2009; Kahneman & Tversky, 1979; Platt & Huettel, 2008). In contrast, most decision situations that we face in our daily lives are not defined well enough to stipulate a set of possible outcomes, leave alone associated probabilities (Camerer & Weber, 1992). While existing research has started to map out the interplay between uncertainty about the outcomes of our decisions for others and uncertainty about the impact of those outcomes on others' well-being (Kappes et al., 2019), characterising social decisions in situations with truly undefined outcomes remains a notoriously hard challenge (Johnson et al., 2020; Tuckett & Nikolic, 2017). An alternative avenue towards advancing our understanding of uncertainty in social decisions is to consider choices with well-defined deterministic outcomes. In the following chapter, I stride this avenue which allows me to focus on the role of uncertainty about others' experience while sidestepping any potential interactions with uncertainty about outcomes for others.

3 Uncertainty's role in harm aversion for others versus self

In the previous chapter, I proposed that uncertainty in social decisions might lead people to adopt precautionary preferences and tested this hypothesis in the context of risky choices. In Chapter 3, I focus on choices with deterministic outcomes for others and seek to explore uncertainty's role in a particular manifestation of precautionary social preferences, namely a stronger aversion to harming others than harming oneself (Crockett et al., 2014, 2015, 2017; L. J. Volz et al., 2017). Recent work has studied how people value harm for others relative to harm for themselves using a task where participants trade off between monetary payoffs and mildly painful electric shocks experienced either by themselves or an anonymous other person (also see section 2.1; Crockett et al., 2014, 2015, 2017; Vlaev et al., 2009). Participants' decisions in this task demonstrate a "hyperaltruistic" valuation of others' pain, signified by less harmful choices for others and fewer shocks delivered to others relative to themselves, despite incurring greater personal sacrifices in terms of monetary payoffs (ibid.). Put differently, people seem to put a greater (subjective) cost on delivering pain to others than pain to themselves (Crockett et al., 2017). This observation cannot be explained from an empathy perspective, which would predict that people value others' pain as a function of their empathizing with them and at maximum similar to their own pain (Batson et al., 1981; T. Singer, Seymour, O'Doherty, et al., 2004). Furthermore, the experimental procedures employed during this previous works went great lengths in eliminating strategic motives such as reciprocity or reputational concerns (see, e.g., Crockett et al., 2014).

One explanation which remains feasible is that people cling to precautionary preferences amidst uncertainty about their decision's impact on others' well-being. Indeed, Crockett and colleagues (2014) observed a positive association between the degree to which participants valued others' pain more than their own pain and the degree to which they were slower in choices for others than themselves (Crockett et al., 2014). Given that greater decisional uncertainty usually is associated with slower response times (De Martino et al., 2013), this observation falls in line with the explanation that enhanced uncertainty in decisions that impact others triggers precautionary preferences. Similarly, decisional uncertainty is commonly reflected in choice noisiness: low uncertainty tends to translate into consistent and deterministic choices whereas high uncertainty tends to produce inconsistent and erratic choices (ibid.). Crockett and colleagues (2014) observed that the degree to which participants were noisier in choices for others than themselves was associated with greater hyperaltruistic valuation of others' pain, and that those who were noisier in choices for others than self were driving the overall hyperaltruistic pattern.

More evidence in support of a relationship between social uncertainty and enhanced aversion to harming others comes from work that has delineated uncertainty about our decisions' outcomes for others from uncertainty about the impact of those outcomes on others' well-being (Kappes et al., 2018). While outcome uncertainty seems to cater self-serving narratives and overly optimistic beliefs about negative outcomes for others not materializing (Dana et al., 2006; Haisley & Weber, 2010), thereby facilitating selfish behaviour coming with costs to others, impact uncertainty seems to lead people to err on the side of caution and to choose whatever minimizes potential harm to others, even at the cost of personal sacrifices (Kappes et al., 2018). In the context of the harm aversion task

discussed above, outcome uncertainty would involve ignorance about one's decision consequences for the other person – for instance, similar to Study 2.1, participants might not know whether or not the shocks associated with a particular choice are actually administered to the other person. Meanwhile, impact uncertainty relates to ignorance about how the decisional consequences will affect others' well-being, specifically the mapping of the number of shocks on the others' subjective unpleasantness in experiencing those shocks.

Importantly, while the original harm aversion studies involved deterministic choice consequences and hence no outcome uncertainty, impact uncertainty was present as it is inherently inevitably in any interaction with others. Indeed, Crockett and colleagues (2014) report that many participants alluded to (what I would refer to as *impact*) uncertainty in explaining their decisions, as illustrated by this remark from a participant: "I knew what I could handle but I was less sure about the other person and didn't want to be cruel" (p. 17323). Together with the observation that impact uncertainty's effect on (pro)social behaviour does not hinge on the presence of outcome uncertainty (i.e., these two variants of uncertainty do not seem to interact; Kappes et al., 2018), this warrants for an experimental test of impact uncertainty's role in hyperaltruistic preferences.

Study 3: Hyperaltruism under experimentally enhanced impact uncertainty

To this end, I ran a laboratory study in which I manipulated uncertainty about the impact of one's decisions on another person's well-being in the harm aversion paradigm. In contrast to the empirical work presented in Chapter 2, outcomes were deterministic to facilitate relatability to previous work and to

buffer potential confounding effects of this different uncertainty variant. To enhance impact uncertainty, I instructed participants that people would differ greatly in how they experience the electric shocks employed in the harm aversion paradigm, with some finding them extremely unpleasant but others not unpleasant at all; and that hence, they could not assume others to experience the shocks similarly to themselves. Thus, by highlighting variability in subjective experience, I sought to reinforce the inherent uncertainty about how one's decisions might impact others' well-being.

I contrasted this *high impact uncertainty* condition with another treatment condition in which participants were explicitly instructed that there is low variability in people's shock experience and that they could hence assume others to have a similar experience as themselves. This *low impact uncertainty* condition allowed to control for the effect of mentioning variability in shock experience, which previous studies have not done. In combination with a third *control* condition that did not mention variability and similarity of shock experience at all, paralleling existing work (e.g., Crockett et al., 2014), this combination of treatments more generally allowed me to relate participants' behaviour in the present study to previous work. For example, small or non-significant differences between the high impact uncertainty and control conditions would indicate that the level of impact uncertainty experienced by participants in previous studies was inherently high.

If impact uncertainty contributes to hyperaltruistic valuation of pain, then explicitly highlighting it might enhance the difference in harm aversion for other versus self, or at least produce differences similar to previously observed ones. Conversely, explicitly reducing impact uncertainty might diminish hyperaltruistic valuation of pain. Following this logic, but more in the spirit of a

manipulation check, self-other differences in indicators of decisional uncertainty, particularly reaction times and choice noisiness, might reflect the level of impact uncertainty: I expected high impact uncertainty to produce slower and noisier choices for other than self, whereas I expected low impact uncertainty to diminish or even cancel such differences.

Study 3 was approved by the University of Oxford's Central University Research Ethics Committee (MSD-IDREC-C1-2015-001). Participants gave their informed consent beforehand and were compensated at an hourly rate of £10 for their participation (with choice-dependent bonus payments paid on top). Together with my supervisor and collaborators Andreas Kappes, Julian Savulescu, and Molly J. Crockett I am preparing a manuscript for publication based on this work.

Design

I used a mixed design in which participants were randomly allocated to one of three between-subjects conditions (*impact uncertainty*: high, low, control) and went through 160 decision rounds of the main task that were randomized in order. Participants did not know there were three different conditions.

Participants

I recruited 123 healthy volunteers from the University of Oxford Department of Experimental Psychology's participant pool and from the general public through poster adverts. A pre-screening survey ensured that recruited participants were between 18 and 35 years of age, had not participated in 10 or more studies at the Department of Experimental Psychology, had not studied psychology or economics as a university degree for more than one year, had no

history of psychiatric illness, and had not previously participated in a study involving electric shocks.

I excluded a total of 10 participants, four out of whom said they had heard the other participant's voice and were substantially influenced by their suspicion about the other's gender during the task; three who said their knowledge of the Milgram experiments influenced them significantly (e.g., such that they believed the other person would not get any shocks); one participant who said they fell asleep during the task; one participant who said they enjoyed consensual pain as part of BDSM practices; and one participant who indicated the thresholding procedure (see below) was not completed. The final sample of N = 113 was 38% male with an average age of 22.25 years (SD = 3.41).

Procedure

Due to unexpected building closures, the study took place across three different laboratory locations in Oxford. Upon signing up for the study, eligible participants were sent a link for an online questionnaire, which they completed prior to their laboratory testing session and which assessed demographics as well as individual trait measures for psychopathy (using a short version of the Self-Report-Psychopathy scale, SRP-II; Williams & Paulhus, 2004), social desirability (Reynolds, 1982), empathy (using the Interpersonal Reactivity Index, IRI; Davis, 1983), trait optimism (using the Life Orientation Test, LOT-R; Scheier et al., 1994), entitlement (Campbell et al., 2004), moral identity (Aquino & Reed, 2002), and a brief personality disorder scale (Moran et al., 2003).

Participants were invited in pairs to the laboratory and arrived at staggered times, 20 minutes apart. Upon arrival, each participant was guided to a separate experimental room and asked to only speak out loud during designated times in order to ensure complete anonymity from the other

participant. After reading the participant information sheet and signing the informed consent form, a two-phase thresholding procedure took place (Vlaev et al., 2009). This procedure allowed to control for differences in skin resistance across participants and to establish participants' subjective pain levels, as well as giving them the opportunity to experience the shocks before the decision task. It should be noted that the electric shocks administered in this and later phases of the experiment were necessarily painful, but even the highest applied levels of electric stimulation were safe and did not cause any significant side effects, which has also been documented in the paradigm's extensive past use (Crockett et al., 2014, 2015, 2017; Vlaev et al., 2009; L. J. Volz et al., 2017).

At the beginning of the thresholding procedure, the experimenter attached an electrode to a prepared area on a participant's left wrist. Next, participants were presented with a visual scale from zero (described as "I can feel it, but it is not painful") to ten (described as "this is highly painful and I could not tolerate another shock at this level") and instructed to use the scale for rating each individual electric shock administered to them in the following first phase of the thresholding. The experimenter then started with a shock at a very low voltage and, after recording the participant's rating, incrementally increased the voltage for administering the next shock until the participant's level 10 was reached. Importantly, the experimenter followed a detailed script for instructing participants in this phase, which also told participants that it was important to determine how much pain they could tolerate and hence for them to push themselves really hard "but not to the point where you might pass out". To encourage participants to rate each shock independently, the experimenter occasionally reduced rather than increased the voltage, or did not change it at all. This initial thresholding procedure was repeated two or three times, allowing

participants to adapt to the shocks and to obtain more consistent ratings. Once a given participant's ratings converged, the second phase of the thresholding procedure was initiated. In this phase, participants received a series of 21 shocks that were obtained by fitting their current-to-rating curve using a sigmoid function and then determining seven shocks that ranged from 40% to 100% percent of their level 10 shock, which were each repeated three times. The order of those 21 shocks was randomized to prevent spillover-effects across pain ratings. Again, participants rated each shock they received. From those ratings, their "strong pain" (level 8) shock intensity was estimated and used at the end of the experiment to deliver outcomes based on participants' decisions in the main task.

The thresholding was followed by a role assignment procedure serving to highlight that another participant was indeed present and to allocate participants to their roles. In particular, participants learned the main task involved two roles, a 'Decider' and a 'Receiver' role, and that drawing differently coloured balls would determine their respective roles. To this end, both participants stood silently on opposite sides of a slightly opened door, first wiggling at each other with a rubber-gloved hand through the door gap and then reaching into a cardboard box to draw one of two differently coloured balls with which they were silently guided back to their respective experimental rooms. Here, participants were told that they drew the Decider ball. Importantly, while both participants were led to believe throughout the study that they were the Decider, both were Decider and Receiver at the same time. Because I was only interested in studying Decider choices, rather than Receiver perceptions about Decider choices, this procedure allowed me to maximise the number of

Deciders while avoiding deception by actually delivering Receiver shocks at the end of the study.

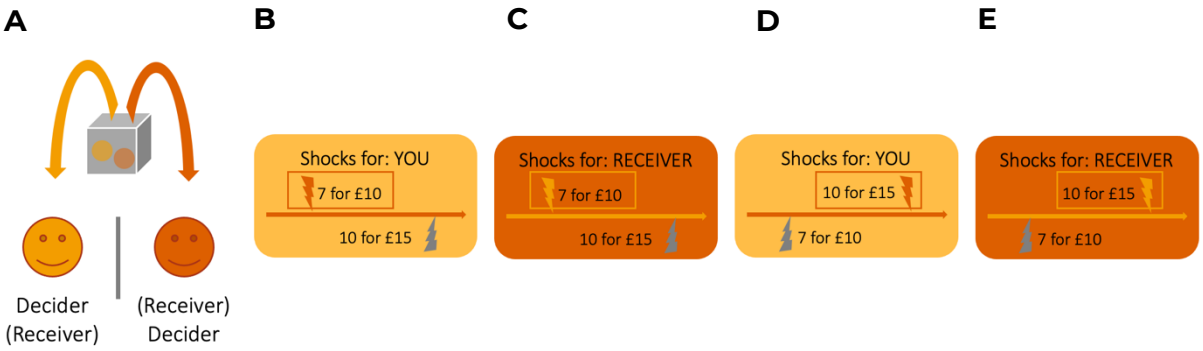


Figure 9. Main experimental task. In the role allocation procedure (A), both participants were assigned the Decider role. Unbeknownst to them, they also were the Receiver for the other participant. In the following task (B-E), they went through 160 rounds in which they chose between a lower number of electric shocks for a lower amount of money, or a higher number of electric shocks for a higher amount of money. Each trial had a highlighted default option (top) which was recorded as choice unless the participant pressed a key within six seconds, in which case the alternative option was highlighted and recorded as choice. Importantly, the money always was for the Decider whereas the shocks were for the Decider on half of rounds (B and D) and for the Receiver on the other half of rounds (C and E). For half of rounds, the default option offered less shocks and money than the alternative option (B and C), for the other half of rounds the default option offered more shocks and money than the alternative option (D and E).

Next, participants were instructed about the main task, the harm aversion paradigm (see *Figure 9*). In this paradigm, participants made a series of choices between a higher amount of money for a higher number of level-8-shocks, or a lower amount of money for lower number of level-8-shocks. There was a total of 160 rounds that were based on previous work (Crockett et al., 2014) and optimized such that shocks and money magnitudes were decorrelated across rounds while at the same time the shocks-money pairings represented the full range of harm aversion observed in previous work. Both shocks and money were limited to vary between 0 and 20 in terms of integers for shocks and in terms of numbers rounded to the nearest 10th for money (for more details see *Study 3* in *Appendix C*: Additional procedures and materials).

Importantly, while the money was always for the Decider, shocks were for the Decider on half of the rounds ("self-rounds") and for the Receiver on the other half of the rounds ("other-rounds"), as indicated by text instructions and different background colours. Each trial had a highlighted choice option that was selected unless participants pressed a key within six seconds. To ensure that choices were independent from each other and not influenced by habituation effects, no shocks or money were delivered during the main task, but one trial was randomly selected and implemented afterwards. These procedures were fully transparent to participants and I ensured they were understood by providing participants with eight practice rounds before starting the main task as well as comprehension questions after the task (see below).

Immediately before starting the main task, participants were presented with the key instructions of the between-subjects manipulation. In particular, those in the control condition only read "The Decision Task will start in 15 sec" while those in the high (low) impact uncertainty conditions read:

> "Please read this information carefully before we start the Decision Task. There have been many studies, conducted in our lab and elsewhere, investigating how people respond to electric shocks. These studies have shown that there is a **high degree of variability** (low uncertainty: **low degree of variability)** in how people experience the shocks. For some people, the experience is extremely unpleasant and for others the experience is not at all unpleasant (low uncertainty: In other words, most people find the experience similar). Because of this, you can NOT assume that the Receiver will experience the shocks in the same way that you do. (low uncertainty: Because of this, you can assume that the Receiver will experience the shocks in the same way that you do.) … The Decision Task will start in 15 sec."

After completing the main task, participants were asked about their perception, attitudes, and beliefs concerning various aspects of the decision task. Among other questions, I asked participants how much they would dislike

1/5/10/20 shocks for themselves (and for the Receiver, asked as separate question) and how much they thought the Receiver would dislike those shocks along with confidence ratings for those answers; how unpleasant they thought a level-8-shock would be for themselves (and for the Receiver, asked as separate question); to what extent they expected the Receiver's experience of shocks being similar to their own experience; how difficult they found the choices when the shocks were for themselves (and for the Receiver, asked as separate question); how selfish they felt about their choices on Receiver rounds; how much empathy they felt towards the Receiver; and what proportion of their task-based payment they were willing to donate for charity (see *Study 3* in *Appendix C*: Additional procedures and materials for the full list of questions with exact wordings).

The computer then randomly selected one round and displayed to the participant how much extra money they earned and how many electric shocks, if any, they would receive based on their choice in this round. At that point, the experimenter was called back into the room and launched a final online debriefing questionnaire. The questionnaire asked participants which role they had played during the experiment, how they decided during the decision task, whether money could persuade them to increase the number of shocks for the Receiver and themselves respectively, and similar questions about trading money against shocks for themselves and the Receiver. Furthermore, participants indicated whether they felt observed during the study; whether they had any doubts about the Receiver being present as well and/or the Receiver potentially receiving electric shocks as a result of their choices; whether they had heard about the Milgram experiments and if so, whether their behaviour in the main task had been influenced by this; what sex they thought

the other person was as well as why they thought so and whether their belief might have influenced their behaviour in the main task. Finally, the randomly chosen decision-round was implemented by administering the designated number of electric shocks at the participant's level-8 voltage and by paying out the designated monetary amount in addition to the showup fee.

Computational model

In addition to statistical analysis of participants' behavioural data, I used computational modelling to examine their decisions in the main task. Computational models allow to identify and quantify latent variables such as decisional uncertainty or subjective value from observable choices (Busemeyer & Diederich, 2014; Crockett, 2016; Smith & Ratcliff, 2004). The computational model I used was developed by Crockett and colleagues (2014) precisely to analyse choices in the harm aversion paradigm and has been shown to outperform alternative models across different variations of the paradigm (Crockett et al., 2014, 2015, 2017). It explains participants' choices by weighting differences in money and shocks by a free parameter κ, capturing harm aversion separately for self- and other-rounds:

$$\Delta V = (1 - \kappa)\Delta m - \kappa \Delta s$$

$$\kappa = \begin{cases} \kappa_{self} & \text{in self-rounds} \\ \kappa_{other} & \text{in other-rounds} \end{cases}$$

Hence, for each round, the model predicts a given participant's subjective value ΔV for switching from the default to the alternative option by scaling the differences in money Δm and shocks Δs between the two options with the harm

aversion parameter κ, which takes different values for self- and other-rounds. With κ capturing the subjective conversion-rate between harm and money, the parameter converges towards 0 for a minimally harm-averse decision-maker who would take any number of shocks for an additional sum of money. Meanwhile, κ converges towards 1 for a maximally harm-averse decision-maker who would forgo any additional amount of money to receive one less shock. The resulting trial-by-trial differences in subjective value are transformed into probabilities for choosing the harmful option using a softmax function that accounts for choice noisiness across decision-rounds by fitting an inverse temperature parameter, β:

$$p(harm) = \frac{1}{1 + e^{-\beta \Delta V}}$$

As β represents the slope of the function relating decision values to choices, higher values indicate a steeper slope and thus more deterministic and consistent choices. Meanwhile, as β approaches 0, the slope of the softmax function becomes less steep, indicating more noisy choices. Because manipulating impact uncertainty (i.e., uncertainty about others' shock experience) conceivably might affect decision uncertainty in other- but not self-rounds, I planned to fit a model with separate β parameters for self- and other-rounds and to compare it to a model with a single β parameter. To identify the parameter combination ($\kappa_{(self/other)}$, $\beta_{(self/other)}$) that maximises the likelihood of the total of observed choices for a given participant, I used nonlinear optimization implemented through MATLAB's fmincon function and confirmed the parameter's reliability with multiple random starts of optimization. To summarise, the described computational model provides a level of description that goes beyond the observed binary choices, by quantifying participants' harm

aversion and choice noisiness in parameters that are in turn comparable across experimental conditions.

Results

All analyses were conducted in *R* (R Core Team, 2020). I used robust analysis techniques implemented with the *WRS2* (Mair & Wilcox, 2020) and *MASS* (Venables & Ripley, 2002) packages to account for non-symmetric distributions of observations as appropriate. Assumptions of equal variances across conditions were not violated. Figures were generated using the packages *ggstatsplot* (Patil, 2018) and *ggplot2* (Wickham, 2016). To facilitate readability, I will refer to (high/low) impact uncertainty as (high/low) uncertainty in the following.

Manipulation check. To verify the effectiveness of the uncertainty manipulation, I first compared participants' ratings about the similarity of their and the Receiver's shock experience across conditions. Since participants in the *low/high uncertainty* condition were told that there is low/high variability in how people experience shocks and that they hence could/could not assume the Receiver would experience the shocks in the same way, I expected higher similarity ratings in the low relative to the high uncertainty group. This was the case at the descriptive level (see *Figure 10*), but the overall effect of condition was not significant ($F_{trimmed}(2, 59.17) = 2.38$, $p = .101$, $\xi = 0.27$, 95% CI [0.07, 0.45]), nor was the pairwise comparison between the low and high uncertainty group when all three possible post-hoc comparisons were carried out and corrected for multiple testing ($p = .116$). However, focusing only on the two treatment groups and ignoring the control group, suggested that the manipulation did affect similarity ratings in the expected direction, $t_{yuen}(60.68) = 2.11$, $p = .039$, $\delta_t = 0.47$, 95% CI [0.11, 1.08].

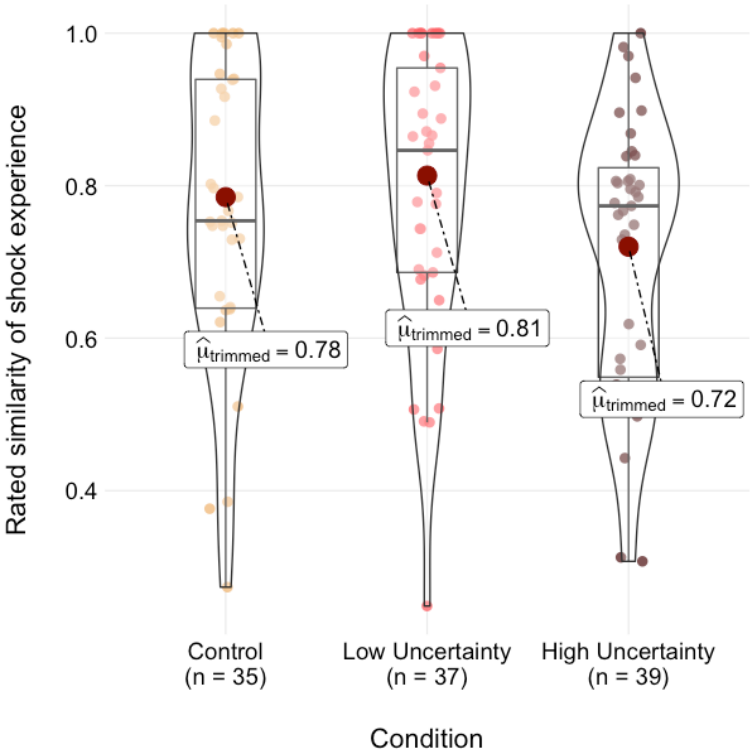


Figure 10. Rated similarity of shock experience. Participants answered the questions "To what extent do you think the Receiver's experience of the shocks is SIMILAR to your own experience of the shocks" on a continuous rating-scale from "not at all similar" to "very similar". Two ratings were missing for the low uncertainty group.

Absolute differences in unpleasantness ratings of a level-8-shock for other versus self were larger in the high uncertainty group (|unpleasantness8$_{other}$ - unpleasantness8$_{self}$|: $\hat{\mu}_{trimmed}$ = 0.13) relative to the low uncertainty ($\hat{\mu}_{trimmed}$ = 0.08) and control ($\hat{\mu}_{trimmed}$ = 0.08) groups, but only at the descriptive level, p = .390. Meanwhile, absolute differences in confidence ratings for these unpleasantness ratings varied significantly across conditions ($F_{trimmed}(2, 57.37) = 3.79$, p = .029, ξ = 0.30, 95% CI [0.07, 0.50]), driven by larger absolute differences for the high relative to the low uncertainty group, $\hat{\Psi}$ = 0.13 p = .033 (other contrasts ps ≥ .189). A follow-up analysis confirmed this effect went into the expected direction, namely the degree to which participants felt less confident about the unpleasantness of a level-8-shock for other than self (confidence8$_{other}$ - confidence8$_{self}$) was larger in

the high uncertainty group ($\hat{\mu}_{trimmed}$ = -0.20) relative to the low uncertainty group ($\hat{\mu}_{trimmed}$ = -0.11). Together, these results suggest that participants in the high relative to the low uncertainty group explicitly rated the shock experience for self and other as less similar and felt less confident when estimating shock unpleasantness for others. This indicates that the impact uncertainty manipulation affected the perceived similarity of shock experiences and confidence about shock experiences in the expected direction, although neither treatment condition differed much from the control condition and effects were overall small, in some cases only at trend-level.

Behavioural analysis

Shock unpleasantness. On a continuous rating scale from "0 – not at all unpleasant" to "1 – very unpleasant", participants rated a level-8-shock as more unpleasant for other ($\hat{\mu}_{trimmed}$ = 0.80) than self, $\hat{\mu}_{trimmed}$ = 0.71; $t_{paired\ yuen}(66) = -5.60$, $p < .001$, $\delta_t = 0.27$, 95% CI [-0.10, -0.05]. A robust two-way within-between-subjects ANOVA confirmed that this main effect of a self-other-difference remained significant ($p < .001$) when controlling for the effect of condition ($p = .008$) and was not qualified by an interaction with condition, $p = .896$. At the descriptive level, participants in the high uncertainty group reported lower unpleasantness ratings for a level-8-shock delivered to themselves ($\hat{\mu}_{trimmed}$ = 0.63) relative to the low uncertainty ($\hat{\mu}_{trimmed}$ = 0.75) and control ($\hat{\mu}_{trimmed}$ = 0.74) groups, $F_{trimmed}(2, 58.75) = 2.28$, $p = .112$, $\xi = 0.27$, 95% CI [0.08, 0.52]. Similarly, participants in the high uncertainty group reported lower unpleasantness ratings for a level-8-shock delivered to the Receiver ($\hat{\mu}_{trimmed}$ = 0.73) relative to the low uncertainty ($\hat{\mu}_{trimmed}$ = 0.83; $p_{pairwise} = .037$) and control ($\hat{\mu}_{trimmed}$ = 0.82; $p_{pairwise} = .037$) groups, while the latter two groups did not differ ($p = .782$), $F_{trimmed}(2, 58.75) = 3.53$, $p = .036$, $\xi = 0.33$, 95% CI [0.14, 0.52].

Harmful choices. I used a generalized regression model with fdr-correction for multiple comparisons to test whether harmful choices (defaulting to or choosing the option associated with more shocks) differed across conditions and choice-target (self versus other), and whether these two predictors interacted. Significant main effects for condition ($\chi^2(2) = 60.65$, $p < .001$) and choice-target ($\chi^2(1) = 9.73$, $p = .002$) were qualified by an interaction between the two predictors, $\chi^2(2) = 7.47$, $p = .024$. In particular, aligning with hyperaltruistic valuation of pain observed in previous work (Crockett et al., 2014, 2015; L. J. Volz et al., 2017), participants chose the harmful option less often when the shocks were for other than self, $b = -0.18$, $SE = 0.06$, $z = -3.12$, $p = .002$, OR = 0.84, 95% CI [0.75, 0.94]. Relative to the control condition, participants in the low uncertainty condition chose the harmful option less often ($b = -0.18$, $SE = 0.05$, $z = -3.25$, $p = .001$, OR = 0.84, 95% CI [0.75, 0.93]) whereas participants in the high uncertainty condition chose the harmful option more often both relative to the control ($b = 0.24$, $SE = 0.06$, $z = 4.31$, $p < .001$, OR = 1.28, 95% CI [1.14, 1.43]) and low uncertainty groups ($b = 0.42$, $SE = 0.05$, $z = 7.74$, $p < .001$, OR = 1.52, 95% CI [1.37, 1.70]).

Importantly, these main effects were qualified by an interaction whereby, relative to the control group, self-other-differences in harmful choices were larger under high uncertainty ($b = -0.21$, $SE = 0.08$, $z = -2.73$, $p = .006$, OR = 0.81, 95% CI [0.69, 0.94]) and at trend-level for the low uncertainty group, $b = -0.11$, $SE = 0.08$, $z = -1.41$, $p = .157$, OR = 0.90, 95% CI [0.77, 1.04]. Even if shifted because of the main effect for condition, self-other-differences in harmful choices were comparable between the low and high uncertainty groups, $b = -0.11$, $SE = 0.08$, $z = -1.41$, $p = .159$, OR = 0.90, 95% CI [0.78, 1.04]. On self-rounds, participants in the high uncertainty group chose the harmful option more often (73% of the time) relative to the low uncertainty (63%) and control groups (67%; see *Figure 11*).

Meanwhile, on other-rounds, they chose the harmful option as often (64%) as participants in the control condition (63%) and only slightly more often than those in the low uncertainty group (57%; see *Figure 11*). These results were robust to including various task parameters as controls, as detailed in *Table 4*.

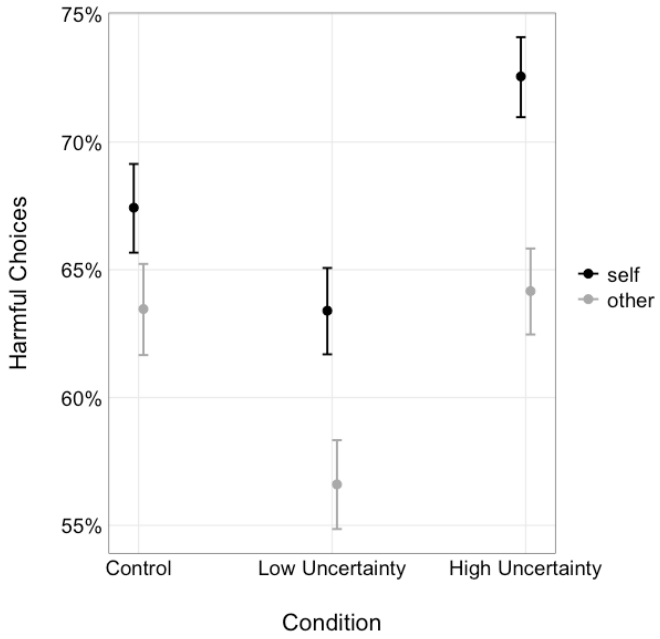


Figure 11. Predicted values for harmful choices predicted by condition x choice-target (i.e., self versus other) interaction in generalized regression model of Study 3. Error bars represent 95% confidence intervals.

Number of shocks. Robust regression analysis indicated that the total number of shocks assigned differed across conditions ($F_{trimmed}(2) = 29.92$, $p < .001$) and for choice-target, $F_{trimmed}(1) = 34.70$, $p < .001$. These main effects were not qualified by an interaction, $p = .661$. In particular, participants overall assigned fewer shocks to other relative to self, $b = -0.50$, $SE = 0.18$, $t = -2.72$, $p = .006$. Participants in the high uncertainty condition assigned more shocks relative to both the control ($b = 0.59$, $SE = 0.18$, $t = 3.26$, $p = .001$) and low uncertainty conditions, $b = 1.03$, $SE = 0.18$, $t = 5.91$, $p < .001$. Participants in the low uncertainty condition assigned fewer shocks relative to the control condition, $b = -0.45$, $SE = 0.18$, $t = -2.48$, $p = .013$.

Table 4. Generalized regression model predicting harmful choices (Study 3)

Variables	*b* (*SE*)	*p*	OR	95% CI_{OR}
(Intercept)	2.00 (0.10)			
Unpleasantness (level-8 shock) for self	0.06 (0.12)	.652	1.06	[0.83, 1.35]
Unpleasantness (level-8 shock) for other	-1.37 (0.16)	< .001	0.25	[0.18, 0.35]
*Δ*shocks	0.00 (0.00)	.274	1.00	[0.99, 1.00]
*Δ*money	0.00 (0.00)	.841	1.00	[1.00, 1.00]
max(shocks)	-0.05 (0.00)	< .001	0.95	[0.95, 0.96]
Other	-0.19 (0.06)	.001	0.83	[0.74, 0.93]
Low uncertainty	-0.18 (0.06)	.002	0.83	[0.75, 0.93]
High uncertainty	0.15 (0.06)	.011	1.16	[1.03, 1.30]
*Δ*shocks x *Δ*money	0.01 (0.00)	< .001	1.01	[1.01, 1.01]
Other x *Δ*money	0.00 (0.00)	.822	1.00	[0.99, 1.01]
Other x low uncertainty	-0.12 (0.08)	.137	0.89	[0.76, 1.04]
Other x high uncertainty	-0.23 (0.08)	.005	0.80	[0.68, 0.93]

Note. Estimates from generalized regression model. '*Δ*shocks' denotes the difference in number of shocks associated with the default and the alternative option respectively. '*Δ*money' denotes the difference in monetary payoffs between the default and the alternative option respectively. 'max(shocks)' denotes the maximum number of shocks in a given round. 'Other' denotes the choice-target (self versus other).

Response times. Response times differed across conditions ($F(2) = 6.26$, $p = .002$) and choice-target ($F(1) = 10.22$, $p = .001$) even though these main effects were qualified by an interaction, $F(2) = 2.98$, $p = .050$. Overall, participants were slower to respond on rounds with shocks for other than shocks for self, $b = 132.93$, $SE = 37.60$, $t = 3.54$, $p < .001$. This slowing effect (rt_{other} - rt_{self}) was, however,

significantly diminished for the low uncertainty (M = 5.70) relative to the control group (M = 132.93; other x low uncertainty: b = -127.23, SE = 52.12, t = -2.44, p = .015), while it was not qualified in the high uncertainty group (M = 68.07; p = .211). These effects were largely robust to controlling for other task parameters (see *Table 5*).

Table 5. Regression model predicting response times (Study 3)

Variables	b (SE)		p	η^2	95% CI
(Intercept)	1847.89 (36.80)				
Δshocks	10.29 (1.71)	<	.001	0.001	[0.001, 0.005]
Δmoney	-34.61 (1.75)	<	.001	0.049	[0.041, 0.058]
max(shocks)	23.87 1.96	<	.001	0.016	[0.012, 0.022]
Other	124.38 36.36		.001	0.001	[0.000, 0.002]
Condition				0.002	[0.000, 0.004]
Low uncertainty	-10.57 35.41		.766		
High uncertainty	-56.75 35.41		.109		
Δshocks x Δmoney	-0.34 0.18		.056	0.000	[0.000, 0.002]
Other x Δmoney	4.52 2.02		.025	0.001	[0.000, 0.002]
Target x condition				0.001	[0.000, 0.002]
Other x low uncertainty	-137.11 50.33		.006		
Other x high uncertainty	-85.81 50.03		.086		

Note. Estimates from regression model. 'Δshocks' denotes the difference in number of shocks associated with the default and the alternative option respectively. 'Δmoney' denotes the difference in monetary payoffs between the default and the alternative option respectively. 'max(shocks)' denotes the maximum number of shocks in a given round. 'Other' denotes the choice-target (self versus other).

Computational modelling

To further quantify and compare differences in harm aversion but also choice noisiness across choice-target and conditions, I estimated the computational models with single and separate noise (β) parameter(s) described above. To compare the models with single and separate β parameters, I computed participant-wise Bayesian Information Criterion (BIC) scores for each model. The BIC integrates a model's error term (i.e., the negative log-likelihood) with the number of rounds and parameters, such that it penalises models with more parameters. Group BIC scores, obtained by summing up the individual scores, indicated that the simpler model with a single β parameter described a better fit to the data with a BIC of 10,143 compared to a BIC of 10,341 for the model with separate β parameters. Thus, the following analyses are based on the on the simpler model with a single β parameter, which correctly predicted 72% of participants' choices, 95% CI [0.69, 0.75].

Harm aversion. Participants in the high uncertainty group were less averse to harm for themselves (κ_{self}: $\hat{\mu}_{trimmed}$ = 0.17) relative to the control (κ_{self}: $\hat{\mu}_{trimmed}$ = 0.22) and low uncertainty (κ_{self}: $\hat{\mu}_{trimmed}$ = 0.27) groups at trend-level, $F_{trimmed}(2, 60.29) = 2.57$, $p = .085$, $\xi = 0.33$, 95% CI [0.08, 0.53]. Meanwhile, there was no evidence for differences in harm aversion for other, $p = .256$. Consistent with participants making less harmful choices for others, they placed a higher cost on shocks to others than shocks to self as indicated by higher κ_{other} ($\hat{\mu}_{trimmed}$ = 0.28) relative to κ_{self} ($\hat{\mu}_{trimmed}$ = 0.22), $t_{paired\ yuen}(90) = -4.25$, $p < .001$, $\delta_t = -0.43$, 95% CI [-0.61, -0.26]. This hyperaltruistic pattern persisted across conditions (see *Figure 12* which includes statistics on self-other differences in harm aversion for each respective condition). Even though the effect size for hyperaltruism was smaller in the high uncertainty group, robust regression analysis indicated no significant

differences across conditions, p = .612. This result was robust to controlling for demographics, empathy towards the Receiver, as well as trait differences in empathy, psychopathy, and entitlement.

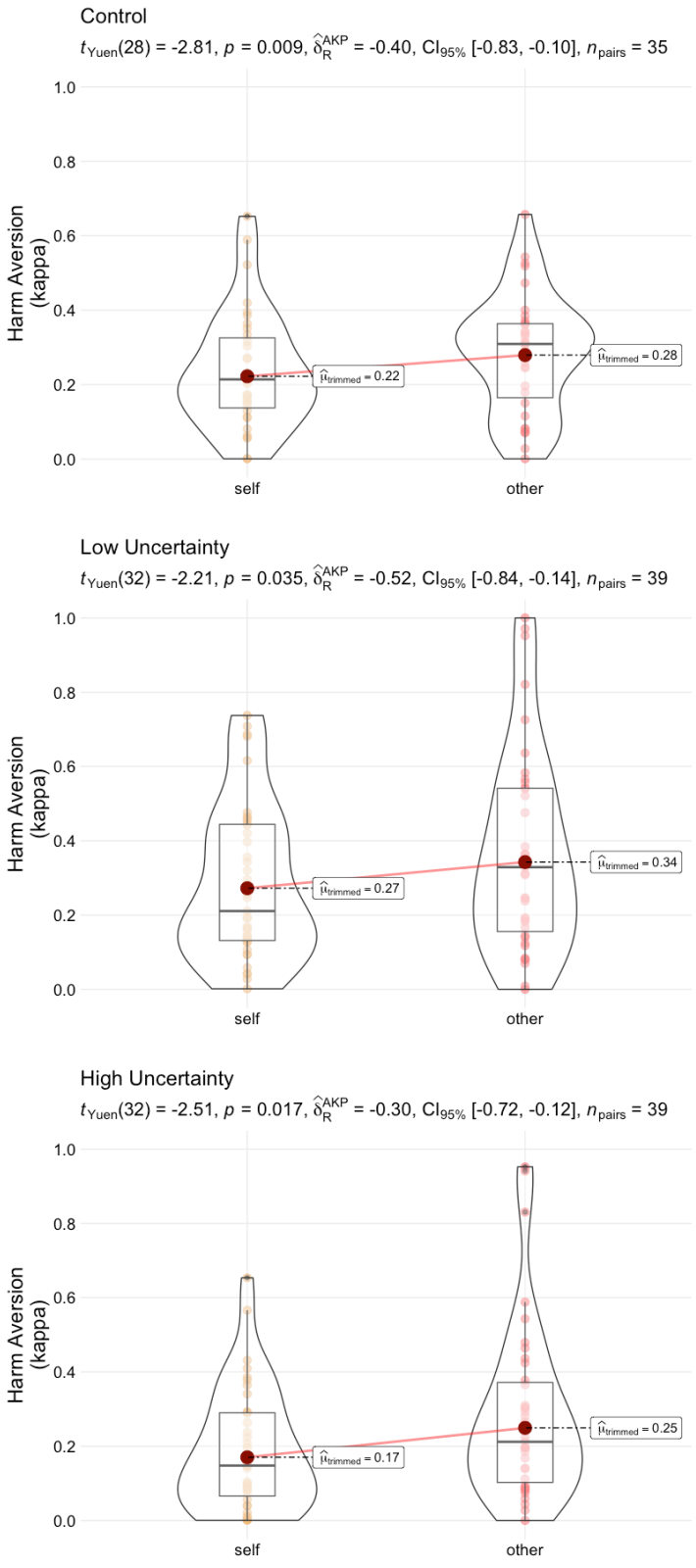


Figure 12. Distributions of and robust statistical tests for differences in harm aversion (κ) for self versus other across the respective conditions. Higher κ values indicate stronger harm aversion. Trimmed means indicated by red annotated dots.

The degree by which participants were hyperaltruistic (κ_{other} - κ_{self}) was correlated with the degree of slowing for others, rt_{other} - rt_{self}; ρ_{pb} = .27, p = .003, 95% CI [.04, .49]. In particular, this association was present in the control (ρ_{pb} = .54, p < .001, 95% CI [.26, .75]) and high uncertainty groups (ρ_{pb} = .31, p = .050, 95% CI [- .08, .64]), but absent in the low uncertainty group (ρ_{pb} = .01, p = .969, 95% CI [- .41, .40]), aligning with the diminished slowing effect observed for the latter group. Furthermore, hyperaltruism was correlated with the degree to which participants explicitly rated the other's shock experience as more unpleasant than their own, ρ_{pb} = .35, p < .001, 95% CI [.17, .49]. This association was particularly pronounced in the high uncertainty group (ρ_{pb} = .47, p = .003, 95% CI [.19, .67]), but a similar trend also appeared in the control (ρ_{pb} = .33, p = .055, 95% CI [- .02, .62]) and low uncertainty groups, ρ_{pb} = .19, p = .254, 95% CI [- .09, .53].

In line with previous work suggesting that harm aversion for others may be anchored in harm aversion for self (Crockett et al., 2014; Epley et al., 2004), κ_{other} and κ_{self} were substantially correlated, ρ_{pb} = .65, p < .001, 95% CI [.50, .77]. At the descriptive level, this correlation was reduced in the high uncertainty (ρ_{pb} = .47, p = .002, 95% CI [.08, .75]) relative to the low uncertainty (ρ_{pb} = .75, p < .001, 95% CI [.56, .89]; 95% $CI_{difference}$[- .10, .65]) and control (ρ_{pb} = .73, p < .001, 95% CI [.51, .86]; 95% $CI_{difference}$[- .09, .65]) groups.

Choice noisiness. Because observations for the inverse temperature parameter β were right-skewed, I used log-transformed values to analyse choice noisiness. I observed a trend for choices of participants in the low uncertainty group being less noisy (log(β): $\hat{\mu}_{trimmed}$ = 2.24) than choices in the high uncertainty (log(β): $\hat{\mu}_{trimmed}$ = 1.81) and control (log(β): $\hat{\mu}_{trimmed}$ = 1.34) groups, $F_{trimmed}$(2, 61.32) = 2.55, p = .087. Participants with more deterministic choices were less harm-averse in choices for self, ρ_{pb} = - .37, p < .001, 95% CI [- .56, - .20]. The association

between choice noisiness and harm aversion was not significant for κ_{other} (ρ_{pb} = -.17, p = .064, 95% CI [- .40, .02]), and there was no association between choice noisiness and hyperaltruism ($\kappa_{other} - \kappa_{self}$: ρ_{pb} = .03, p = .763, 95% CI [- .17, .22]).

To further examine the relationship between choice noisiness and harm aversion across self and other decisions, and to explore the hypothesis that manipulating impact uncertainty about others' shock experiences might affect uncertainty in choices for other but not in choices for self, I conducted an exploratory analysis with separate β parameters for other and self obtained in the more complex model. I did not observe significant differences across conditions for log(β_{self}), log(β_{other}), or their difference score (ps ≥ .166). At the descriptive level, choices for others were noisier (log(β_{other}): $\hat{\mu}_{trimmed}$ = 1.80) than choices for self, log(β_{self}): $\hat{\mu}_{trimmed}$ = 2.14; $t_{paired\ yuen}(90)$ = 1.58, p = .119. Looking at the separate experimental groups, this difference was significant only in the high uncertainty group ($\Delta\hat{\mu}_{trimmed}$ = 0.76; $t_{paired\ yuen}(32)$ = 2.31, p = .027, δ_t = 0.26, 95% CI [0.09, 1.42]) but not in the control ($\Delta\hat{\mu}_{trimmed}$ = 0.18; $t_{paired\ yuen}(28)$ = 0.41, p = .683) and low uncertainty groups ($\Delta\hat{\mu}_{trimmed}$ = 0.04; $t_{paired\ yuen}(32)$ = 0.13, p = .898). Robust regression analysis confirmed this pattern but the interaction between 'condition' and 'choice-target' was not significant (ps > .293).

Participants who were less noisy in their choices for others (higher log(β_{other})) were less averse to harming others (lower κ_{other}), ρ_{pb} = - .20, p = .036, 95% CI [- .42, - .02]. Again, this relationship was strongest in the high uncertainty (ρ_{pb} = - .53, p < .001, 95% CI [- .73, - .24]) and control groups (ρ_{pb} = - .26, p = .033, 95% CI [- .64, - .03]), whereas it was not present in the low uncertainty group, ρ_{pb} = .08, p = .643, 95% CI [- .25, .39]. Meanwhile, less noisy choices for self were associated with less harm aversion for self across all conditions (ρs_{pb} ≤ - .37, ps ≤ .035). The degree to which participants were noisier in choices for other than self (log(β_{self})-

log(β_{other})) was associated with the degree of hyperaltruism (κ_{other} - κ_{self}) at trend-level, ρ_{pb} = .14, p = .151, 95% CI [- .06, .32]. This trend was strongest for the control group (ρ_{pb} = .33, p = .055, 95% CI [- .05, .64]), followed by the high uncertainty (ρ_{pb} = .16, p = .321, 95% CI [- .14, .51]) and low uncertainty (ρ_{pb} = - .02, p = .886, 95% CI [- .38, .30]) groups respectively.

Individual differences and control variables

Willingness to pay & charitable donation. Participants' willingness to pay (WTP) for receiving one less shock was lower for self ($\hat{\mu}_{trimmed}$ = 0.46) than for other ($\hat{\mu}_{trimmed}$ = 0.78), $t_{paired\ yuen}$ (87) = -3.40, p = .001, δ_t = -0.26, 95% CI [-0.41, -0.14]. The degree by which participants were more willing to pay for reducing the other's versus their own shocks (WTP_{other} - WTP_{self}) did not differ across conditions, p = .573. Higher WTP for reducing the other's shocks was positively correlated with harm aversion for other (ρ_{pb} = .50, p < .001, 95% CI [.33, .64]) and hyperaltruism (ρ_{pb} = .48, p < .001, 95% CI [.30, .63] and these associations did not vary across conditions (ps ≥ .157). Similarly, participants' WTP for reducing their own shocks was positively correlated with harm aversion for self, ρ_{pb} = .55, p < .001, 95% CI [.40, .68] and this association did not vary across conditions (ps ≥ .174).

Most participants (65%) did make a charitable donation that was on average 26.23% of their task-based payment on average. These rates did not differ across conditions (p = .600). Charitable donations were correlated with the degree to which participants were more averse to harm for other than self (κ_{other} - κ_{self}: ρ_{pb} = .43, p < .001, 95% CI [.23, .60]) and with harm aversion for other (κ_{other}: ρ_{pb} = .43, p < .001, 95% CI [.27, .58]), but not with harm aversion for self (κ_{self}: ρ_{pb} = .19, p = .050, 95% CI [.02, .36]).

Empathy. Participants overall felt empathic towards the Receiver ($\hat{\mu}_{trimmed}$ = 0.67, 95% CI [0.64, 0.74]) and there were no differences across conditions (p = .398). Those who felt more empathic towards the Receiver were more averse to harming them in absolute terms (κ_{other}: ρ_{pb} = .40, p < .001, 95% CI [.23, .55]) as well as relative to harm aversion for themselves, κ_{other} - κ_{self}: ρ_{pb} = .38, p < .001, 95% CI [.20, .53]. Notably, these associations were not significant in the high uncertainty group, κ_{other}: ρ_{pb} = .22, p = .171; κ_{other} - κ_{self}: ρ_{pb} = .21, p = .203. Meanwhile, higher trait empathy measured by the IRI's perspective taking subscale was associated with more harm aversion for other in the high uncertainty group both in absolute (κ_{other}: ρ_{pb} = .37, p = .028, 95% CI [.07, .63]) and relative terms (κ_{other} - κ_{self}: ρ_{pb} = .53, p = .001, 95% CI [.26, .72]), but not in the low uncertainty (ρ_{pb} ≤ .23, ps ≥ .162) and control (ρ_{pb} ≤ - .12, ps ≥ .085) groups.

Other individual differences. Participants who felt more entitled were less harm-averse for others (κ_{other}: ρ_{pb} = - .23, p = .016, 95% CI [- .41, - .07]) and less hyperaltruistic, κ_{other} - κ_{self}: ρ_{pb} = - .30, p = .002, 95% CI [- .45, - .12]. This association was significantly stronger in the high uncertainty group (ρ_{pb} = - .58, p < .001, 95% CI [- .78, - .31]) relative to the control (ρ_{pb} = - .10, p = .571, 95% CI [- .85, - .04]; 95% $CI_{difference}$[- .09, .65]) and low uncertainty (ρ_{pb} = - .23, p = .167, 95% CI [- .52, .08]; 95% $CI_{difference}$[- .74, .00]) groups. Individual differences in subclinical psychopathic traits were negatively associated with harm aversion for other (κ_{other}: ρ_{pb} = - .21, p = .028, 95% CI [- .38, - .02]), and at trend-level with harm aversion for self (κ_{self}: ρ_{pb} = - .17, p = .080, 95% CI [- .35, .05]) and hyperaltruism, κ_{other} - κ_{self}: ρ_{pb} = - .18, p = .068, 95% CI [- .36, .00]. These associations did not differ across conditions (ps ≥ .290). Relative to female participants, male participants showed less harm aversion for self (t_{yuen}(59.48) = 2.16, p = .035, δ_t = -0.43, 95% CI [-0.82, -0.04]) and at trend-level for others (t_{yuen}(53.70) = 1.80, p = .077, δ_t = -0.37, 95% CI [-0.78, 0.05]), while

hyperaltruism did not differ across male and female participants (p = .550). Differences in social desirability, trait optimism, or income were not associated with harm aversion or hyperaltruism (ps ≥ .184).

Discussion

Using a harm aversion paradigm, which allows quantifying how much people value harm for others versus self, I tested the effect of explicit high (versus explicit low versus non-explicit) impact uncertainty in decisions for others. While hyperaltruism was not modulated by the experimental manipulation, larger self-other differences in reaction times and choice noisiness under high uncertainty point towards a modulating effect of impact uncertainty in choices for others. In line with the intended effect of the manipulation, participants in the high (low) uncertainty group perceived the Receiver as less (more) similar to themselves and felt more (less) uncertain about their rating of how unpleasant the Receiver would experience a level-8-shock. Relative to the low uncertainty and control groups, the correlation between harm aversion for self and harm aversion for other was diminished in the high uncertainty group, conforming to the instructions which told participants they could not expect the Receiver to have a similar shock experience as themselves.

Exploratory analyses drawing upon a computational model estimating separate parameters of decision noisiness for self and other indicated that choices for others were significantly noisier than choices for self in the high uncertainty group but not in the low uncertainty and control groups. As higher choice noisiness indicates less confidence (T. Singer, Seymour, O'doherty, et al., 2004), these exploratory findings suggest that participants felt less sure about their choices under high impact uncertainty. In view of these effects being small

in size and in some cases only apparent at trend-level, evidence for the effectiveness of the uncertainty manipulation remains inconclusive.

Hyperaltruism was present across all conditions and was not modulated by the impact uncertainty manipulation. Meanwhile, several differences in harm aversion across treatment groups emerged at the behavioural as well as computational levels of analysis: I observed higher rates of harmful choices in the high uncertainty relative to the other two groups. Notably, these seemed to be driven by more harmful choices for self under high uncertainty, whereas harmful choices for other were similar across the high uncertainty and the control groups. Consistent with this observation, modelling estimates of harm aversion for self tended to be lower in the high uncertainty group relative to the low uncertainty and control groups. Supporting a relationship between impact uncertainty and (hyper)altruism, I observed that participants who were noisier in choices for other displayed more harm aversion for other – especially so, if they were in the high uncertainty or control group.

While the stronger relationship between decision noisiness and (hyperaltruistic) harm aversion for others under high impact uncertainty supports a relationship between uncertainty and hyperaltruism, the experimental treatment did not directly affect hyperaltruism. If anything, high impact uncertainty seems to have shifted harm aversion for self and other towards a higher pain tolerance. I had expected the impact uncertainty manipulation to affect harm aversion for other, but not for self, in line with the theoretical motivation and instrumentalization which pointed to variability in others' shock experience. However, the summarised observations suggest that highlighting impact uncertainty this way might have shifted representations of shock aversiveness for self, and to a slightly smaller extent for other, towards a

higher tolerance. This interpretation is supported by lower unpleasantness ratings for a level-8-shock for self (and to a smaller extent for other) in the high uncertainty group relative to the control and low uncertainty groups. It is noteworthy that, under high uncertainty, lower reported shock-unpleasantness for self translated into less harm aversion and more harmful choices for self. In contrast, lower reported shock-unpleasantness for other did not translate into less harm aversion for others, nor into more harmful choices for others under high uncertainty. This dissociation was also signified by a reduced relationship between harm aversion for self and for other in the high uncertainty group relative to the control and low uncertainty groups.

In hindsight, one explanation for this unexpected effect of the (high) impact uncertainty manipulation on experienced shock unpleasantness might lie in the instructions' focus on variability and similarity of people's shock experience, with uncertainty about possible impacts on the counterpart's well-being remaining peripheral. Rather than prompting uncertainty about their decision's impact on the Receiver in particular, this operationalisation might have actuated uncertainty about the range of possible shock experiences more generally and including participants' own ranking within this distribution. In combination with research suggesting that impact uncertainty is only effective when it is salient (Kappes et al., 2018, Study 7) and when it appeals to impacts on others' well-being that are relevant to the specific decision (Kappes et al., 2018, Studies 4 and 5) – findings that were established after this study was devised and data collected –, it seems plausible that the manipulation did not work as intended. Furthermore, results on individual differences suggested that the high impact uncertainty manipulation might have had a stronger effect on people with a general tendency to spontaneously adopt others' viewpoints (i.e., those

scoring high on trait perspective taking) and those feeling less deserving and entitled to being better off than others (i.e., those scoring low on trait entitlement).

As this latter set of findings demonstrates, more remains to be learned about the mechanisms and boundary conditions of impact uncertainty. As insights about impact uncertainty's mechanisms were even less defined at the outset of this project than at present, it might have been recommendable to use a stronger manipulation of impact uncertainty in the spirit of a hammer rather than a nudge approach. For instance, presenting the manipulation instructions for 15 seconds might have impeded demand effects as planned, but it might have been too brief to significantly boost impact uncertainty's salience as intended. Based on work that we conducted in the meantime (Kappes et al., 2018), I would also contend that, because impact uncertainty centers around subjective utilities over possible states of the world, manipulations should appeal to subjective utilities. The present impact uncertainty manipulation did briefly touch upon those in mentioning that "For some people, the experience is extremely unpleasant and for others the experience is not at all unpleasant", but the instructions put more emphasis – even literally – on the underlying possible states of the world in stating that "there is a **high degree of variability** in how people experience these shocks". A manipulation check asking participants more directly about how (un)certain they felt about the other's shock experience might have provided more explicit and interpretable insights about the manipulation's effectiveness.

Furthermore, effective manipulations of impact uncertainty have specified the odds of different possible states of the world, usually drawing upon maximally ambiguous chances (Kappes et al., 2018). In the context of the harm

aversion paradigm, this could be accomplished by telling participants there is a 50% chance of being paired with a Receiver who experiences the shocks as extremely painful and unpleasant, and a 50% chance of being paired with a Receiver who experiences the shocks as not painful or unpleasant at all. Indeed, exploring whether and when impact uncertainty's effect prevails with unknown probabilities (and how robust it is across different probability levels) in a systematic way might be an interesting avenue for future research, even though avoiding deception would become challenging within the harm aversion paradigm. While I think that the above conceptual points concerning the impact uncertainty manipulation provide the focal limitations of the present study and as such contributed to a somewhat heterogeneous and ambiguous set of results, I would also aim for a larger sample size in follow-up work; not at least to achieve the necessary power for examining potential qualifying conditions of the correlational results reported here.

Finally, behavioural and computational indices of a relationship between uncertainty and harm aversion for others (i.e., self-other differences in response time, noisiness, and unpleasantness ratings that relate to hyperaltruism) were more similar to the control condition under high uncertainty than under low uncertainty. So even when uncertainty about others' experience was not explicitly highlighted (as in the control condition), people behaved as if it was. This could, in principle, imply ceiling effects concealing impact uncertainty's role in participant's choices; but given that decisions under low uncertainty rarely differed significantly from the other two, I would expect such ceiling effects to be minor if present at all.

To summarise, the present findings support a robust tendency of being more averse to harming others than oneself. This precautionary pattern likely

requires a multifactorial explanation, in which enhanced uncertainty about others' experience plays a supporting – but not main – role that seems to interact with other factors such as individual differences in empathy or self-serving motives (e.g., a superior tolerance for pain relative to others). It will be an important avenue for future research to further characterise impact uncertainty per se, its influence on hyperaltruism, and its relationship with other determinants of social behaviour.

As we increasingly interact not only with other human beings but also with artificially intelligent agents, the question emerges how uncertainty about their inner workings shapes our interactions with them. For instance, Slater and colleagues (2006) have demonstrated that people seem to treat virtual agents similar to human agents in a setting requiring them to give electric shocks to their artificial counterparts. While the paradigm deployed in this study focused on obedience (following Milgram, 1974), rather than relative harm aversion for self and (an artificial) other, the study highlights that people may treat artificial agents like other social agents. The following chapter will thus seek to explore how humans respond to uncertainty in a broader category of social interactions including artificial agents.

4 Uncertainty in human-machine interactions

We increasingly create spaces where we interact with artificial rather than human decision-makers. This development raises the question how uncertainty shapes those interactions. Take, for instance, the domain of medical diagnosing, which has traditionally been fielded by highly educated medical doctors who are often entrusted with no less than a human life. Where most human medical doctors struggle to identify and predict complex medical conditions from people's health records (e.g., schizophrenia or cancer), a readily available algorithm, named 'deep patient', has strived and indeed performed at superhuman levels (Miotto et al., 2016). However, these high-stakes gains in terms of performance usually come at a high-stakes cost: the most powerful machines available to date, among them deep patient, rely on black-box models that are often inscrutable not only to end-users but even to their designers (Adadi & Berrada, 2018; Gunning et al., 2019; Murdoch et al., 2019; Waldrop, 2019). Also in the realm of medical decisions, uncertainty about machines' inner workings is viewed as a major obstacle to the widespread adaptation of tools such as deep patient (Hardesty, 2016; Price, 2018).

The dilemma whereby machine performance comes at the cost of uncertainty about its inner workings (Gunning et al., 2019) has evolved from "the dark secret at the heart of AI" (W. Knight, 2017) to a fiery topic absorbing developers, politicians, practitioners, and academics alike. It is usually referred to as the problem of "*interpretable AI*" – sometimes also "explainable AI", "intelligible AI", or "transparent AI" (Lipton, 2017; Mittelstadt et al., 2019; Rudin, 2019) – and hence I will use this terminology in the following to refer to the

flipside of uncertainty about machines' inner workings. Currently available AI technologies leverage a range of methods to make predictions, from simple linear regression models to highly complex deep learning models. Simpler models are generally more interpretable, in that it is straightforward to understand why and how the AI arrives at its decisions (Murdoch et al., 2019). For example, an AI system relying on linear regression might predict a health outcome from a limited set of variables (e.g., age, weight, gender) in a way that is easy to explain in simple language. More complex AI systems, such as those relying on deep learning, can be difficult or even impossible to interpret, not only to end-users, such as policy makers and citizens, but even to their engineers (Gunning et al., 2019; Miller, 2019; Murdoch et al., 2019; Waldrop, 2019). For instance, a deep learning system might predict health outcomes based on high-dimensional interactions among hundreds of variables – patterns that are impossible for human minds to grasp.

AI research initially focused on optimizing AI performance, aiming to design systems that make the most accurate predictions, regardless of whether those predictions are interpretable. More recently, however, stakeholders suggested AI interpretability is important in its own right: from medical doctors rejecting the adoption of AI systems due to lack of insight about how they work (Hardesty, 2016), to business executives expressing concerns that "AI's inner workings are too opaque" (Kahn, 2018), to social scientists proposing an "imperative of interpretable machines" (Stoyanovich et al., 2020) and the European Union establishing a right to obtain "meaningful information about the logic involved" in AI decisions (European Parliament & European Council, 2016). These concerns have sparked debates about how much interpretability should be prioritized relative to overall AI performance, given that

interpretability oftentimes (Lipton, 2017; but not always, see Rudin, 2019) comes at the cost of accuracy (Doshi-Velez & Kim, 2017; Gilpin et al., 2019; Gunning et al., 2019; Murdoch et al., 2019). While much of this recent attention has focused on the technical feasibility of interpretable AI (Lipton, 2017; Mittelstadt et al., 2019; Rudin, 2019), little is known about the public demand for interpretable AI. To address this gap, I conducted five empirical studies investigating whether and how much people without expertise in AI actually care about AI interpretability – per se and when it comes at a tradeoff with performance – across a variety of real-world applications.

Because explanation is a central component of interpretable AI (Miller, 2019; Murdoch et al., 2019), psychological research on explanation provides a useful starting point for predicting public demand for interpretable AI. Decades of research documents explanation as a fundamental human need (De Graaf & Malle, 2017; Heider, 1958; Keil, 2006; Lombrozo, 2011; Malle, 2006). Explanations facilitate understanding and guide subsequent learning, prediction, and feelings of control (Lombrozo, 2006). In serving these functions, explanation is essential for establishing trust (Heider, 1958; Langer et al., 1978; Lombrozo, 2006). If explanations play a similar role in human-machine interactions, interpretability will be a necessary precondition for establishing trust in AI (Miller, 2019; Stoyanovich et al., 2020). Indeed, a recent study provides initial evidence: when people perceived an AI joke-recommender system as opaque, they avoided relying on its recommendations, even when they knew that the AI outperformed humans in terms of accuracy (Yeomans et al., 2019).

This raises two immediate questions: does demand for interpretability generalize beyond recommender systems and does it generalize beyond the particular case of jokes? By their very nature, jokes do not rank among the most

serious AI applications. If people are evidently concerned about interpretability even in this domain, their desire to understand machine-made outputs might plausibly be enhanced for applications involving higher stakes, such as medical care, criminal justice, or social benefits. It is these instances, where interpretability's functions (Lombrozo, 2006; Miller, 2019) of predicting, auditing, and controlling underlying decision-making processes become most crucial (Lombrozo, 2006; Joy Lu et al., 2019). Considering low- versus high-stake versions of the same domain reinforces the intuition: you would probably care more about understanding the decision-making process that leads to being accepted or rejected for a salaried permanent job-position, compared to an unpaid honorary job-position. And indeed, decades of research have documented magnitude-sensitivity as a fundamental driving force of valuation at the cognitive level (Kahneman & Tversky, 1979; Platt & Huettel, 2008). Testing whether such valuation-differentials for low- relative to high-magnitude outcomes translate into enhanced demand for interpretability in high- relative to low-stakes AI applications, is an outstanding step.

Many emerging AI applications are designed to determine access to scarce but desirable resources, such as jobs, financial loans, or medical care. Ample theoretical and empirical work demonstrates that people demand explanations for decisions involving the allocation of resources (Brockner & Wiesenfeld, 1996; Lind & Tyler, 1988; Skitka et al., 2003), especially when those resources are scarce (Brockner & Wiesenfeld, 2005; Lerner & Lerner, 2013), in order to ensure that the allocation procedure was fair. Importantly, existing literature suggests that people's concerns about fairly allocating scarce resources are dissociable from concerns about stakes (Skitka et al., 2003). This

suggests that people may be more concerned about AI interpretability in applications that allocate scarce resources, independent of the stakes at hand.

Thus, as a first step towards characterizing people's preferences for interpretability in AI, I surveyed demand for interpretability across a variety of real-world applications where AI systems either make recommendations to a human decision-maker, or take decisions on behalf of a human decision-maker (Studies 4.1A and 4.1B). Next, I investigated the causal impact of stakes and scarcity on demand for interpretability across several AI applications (Study 4.2). Since AI interpretability sometimes comes at the cost of accuracy, I further characterized preferences for interpretability when it directly trades off with accuracy (Studies 4.3-4.4). Finally, I examined the robustness of my findings across different stakeholder perspectives and different measurements (Study 4.4).

All studies were approved by the Yale University Human Subjects Committee (approval number: HSC 2000022385) and participants in each study gave their informed consent beforehand. Participants were recruited via MTurk (Studies 4.1A, 4.2-4.4) or Prolific (Study 4.1B) and paid at an hourly rate of $7.25 in line with the US minimum wage. A pre-registration for Study 4.1B with the Open Science Framework is accessible here. Together with my supervisor and collaborators Lan Luo, Elisa Celis, and Molly J. Crockett I am currently working on revisions of the manuscript that we have been invited to resubmit at *Nature Communications*.

Study 4.1A: Demand for AI interpretability across a variety of applications

As a first step, I wanted to examine whether there is positive demand for interpretability in AI among the general public and whether it generalizes across a variety of applications. Because currently deployed AI systems often merely provide recommendations for human users to implement, rather than making decisions autonomously as they may in the near future, I tested whether demand for interpretability in AI varies along these dimensions. Based on past work suggesting that people demand more explanation from intentional agents (De Graaf & Malle, 2017), I predicted that people would demand more interpretability for AI systems deciding in an autonomous capacity relative to systems providing recommendations.

I also explored whether demand for interpretability is stronger in settings with higher stakes and scarcity (e.g., allocating medical treatment) than in settings with lower stakes and scarcity (e.g., assisting in shopping) in line with existing research suggesting that people demand explanations more for high-stakes than low-stakes decisions (Lombrozo, 2006; Miller, 2019; Rudin, 2019) and more for decisions involving the allocation of scarce resources (Brockner & Wiesenfeld, 2005, 1996; Lerner & Lerner, 2013; Lind & Tyler, 1988; Skitka et al., 2003).

Design

I used a mixed design with *AI autonomy* (*recommend* versus *decide*) as between-subjects factor and a set of 20 AI applications that varied within-subjects.

Participants

I recruited 200 US participants via MTurk. I excluded one duplicate response and 29 participants who failed a comprehension check on their second attempt from the analyses, leaving a final sample of N = 170. The final sample included 107 males, 60 females, and 3 "other" with an average age of 36.94 (SD = 11.83).

Procedure

Participants learned that I was interested in their attitudes towards AI, defined as follows (B. Zhang & Dafoe, 2019): "Artificial Intelligence (AI) refers to computer systems that make predictions, recommendations, or decisions by learning from existing datapoints. This computerized process is automated and occurs without explicit human instructions". They proceeded to questions about their intuitions on "the extent to which people [with/without] training in computer science can explain how an AI reaches certain predictions, recommendations, or decisions in certain cases" (5-point rating scale from "cannot explain at all" to "can explain fully") and their general support for AI (5-point rating scale from "strongly oppose" to "strongly support"). Next, I provided them with a precise definition of explainability reading "By explainable we mean that an AI's [decision/recommendation] can be explained in non-technical terms. In other words, it is possible to know and to understand how an AI arrives at its [decision/recommendation]". Following past work in psychology and philosophy of science (Miller, 2019), I used the more intuitive term 'explainable' rather than 'interpretable' while ensuring that the definition aligned with both terms' prevalent use in existing work on interpretable AI (Murdoch et al., 2019).

The definition was followed by a simple comprehension check that all participants were ultimately allowed to pass, though the final sample excluded

those who failed on two attempts (probe: “What is meant by ‘explainable AI’?”; correct response: “That an AI’s [decision/recommendation] can be explained in non-technical terms”). Participants then learned that they would go through a series of AI applications, for each of which they answered “how important is it that the AI in this application is *explainable*, even if it performs accurately?” on a dichotomous 5-point rating scale from “not at all important” to “extremely important”.

To test the role of AI autonomy in demand for interpretability, half of the participants were randomized to receive a *recommend* version that described an AI system making recommendations to a human decision-maker, while the other half of participants received a parallel *decide* version that described an AI system deciding on behalf of a human user. For example, the *recommend* version of the ‘medical treatment’ application read “An AI recommends to a doctor what disease a patient might be suffering from”, whereas the corresponding *decide* version read “An AI establishes on behalf of a doctor what disease a patient might be suffering from” (see *Study 4.1A* in *Appendix C*: Additional procedures and materials for the full set of 27 applications). Each participant saw a random subset of 20 out of these applications, presented in randomised order on subsequent pages with a reminder about explainability’s definition displayed at the bottom (“Reminder: ‘Explainable’ means that the AI's recommendation can be explained in non-technical terms”). On one round during this main task, participants were presented with an attention check for which the label read “AI application: detecting inattention” and the vignette text “If you read this, please ignore the below question and move the slider to the 'moderately important' scale point”.

After the main task, participants indicated how important they considered the respective motives "explain to justify/to verify/to improve/to discover" (B. Zhang & Dafoe, 2019). I then probed them again on their support for AI and asked how likely they considered it that their occupation would be replaced by AI at some point, and whether they had any computer science knowledge. The survey concluded with standard demographics.

Results

I first examined the effect of AI autonomy (recommend versus decide) on participants' demand for interpretability using ordinal regression analysis. As predicted, the *decide* condition elicited a small but significant increase in demand for interpretability relative to the *recommend* condition across applications, $\hat{\beta}$ = 0.15, p = .015, 95% CI [0.03, 0.27]. However, rank-orderings of applications according to interpretability demand did not change substantially between *recommend* and *decide* conditions (see Study 4.1A in *Appendix D*: Additional results). Summarizing across *recommend* and *decide* conditions, participants viewed interpretability as important (M = 3.59), averaging significantly higher than the scale's "moderately important" midpoint, $t(3,575)$ = 27.00, p < .001, 95% CI [3.54, 3.63].

In addition, I observed substantial variation in demand for interpretability across AI applications, with the highest demand for 'parole reviewing' (M = 4.61) and 'allocating medical treatment' (M = 4.59), followed by applications such as 'political news reporting' (M = 3.35), and lowest demand for 'organizing pictures' (M = 2.34; *Figure 13*).

Next, I considered whether variation across applications in terms of their stakes and scarcity predicted variation in demand for interpretability. I first considered the effects of stakes by designating each application as "low",

"medium", or "high stakes" (see *Figure 13*). Entering stakes as the main effect into an ordinal regression analysis confirmed increased demand for interpretability in high-stakes ($\hat{\beta}$ = 2.62, p < .001, 95% CI [2.43, 2.81]) and medium-stakes applications ($\hat{\beta}$ = 1.69, p < .001, 95% CI [1.53, 1.85]) relative to low-stakes applications, with demand in high-stakes applications also exceeding demand in medium-stakes applications, $\hat{\beta}$ = 0.93, p < .001, 95% CI [0.78, 1.09]. Similarly, I assigned each application a binary classification indicating whether the AI allocated a scarce resource or not (see *Figure 13*). Ordinal regression analysis with scarcity entered as the main effect confirmed increased demand for interpretability in applications that potentially allocated scarce resources, relative to those that did not, $\hat{\beta}$ = 0.84, p < .001, 95% CI [0.71, 0.97].

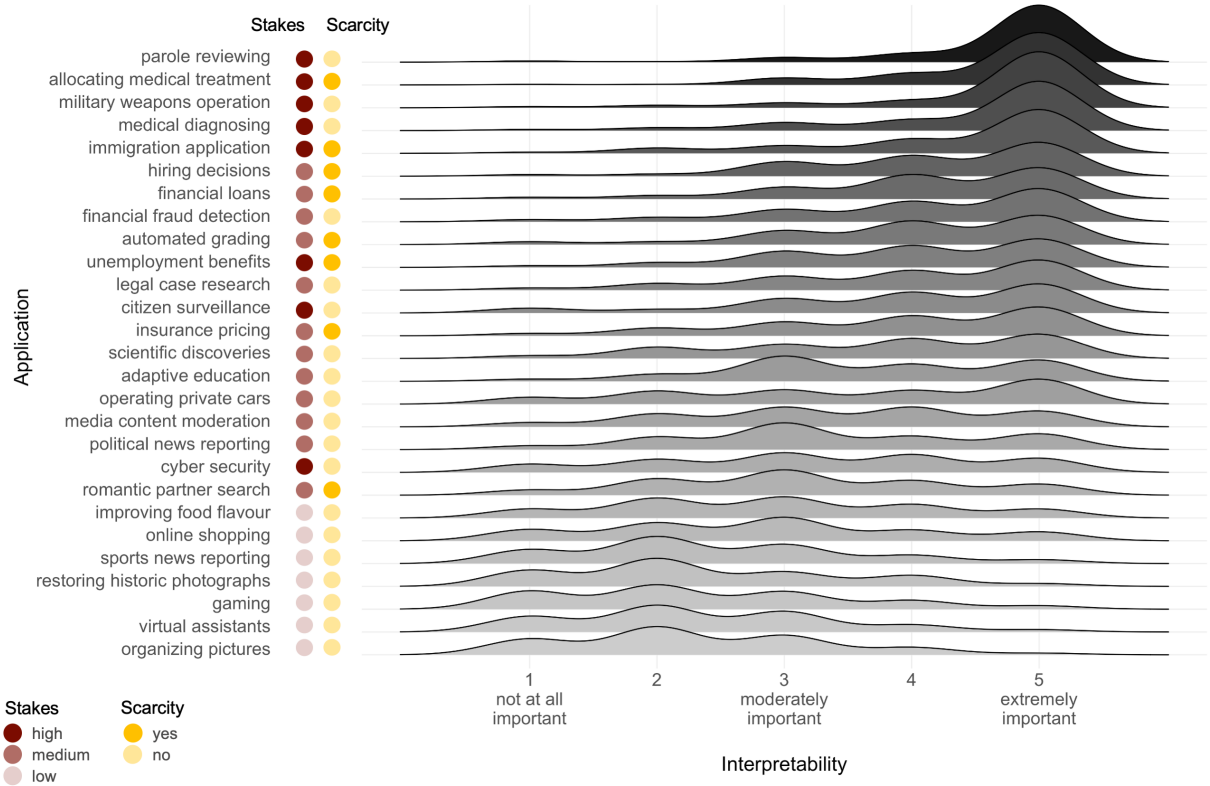


Figure 13. Demand for interpretability across real-world AI applications. Joyplot visualizes the distributions of interpretability ratings, averaged across *recommend* and *decide* versions. Participants responded to the question "how important is it that the AI in this application is explainable, even if it performs accurately?" on a 5-point rating scale (1 = not at all important, 5 = extremely important).

Discussion

The results of Study 4.1A demonstrate a positive demand for interpretable AI, which generalized across more than two dozen AI applications. Demand for interpretability was slightly higher for AI with direct agency relative to AI making recommendations to human decision-makers. As an intermediate step in further characterising demand for interpretable AI and its driving forces, I aimed to replicate the main findings of the present study in representative non-expert samples from two populations (i.e., US and UK) and with slightly amended materials.

Study 4.1B: Replicating demand for AI interpretability in representative US and UK samples

Study 4.1B probed whether the previous study's findings would replicate in samples from the US and UK that were representative in terms of gender, age, and ethnicity. I dropped the manipulation of AI autonomy and instead focused on testing whether the observed (positive) attitudes towards interpretability in AI were robust to varying the language used to probe them and to using a validated categorization of the applications in terms of involved stakes and scarcity (see Study 4.1B Results). I also added two AI applications (see Study 4.1B in *Appendix C*: Additional procedures and materials) so that the full set comprised 29 vignettes.

Design

I used a descriptive design in which all participants were presented with the same 29 different AI applications in randomized order.

Participants

I recruited 300 participants for each national sample. Both US and UK samples were recruited on Academic Prolific, using the platform's feature for collecting representative samples that match census data in terms of age by sex by ethnic group proportions.

For the US sample, Prolific stopped sampling at 293 participants (this occurs when it takes too long to fill the missing cases for the required combination of quotas). Of those, in line with the pre-registration, I excluded 33 participants who failed comprehension checks on both attempts, or who failed all attention checks. The resulting final sample of N_{US} = 258 was 48% male with an average age of 45.53 (*SD* = 16.41).

For the UK sample, Prolific stopped sampling at 298 participants. Of those, in line with the pre-registration, I excluded 52 participants who failed comprehension checks on both attempts, or who failed all attention checks. The resulting final sample of N_{UK} = 246 was 48% male with an average age of 45.57 (*SD* = 15.94).

Procedure

Participants received the same definitions and instructions as in Study 4.1A. However, across instructions, comprehension checks, and dependent variable questions, I replaced the word "explainable" with "understandable" as a robustness check. Thus, as participants went through the complete set of 29 AI applications, they answered the question "In this case, how important is it that the AI is understandable?" for each application. In order to allow more fine-grained responses, I slightly changed answer format from a discrete measure to a continuous slider with five tick marks and the same three labels as in Study 4.1A ("not at all important", "moderately important", "extremely important"). To

comply with the data collection policy of Prolific, participants were not screened out during the survey when they met the pre-registered exclusion criteria (i.e., failing comprehension checks on both attempts and/or all attention checks). I added a comprehension check to verify that participants understood they should assume the AI to perform accurately (probe: "What should you do in the following task?"; correct response: "Consider how important it is that an AI is understandable, even if it performs accurately") and an attention check that was not related to the main task. Analogously, I extended the reminder displayed below the answer variable so that it read "'understandable' means that the AI's decision can be explained in non-technical terms. Please consider how important it is that the AI is understandable, even if it performs accurately". I collected the same control and demographic variables as in Study 4.1A.

Results

Following the pre-registered analysis-plan, I first tested whether attitudes towards interpretability (~understandability) in AI – aggregated across applications – exceeded the scale-midpoint "moderately important" for each respective national sample. Because demand did not differ across national samples, for the sake of readability the following reporting will deviate from the pre-registration by focusing on whether stakes and scarcity predicted attitudes towards interpretability in AI in the overall sample, aggregating across US and UK populations (but see Study 4.1B *Appendix D*: Additional results for results on the separate national samples).

To validate the post-hoc categorization of applications for these attributes, I invited nine independent raters (i.e., who were blind to the study hypotheses) to categorize each application according to the involved stakes (low/medium/high) and scarcity (yes/no). Aggregating across vignettes, 70% of

raters agreed in their stakes categorizations and 84% of raters agreed in their scarcity categorizations.

US sample. Participants viewed interpretability as important (M = 3.70, SD = 1.24), averaging significantly higher than the scale's "moderately important" midpoint, $t(7,481) = 49.32$, $p < .001$, 95% CI [3.68, 3.73], Cohen's d = 0.57.

UK sample. Again, participants viewed interpretability (~ understandability) as important (M = 3.68, SD = 1.26), averaging significantly higher than the scale's "moderately important" midpoint, $t(7,133) = 46.02$, $p < .001$, 95% CI [3.66, 3.71], Cohen's d = 0.54.

Overall sample. Participants' valuation of interpretability did not differ across national samples, p = .330. Hence, the following analyses were run on the aggregate sample. Because of multicollinearity issues, I had to estimate separate regression models for stakes and scarcity. I observed a significant main effect for stakes, $F(2, 14613) = 2236.80$, $p < .001$, $\eta^2 = 0.23$, 95% CI [0.22, 0.24]. Relative to applications involving low stakes, demand for interpretability was stronger for applications involving high (b = 1.50, $p < .001$, 95% CI [1.46, 1.55]) or medium stakes (b = 0.94, $p < .001$, 95% CI [0.89, 0.99]), and it was stronger amidst high than medium stakes, b = 0.56, $p < .001$, 95% CI [0.52, 0.60]. Similarly, a significant main effect for scarcity ($F(1, 14614) = 583.93$, $p < .001$, $\eta^2 = 0.04$, 95% CI [0.03, 0.05]) indicated increased demand for interpretability in applications potentially allocating scarce resources, relative to those that did not, b = 0.57, $p < .001$, 95% CI [0.52, 0.62].

Discussion

The results of Studies 4.1A and 4.1B suggest that stakes and scarcity might drive demand for interpretability in AI. However, these findings are correlational; the applications I tested varied on a number of other dimensions; and stakes and

scarcity designations were assigned to the applications post-hoc. Even though I sought to minimise researcher bias by using an independently validated rating in Study 4.1B, which replicated the results from Study 4.1A, the two attributes covaried in the sense that almost all applications involving high scarcity also involved high stakes. Indeed, there was no application involving low stakes but high scarcity across to the validated ranking. In the following study I thus sought to address these limitations by using an experimental approach to test the hypothesis of stakes and scarcity independently driving demand for interpretability in AI.

Study 4.2: How stakes and scarcity drive demand for interpretability in AI

To examine whether demand for interpretable AI is driven by stakes and scarcity, I independently manipulated the two factors in five AI applications: allocating vaccines, prioritizing hurricane first responders, reviewing insurance claims, making hiring decisions, and prioritizing standby flight passengers. Because Study 4.1B had confirmed that demand for interpretability is robust to varying the language used to probe it, I went back to the more common terminology of 'explainable' AI.

Design

I used a 2 (*stakes*: low versus high) by 2 (*scarcity*: low versus high) within-subjects design in which participants went through the resulting four versions of five different vignettes about AI applications (allocating vaccines, prioritizing first responders, reviewing insurance claims, making hiring decisions, allocating

standby passengers), indicating for each one their preference for explainability in the involved AI.

Participants

I recruited 120 US participants via MTurk. 36 participants failed comprehension checks on two attempts and were excluded from all analyses, resulting in a final sample of *N* = 84. The final sample included 46 males, 36 females, 1 "other", and 1 "prefer not to say" with an average age of 37.38 (*SD* = 11.76).

Procedure

Participants were presented with the same definitions of AI and explainability as in Study 4.1A. and the same comprehension checks as in Study 4.1B. All participants were allowed to pass these comprehension checks but those who failed them on their second attempt were excluded from the analyses. After the comprehension check, participants proceeded to the main task in which they were presented with the different (versions of) AI applications and for each one rated how important they thought it was that the involved AI was explainable. For each application, participants were presented with the four different versions in randomized order (see *Table 6* for an example). As in Study 4.1B, I used a continuous slider with five tick marks and three labels reading "not at all important", "moderately important", and "extremely important" as dependent variable. I fully randomized the order of applications across participants, as well as the order of versions within each application (high/low stakes and scarcity).

I collected the same additional variables as in Studies 4.1A/B and added a question asking whether participants had heard about bias in AI and if so whether they thought their answers had been influenced by this.

Table 6. Exemplary AI application with its four versions according to the stakes/scarcity manipulation. Participants were presented with one version at a time and the different versions were randomized in order.

It is flu season. An AI decides whether or not a citizen will get a vaccine.

	High scarcity	Low scarcity
High stakes	In this case, **the vaccine supply is <u>very limited</u>**. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, **the vaccine cannot be stocked in large quantities**. This particular vaccine protects against a **<u>deadly</u>** variant of the flu.	In this case, **the vaccine supply is <u>abundant</u>**. Because the vaccine is very cheap, easy to produce, and can be stored at any temperature, **the vaccine can be stocked in large quantities**. This particular vaccine protects against a **<u>deadly</u>** variant of the flu.
Low stakes	In this case, **the vaccine supply is <u>very limited</u>**. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, **the vaccine cannot be stocked in large quantities**. This particular vaccine protects against a **<u>mild</u>** variant of the flu.	In this case, **the vaccine supply is <u>abundant</u>**. Because the vaccine is very cheap, easy to produce, and can be stored at any temperature, **the vaccine can be stocked in large quantities.** This particular vaccine protects against a **<u>mild</u>** variant of the flu.

Results

I first ran a linear mixed-effects model predicting participants' demand for interpretability, with 'stakes' (low versus high), 'scarcity' (low versus high), and their interaction entered as fixed effects while 'subject' and 'application' were entered as random intercepts. Using type II Wald chi-square tests to probe the fixed effects' significance, I observed significant main effects for stakes ($\chi^2(1) = 348.48, p < .001$) and scarcity ($\chi^2(1) = 110.98, p < .001$) on demand for interpretability, which were not qualified by an interaction, $\chi^2(1) = 0.10$, $p = .754$. Participants indicated a stronger demand for interpretability for high- relative to low-stakes

cases (b = 0.85, p < .001, d = 0.33, 95% CI [0.25, 0.40]) and for high- relative to low-scarcity cases, b = 0.49, p < .001, d = 0.19, 95% CI [0.11, 0.26]. These effects were robust when I added gender, age, education, income, pre- and post-task support for AI, and computer science knowledge to the model (*Table 7*).

Table 7. Random intercept model predicting interpretability ratings (Study 4.2)

Variables	b (SE)	p	d	95% CI_d
(Intercept)	2.84 (0.61)			
Stakes	0.85 (0.06)	< .001	0.32	[0.24, 0.40]
Scarcity	0.48 (0.06)	<.001	0.18	[0.11, 0.26]
Gender (female)	0.24 (0.14)	.104	0.04	[-0.08, 0.16]
Age	-0.01 (0.01)	.071	0.04	[0.02, 0.07]
Education	0.06 (0.06)	.330	0.02	[-0.05, 0.10]
Income	0.02 (0.04)	.675	0.01	[-0.05, 0.07]
Support (pre-task)	0.21 (0.35)	.551	0.01	[-0.17, 0.20]
Support (post-task)	-0.11 (0.34)	.740	0.01	[-0.17, 0.19]
Knowledge about bias in AI	-0.07 (0.18)	.685	0.01	[-0.12, 0.14]
Some CS experience	-0.23 (0.3)	.441	0.02	[-0.15, 0.19]
College degree in CS	-0.01 (0.3)	.965	0.00	[-0.17, 0.17]
Undergraduate degree in CS	0.01 (0.3)	.971	0.00	[-0.17, 0.17]
Graduate degree in CS	0.38 (0.4)	.336	0.02	[-0.17, 0.22]
Stakes x scarcity	-0.03 (0.09)	.757	0.01	[-0.08, 0.10]

Note. Estimates from random intercept model predicting interpretability ratings by stakes, scarcity, and various control variables. CS denotes Computer Science.

As a further robustness check, I ran separate mixed effects models for each AI application. As illustrated in *Figure 14*, the positive effects of stakes and scarcity on demand for interpretability generalized across all applications even though the effect for scarcity was not significant in the 'hiring' application ($\chi^2(1) = 1.87$, $p = .172$) and weak in the 'standby passenger' application ($\chi^2(1) = 4.12$, $p = .042$).

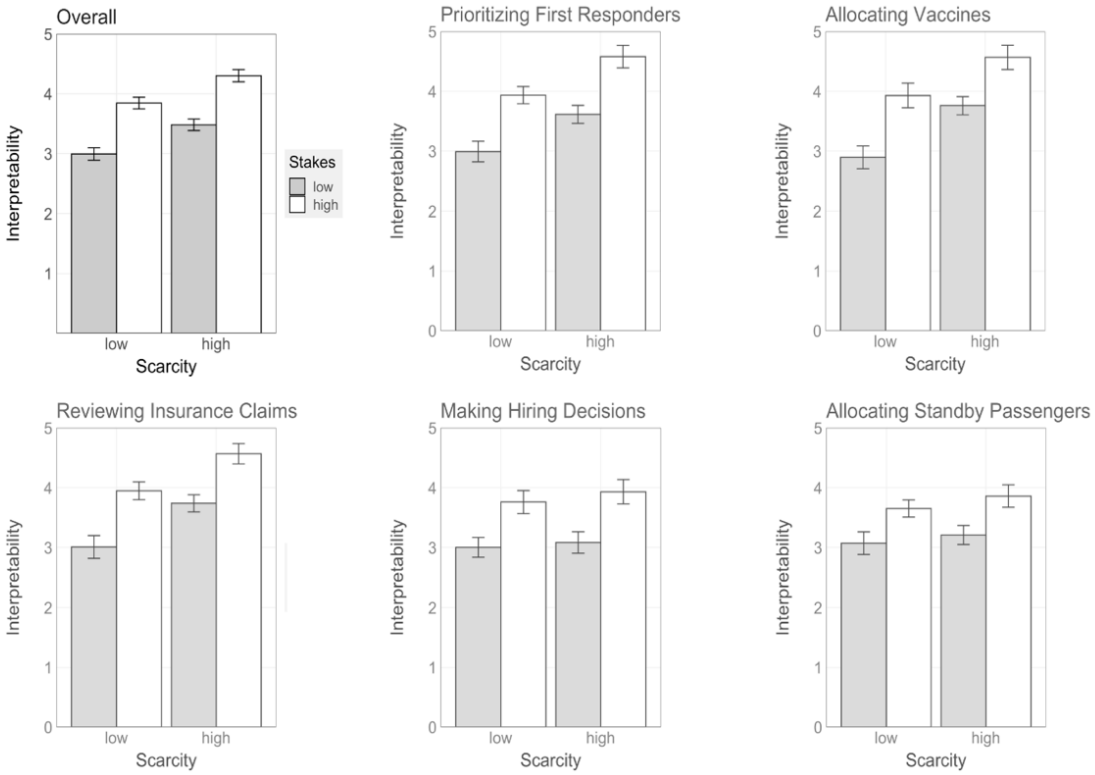


Figure 14. Barcharts illustrating main results for Study 4.2. Y-axis represents participants' responses to the question "How important is it that the AI is explainable, even if it performs accurately?" on a continuous slider-scale from "not at all important" (1) to "extremely important" (5). The first panel aggregates across all five applications.

Discussion

Studies 4.1A/B and 4.2 established a robust demand for interpretability in AI, generalizing across a wide range of applications, that was higher when applications involved high stakes and scarce resources. In these studies, I held the level of accuracy constant by explicitly instructing participants to rate

interpretability's importance for a given application "even if the AI performs accurately". Because it has been widely argued that, in practice, interpretable AI may require trading off interpretability against accuracy (Doshi-Velez & Kim, 2017; Gilpin et al., 2019; Gunning et al., 2019; Murdoch et al., 2019), in Study 4.3 I sought to investigate people's preferences for interpretability in AI when it explicitly comes at the cost of accuracy.

Study 4.3: How stakes and scarcity impact tradeoffs between AI interpretability and accuracy

To examine how stakes and scarcity impact demand for interpretability when it comes at the cost of accuracy, I replicated Study 4.2 with one key modification: participants responded on a slider measure where one end represented a 'completely accurate' but 'not at all explainable' AI, whereas the other end represented a 'not at all accurate' but 'completely explainable' AI (the "tradeoff slider"; see *Figure 15A*). As I was interested in people's a priori preferences, I continued using the scenario format from Studies 4.1A/B and 4.2, where I did not specify the outcome of the machine-made decisions. Previous work from psychology suggests people might perceive AI-accuracy as a proxy, or at least precondition, for ensuring favorable outcomes (Brockner & Wiesenfeld, 2005, 1996), which would suggest an overall preference for accuracy over interpretability.

Participants

I recruited 120 US participants via MTurk. To reduce the dropout from failed comprehension checks that I had observed in Study 4.2, only those participants who successfully passed the comprehension checks proceeded to

the full survey. Participants who failed to pass either comprehension check on their second attempt were paid a compensation fee for their time until that point and were replaced. This left a final sample of *N* = 116 participants after exclusions due to duplicate submission or incomplete data. The final sample included 79 males, 36 females, and 1 "prefer not to say" with an average age of 36.51 (*SD* = 10.21).

Design

Again I used a 2 (*stakes*: low versus high) by 2 (*scarcity*: low versus high) within-subjects design in which participants went through the resulting four versions of five different vignettes about AI applications (allocating vaccines, prioritizing first responders, reviewing insurance claims, making hiring decisions, allocating standby passengers), indicating for each one their preference over a tradeoff between explainability and accuracy in the involved AI.

Procedure

Participants first read definitions of AI and explainability. In order to align the instructions across studies and to reflect the tradeoff dependent variable used in the present study, I slightly amended the definition of AI explainability such that it read "By 'explainable' we mean that it is possible to understand how an AI arrives at its decision. In other words, the more explainable an AI, the more one can understand about how it makes its decisions. The less explainable an AI, the less one can understand about how it makes its decisions". Additionally, participants were presented with a definition of 'accuracy' reading as follows: "By 'accurate' we mean that an AI's decisions are correct. In other words, the more accurate an AI, the fewer mistakes it makes when performing decisions. The less

accurate an AI, the more mistakes it makes when performing decisions". The instructions also highlighted that, ideally, one would want an AI that is both completely accurate and completely explainable, but that, in practice, a more accurate AI is often less explainable; vice versa, a more explainable AI is often less accurate (e.g., Murdoch et al., 2019). Correspondingly, I amended the comprehension checks and added two additional questions ensuring that participants had understood the definition of accuracy and the tradeoff-slider's format.

Participants then were presented with the five applications and their four respective versions in randomized order. On each round, they used the tradeoff-slider (*Figure 15*A) to indicate their preferences over interpretability and accuracy by moving the slider's button from the middle position to either end. The ordering of the accuracy and explainability labels, and correspondingly the valence of the slider measure's ends, was counter-balanced across participants. Participants were reminded that "Accuracy depends on correctly determining which standby passengers need to board urgently. Explainability depends on understanding the criteria used to determine which standby passenger needs to board urgently", with the reminder-order mirroring the order of the two attributes encountered in previous instructions and the slider set-up assigned to a given participant. In order to explore additional explanatory factors in people's preferences for accurate and interpretable AI, I added randomly ordered questions probing participants' perception of (i) the reversibility for the AI's decisions, (ii) the level of expertise required for a human to perform the AI's decision, (iii) the likelihood of being personally affected, and (iv) the number of people affected. Participants answered these measures separately for each of

the applications from the main task. Otherwise, I collected the same additional variables as in Study 4.2.

Results

I coded participants' responses such that positive values represent a preference for interpretability over accuracy and negative values indicate a preference for accuracy over interpretability. These data revealed an overall preference for accuracy over interpretability, signified by a mean rating of M = -0.36 that differed significantly from the indifference point of 0, $t(2,319) = -12.62$, $p < .001$, 95% CI [-0.42, -0.30].

Next, I ran a linear mixed effects model predicting participants' tradeoff preferences, with 'stakes', 'scarcity', and their interaction entered as fixed effects while I entered 'subject' and 'application' as random intercept effects. Using type II Wald chi-square tests to probe the fixed effects' significance, I observed significant main effects for stakes ($\chi^2(1) = 59.45$, $p < .001$) and scarcity ($\chi^2(1) = 26.66$, $p < .001$) on tradeoff preferences, which were not qualified by an interaction, $\chi^2(1) = 0.54$, $p = .461$ (*Figure 15B*). Overall, participants preferred accuracy over interpretability, and this preference was amplified by the same conditions that impacted preferences for interpretability in Study 4.2. That is, participants' preferences for accuracy over interpretability were more pronounced for high relative to low stakes cases (b = -0.42, $p < .001$, d = 0.12, 95% CI [0.05, 0.20]) and for cases involving high relative to low scarcity, b = -0.29, $p < .001$, d = 0.09, 95% CI [0.01, 0.16]. Main effects for stakes and scarcity were robust (stakes: $p < .001$; scarcity: $p < .001$) to controlling for AI- and task-related covariates, in particular the ordering of accuracy and interpretability across instructions and response-variable (p =. 261), pre- and post-task support for AI (ps ≥. 590), computer science knowledge (ps ≥. 067), and demographic variables (ps ≥. 391). Main effects for

stakes and scarcity also remained significant when I added further explanatory candidates, such as decision-reversibility or personal affectedness, to the model (see *Table A2* in *Appendix D*: Additional results).

Discussion

The results of Study 4.3 suggest that people prioritize AI accuracy over interpretability when the two trade off against one another. Moreover, participants appear to be more inclined to sacrifice interpretability for accuracy under the very same conditions that drive preferences for interpretability (i.e., high stakes and high scarcity). To build on these findings, I next sought to probe participants' preferences for interpretability and accuracy in AI when the two attributes are evaluated separately, as well as from different stakeholder perspectives.

Study 4.4: Stakes and scarcity impact interpretability-accuracy tradeoffs across stakeholder perspectives

Taken together, Studies 4.1A/B to 4.3 provide evidence that people value interpretability in AI, although they prioritize accuracy over interpretability when they trade off with one another. Moreover, stakes and scarcity drove preferences for accuracy over interpretability in the same direction as their impact on preferences for interpretability alone. In Study 4.4, I sought to replicate and extend these findings by examining how stakes and scarcity affect preferences for interpretability and accuracy in tandem as well as independently, and whether these effects are robust across different stakeholder perspectives.

Design

I used a mixed design with stakeholder *perspective* (agent versus patient) administered as between-subjects factor and *stakes* (low versus high) and *scarcity* (low versus high) administered as fully crossed within-subjects factors.

Participants

I recruited a final sample of *N* = 247 US participants via MTurk following the procedure from Study 4.3. The final sample included 162 males, 84 females, and 1 "other" with an average age of 35.60 (*SD* = 9.64).

Procedure

Participants were presented with the same definitions, instructions, and comprehension checks as in Study 4.3. However, to test robustness across stakeholder perspectives, I varied whether participants evaluated the AI application from the perspective of a responsible *agent*, who manages the AI application, or from the perspective of a *patient* who is directly impacted by the AI's decision. For example, in the 'vaccine allocation' application, participants assigned to the agent condition were asked to evaluate the AI from the perspective of a public health director responsible for managing the AI making vaccination decisions, while participants assigned to the patient condition were asked to evaluate the AI from the perspective of a patient seeking to get a vaccine. Participants first indicated their preferences for interpretability at the expense of accuracy on the tradeoff-slider from Study 4.3 and, subsequently, on a separate page, for interpretability and accuracy separately. Again, the order of the two attributes was counterbalanced across participants.

Results

First, I tested the main effects of stakes and scarcity on preferences for interpretability relative to accuracy when the two attributes traded off against one another. Replicating the results of Study 4.3, participants overall prioritized accuracy over interpretability, as indicated by a mean tradeoff rating of M = -0.35, which differed significantly from the indifference point of 0, $t(4,938) = -17.73$, $p < .001$, 95% CI [-0.39, -0.31]. Similarly, I replicated the main effects for stakes ($\chi^2(1) = 216.30$, $p < .001$) and scarcity ($\chi^2(1) = 60.19$, $p < .001$) on tradeoff preferences, which were not qualified by an interaction ($\chi^2(1) = 0.15$, $p = .698$; *Figure 15C/D*). These main effects followed the previously established pattern whereby participants' preference for accuracy was more pronounced for high- relative to low-stakes cases (b = -0.51, $p < .001$, d = 0.15, 95% CI [0.10, 0.20]) and for cases characterized by high relative to low scarcity, b = -0.28, $p < .001$, d = 0.08, 95% CI [0.03, 0.13].

These results generalized across patient and agent perspectives: when I added a main effect for *perspective* and its interaction with stakes and scarcity to the model predicting tradeoff preferences, main effects for *stakes* ($\chi^2(1) = 216.37$, $p < .001$) and *scarcity* ($\chi^2(1) = 60.21$, $p < .001$) remained significant (*Figure 15C*), while there was no main effect for *perspective* ($\chi^2(1) = 0.17$, $p = .677$) nor any significant two- or three-way-interactions between perspective and the other two predictors (ps > .221).

Next, I examined the effects of stakes and scarcity on preferences for interpretability and accuracy measured independently. I observed a significant overall demand for interpretability (M = 3.30), which exceeded the "moderately important" scale-midpoint, $t(4,938) = 17.68$, $p < .001$, 95% CI [3.27, 3.34]. Similarly, I observed a significant demand for accuracy (M = 3.82), which also exceeded the "moderately important" scale-midpoint, $t(4,938) = 49.82$, $p < .001$, 95% CI [3.78,

3.85]. Demand for accuracy was significantly higher than demand for interpretability, $t(9,861) = 21.76$, $p < .001$, |95% CI| [0.47, 0.65], consistent with the results obtained from the tradeoff measure. Again, I replicated main effects for stakes and scarcity on both separate measures (regressing on accuracy: *stakes:* $\chi^2(1) = 391.22$, $p < .001$; *scarcity:* $\chi^2(1) = 99.73$, $p < .001$; regressing on interpretability: *stakes:* $\chi^2(1) = 13.91$, $p < .001$; *scarcity:* $\chi^2(1) = 8.39$, $p = .004$). As with the tradeoff measure, there was no main effect of stakeholder perspective on either preferences for interpretability or accuracy (*p*s > .253), nor did perspective moderate the main effects of stakes or scarcity on either dependent variable (*p*s > .151).

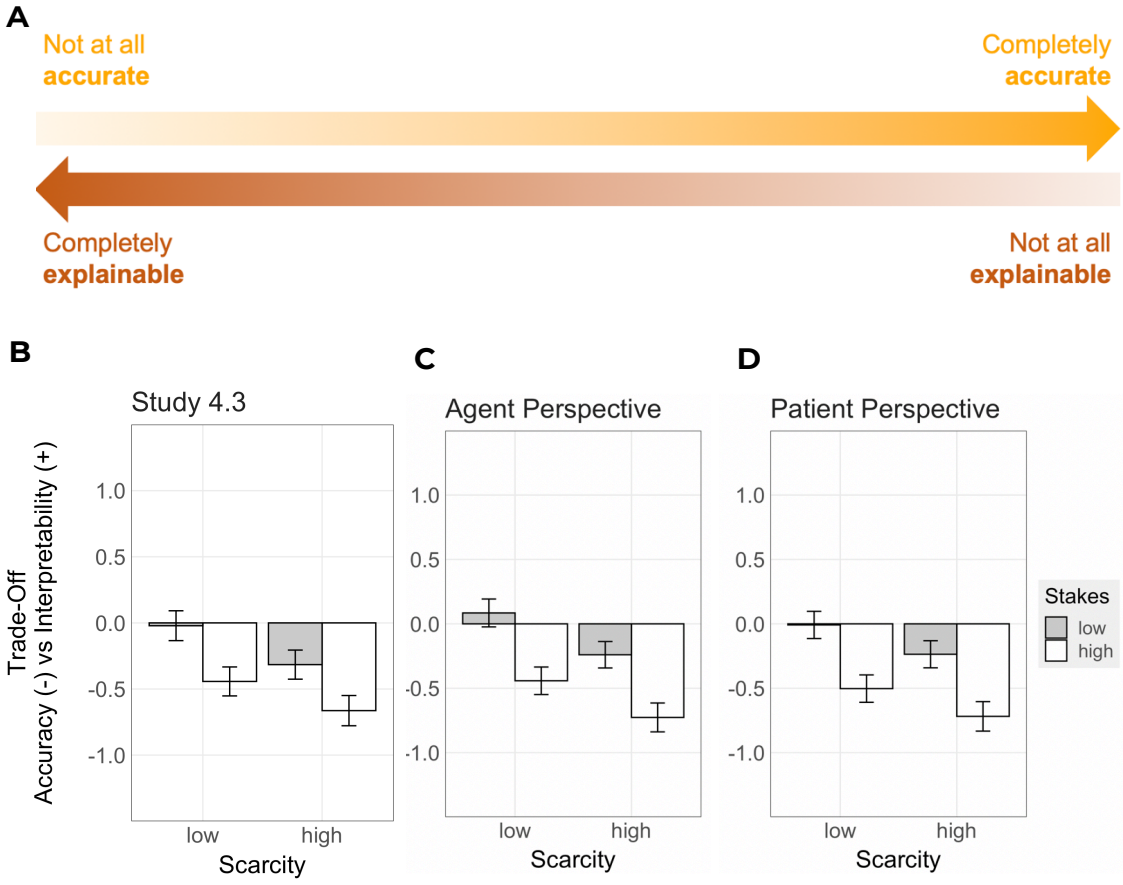


Figure 15. Dependent variable and results for Studies 4.3 and 4.4. (**A**) Labels of dependent measure displayed with a continuous slider on which participants answered "Would you prefer an AI that is …". The order of the accurate/explainable attributes and hence the direction of the slider were counter-balanced across participants. (**B**) Results from Study 4.3. Y-axis represents participants' responses on the slider from (**A**). (**C** and **D**) Results from Study 4.4 broken down by agent and patient perspectives using the slider from (**A**). Error bars represent 95% confidence intervals.

Putting the three measures of preferences into relation, I observed a stronger association between the separate accuracy measure and the trade-off measure ($|r_s(4,939)| = .65$, $p < .001$, |95% CI| [0.66, 0.63]) compared to a weaker association between the separate interpretability measure and trade-off measure, $r_s(4,939) = .57$, $p < .001$, 95% CI [0.55, 0.59]. This might indicate spillover effects of different strengths from the trade-off to the subsequent separate measures, whereby participants' evaluations on the separate measures were influenced by their trade-off preferences and more so for accuracy compared to interpretability. Further support for this interpretation comes from a negative correlation between the two separate measures ($r_s(4,939) = - .34$, $p < .001$, 95% CI [-0.37, -0.32]), which implies that stronger valuations of separately measured accuracy came with lower valuations of separately measured interpretability. Finally, I observed significant order-effects on the trade-off ($\chi^2(1) = 14.24$, $p < .001$) and separate interpretability measures ($\chi^2(1) = 5.99$, $p = .014$), but not on the separate accuracy measure ($\chi^2(1) = 0.41$, $p = .524$) when I added a dummy-predictor to the respective regression models (main effects for stakes and scarcity remained significant). These were driven by a stronger preference for interpretability amongst participants who encountered this attribute first across instructions and measures.

Discussion

The results of Studies 4.3 and 4.4 provide consistent evidence for an overall preference in favour of accuracy over interpretability when these properties are mutually exclusive. However, the pattern of results across the different measures in Study 4.4 might also point towards an advantage of accuracy over interpretability in terms of evaluability (Hsee, 1996): the observation of separate accuracy-ratings remaining unaffected by the order of

attribute-presentation, whereas interpretability-ratings seemed susceptible to the order of presentation, could indicate that accuracy is – relatively seen – easier to evaluate and as such elicits more robust, consistent, and pronounced preferences (ibid.). Meanwhile, people's preferences for interpretability, accuracy, and their trade-off were stable across an immediately affected patient perspective and a less directly affected (and as such potentially more normative) agent perspective. This could indicate that the characterized set of preferences is sufficiently fundamental to be shared by stakeholders standing on different sides, although it will be an important task for future work to extend the range of perspectives as well as to probe preferences among actual stakeholders (e.g., healthcare providers versus healthcare beneficiaries).

Overall discussion

In recent years, academics, policymakers, and developers have debated whether interpretability is a fundamental prerequisite for trust in AI systems (e.g., Stoyanovich et al., 2020). However, it remains unknown whether non-experts – who may ultimately comprise a significant portion of end-users for AI applications – actually care about AI interpretability, and if so, under what conditions. Here, I characterize human preferences for interpretability in AI across five studies. The data demonstrates a robust and positive demand for interpretable AI, which generalized across more than two dozen AI applications and systematically varied along several dimensions. Demand for interpretability was higher for AI applications that directly implemented decisions, as opposed to recommending a course of action to a human (Study 4.1A), for applications making decisions with higher (relative to lower) stakes (Studies 4.1A/B-4.4), and for applications that determined access to scarce (relative to abundant)

resources (Studies 4.1A/B-4.4). However, my findings also demonstrate that demand for interpretability in AI is volatile: When interpretability came at the cost of accuracy, participants prioritized accuracy. Furthermore, participants favoured accuracy over interpretability under the very same conditions that drove demand for interpretability in the first place: when stakes are high and resources are scarce (Studies 4.3-4.4). These findings were robust regardless of whether participants considered the decision from the position of a responsible agent overseeing the AI application, or a patient directly affected by the AI's decision (Study 4.4).

Several implications follow from my observation that demand for interpretability is attenuated when AI makes recommendations to a human decision-maker, rather than deciding on behalf of a human decision-maker. The observed pattern may reflect a false intuition (Elish, 2019) that a human receiving recommendations from an AI understands how these outputs are generated and can respond accordingly, thereby compensating for the recommender system's lack of interpretability. There is a risk that such misattributions of understanding in human-AI interactions (Awad et al., 2019) could amplify recommender-systems' opacity. Consider, for example, a doctor receiving recommendations from a deep learning model on medical diagnoses such as tumor classifications. While the doctor's experience and intuition may lead them to diagnose a dark spot on an X-ray as a malignant tumor, the recommender-system may suggest an "all clear". Interacting with a non-interpretable AI-system, the doctor has no way to cross-check their diagnosis against the machine's recommendation; all they can do is choose between blindly trusting or disregarding the machine's suggestion, with no prospect of gaining insights into the "whys" that could benefit them and patients in future cases.

This scenario becomes all the more worrying given that even (medical) image-recognition-tools, praised among the most robust and capable AI systems available (Price, 2018), have proven vulnerable to so called “adversarial attacks”, where subtle changes in models’ input produce confident but manifestly wrong outputs (Finlayson et al., 2019). In a humorous but nonetheless unsettling demonstration of this vulnerability, merely changing a few photograph pixels was sufficient to trick Google’s famed *Inception* model into identifying a cat as guacamole with 99.99 percent confidence (Athalye et al., 2018). My findings substantiate the worry that recommender-systems’ allure in terms of high predictive power, in combination with people’s relaxed expectations in terms of model interpretability, could exacerbate the perils of machine-assisted decision-making in critical domains such as medical diagnosing.

My findings also highlight that high-stakes applications, such as medical diagnosis, will generally be met with an enhanced demand for interpretability in AI. Notably, this sensitivity to stakes parallels magnitude-sensitivity as a foundational process in the cognitive appraisal of outcomes (Kahneman & Tversky, 1979; Platt & Huettel, 2008). The impact of stakes on demand for interpretability was apparent not only in the experiments that manipulated stakes within a given AI-application, but also in absolute and relative levels of demand for interpretability across applications – take, for instance, ‘hurricane first aid’ and ‘vaccine allocation’ outperforming ‘hiring decisions’, ‘insurance pricing’, and ‘standby seat prioritizing’. Conceivably, this ordering would also emerge if we ranked the applications according to the scope of auditing- and control-measures imposed on human executives, reflecting interpretability’s

essential capacity of verifying appropriate and fair decision processes (Kleinberg et al., 2019; Joy Lu et al., 2019; Mehrabi et al., 2019; Miller, 2019).

Fairness concerns are also salient in 'gatekeeping' settings where decision-makers determine access to scarce resources (Brockner & Wiesenfeld, 1996; Lind & Tyler, 1988; Skitka et al., 2003). Accordingly, I found the demand for interpretability was higher when AI applications allocated resources under conditions of scarcity. These findings build on past work showing that people demand more explanation for decisions involving resource allocation in order to ensure that the allocation process was fair (Kleinberg et al., 2019; Rich et al., 2020; Stoyanovich et al., 2020), demonstrating that such principles also operate in the context of AI applications and substantiating recent calls for interpretability as a safeguard for ethical and fair AI systems.

Despite evidence across all five studies that people demand interpretability in AI, this demand proved volatile. Participants readily sacrificed interpretability for accuracy when the two traded off against one another, and were more inclined to do so for high-stakes applications and those involving the allocation of scarce resources. Prioritizing accuracy over interpretability by seeking "answers first, explanations later" accrues what the legal scholar Jonathan Zittrain (2019) has described as "intellectual debt": answers gained at the expense of understanding. Intellectual debt is risky because lacking understanding of *how* something works can produce negative unintended consequences in complex systems. For instance, if a drug is effective but the underlying mechanism is unknown, prescribing that drug can lead to dangerous side-effects if administered in combination with other drugs. Likewise, accruing intellectual debt in AI systems becomes riskier in settings where multiple AI applications will interact: consider a medical system where AI

diagnosis applications are used in combination with AI applications that decide who gets access to scarce medical treatments. My studies imply that participants endorse the accrual of intellectual debt, sacrificing interpretability for accuracy, under conditions of high stakes and high scarcity – exactly those conditions where negative unintended consequences are likely to produce the most damage.

As much as the present work offers a glimpse into public demand for interpretability in AI, it also highlights the need for deeper insights. One important limitation of the present work is that my experiments relied on relatively broad and simplified definitions of interpretability and accuracy, which may have influenced my findings. For instance, the prominence of "mistakes" in the definition of accuracy ("the more accurate an AI, the fewer mistakes it makes when performing decisions") might have inflated people's valuation of accuracy relative to interpretability in Studies 4.3 and 4.4. I also focused on *stakes* and *scarcity* as two explanatory variables emerging immediately from existing literature on explanation, and only focused on *a priori* appraisals, which might be rather different from *a posteriori* appraisals where outcomes are known.

Furthermore, the conclusions drawn from the presented work are limited to participants from the United States and the United Kingdom. Exploring demand for interpretable AI among other populations is a promising and important topic for future work, especially in light of recent evidence suggesting that expectations towards machine-made decisions can vary substantially across countries and cultures (Awad et al., 2018) and amidst reports about a proliferated impact of AI on the lives of low-income populations outside Western, educated, industrialized, rich, democratic contexts (Pilkington, 2019).

Finally, it remains an open question how preferences for interpretability and accuracy in AI ultimately shape human-machine interactions. A proximate question, primarily of concern to developers and policy-makers, is whether and how these preferences translate into trust in AI and willingness to adopt AI-powered technologies (Stoyanovich et al., 2020; Yeomans et al., 2019). At a more fundamental level, we might ask how uncertainty about machines' inner workings (i.e., lack of interpretability) affects our representations of machines and behavioural responses towards them when we interact with them in traditionally social settings. Already today, many among us are receiving advice on our life-choices from AI-powered tools (Zellers et al., 2021), entrusting algorithms with the managements of our material wealth (Patterson, 2012), or replacing human and animal companions with social robots (Henschel et al., 2021; Wykowska, 2020). The findings reported in this chapter provide an initial step towards addressing these questions and to ultimately not only develop but also broaden our understanding of how uncertainty shapes social decisions writ large.

5 Concluding discussion

Human decisions are fundamentally social, and the social world is fraught with uncertainty. Examples range from individuals' choices about complying with hygiene measures during a raging pandemic in the face of persisting unknowns about the measures' efficacy for preventing others from becoming infected; over surrogate medical choices between treatments bearing different risks for the patient; to policy decisions weighing daring and caution in choices spanning healthcare, the environment, or foreign relations among others. In my thesis, I sought to map out how social decision-making is shaped by uncertainty.

Summary of key findings

Starting out from the reference point of individual decisions under risk, in Chapter 2 I showed that participants systematically erred on the safe side in social decisions when the odds of a desirable outcome were low. Meanwhile, individual decisions more often endorsed the risky option when choosing the safe option was optimal in terms of expected value, to which social decisions were more attuned. Substantiating the proposal that enhanced uncertainty in social decisions might drive such precautionary preferences, I observed greater indecision and less confidence in choices for others than self, whereas motivational differences seemed to play only a minor role.

In Chapter 3, I sought to experimentally manipulate uncertainty about the impact of one's decision on other's well-being and to test whether this variety of *impact uncertainty* might drive a particular manifestation of precautionary social preferences whereby people are more averse to harming others than

themselves. Relative to a baseline and a low uncertainty condition, under high uncertainty I observed larger self-other-differences in two quintessential indicators of uncertainty: choices for others were slower and noisier relative to choices for self; more so under high impact uncertainty than in the other two conditions. However, these effects did not translate into greater harm aversion for others than self under high impact uncertainty. Overall, my observations suggest that enhanced uncertainty about others' experience plays a supporting role for hyperaltruistic valuations of pain. They also highlight the need to further characterise the concept and mechanisms of impact uncertainty to understand its role in hyperaltruism but also social decisions more broadly.

In Chapter 4, I considered uncertainty in social decisions defined more broadly. Specifically, as a first step towards understanding how uncertainty about machines' inner workings shapes our interactions with them, I showed that people value interpretability in AI. As such, my findings highlight that our fundamental need to explain others' minds and actions extends to interactions with machines. And paralleling our quest for answering the most basic question of all – "why?" (Pearl & Mackenzie, 2018) – in realms evading machines (Lombrozo, 2011), participants cared most about interpretability in situations involving high stakes and/or scarcity. In conjunction with emerging evidence that other aspects characterising human social cognition in traditional human-human interactions extend to human-machine interactions, my observations affirm the importance of understanding how uncertainty about machines' inner workings informs our interactions with them in social settings such as advice-sharing, trading, or companionship.

Together, these findings highlight uncertainty's pivotal role in shaping our interactions with other agents, may they be humans or machines. They also

allude to the manifold ways in which uncertainty comes into play. For instance, the empirical observations from Chapter 2 expose an intricate interplay between uncertainty over outcomes for others and uncertainty about how those outcomes might affect others. Chapter 3 points towards individual differences in other-regarding traits, such as empathy, modulating the effect of uncertainty in social decisions. And Chapter 4 reflects how uncertainty about other agents' minds might be deflected by other attributes about those agents. In the following, I reflect where my dissertation work has surfaced the intricate ways in which uncertainty shapes social decisions and where it leaves room for deeper insights.

Uncertainty's contribution to self-other differences in human decision-making

Even though self-regarding choices under uncertainty rank among the most studied facets in the science of decision-making, other-regarding choices under uncertainty remain poorly understood (Polman & Wu, 2020). As we begin to map out social decision-making under uncertainty, individual decisions under uncertainty provide a natural reference point; not least because this approach ties in with past decision-making research that has explored the motivational underpinnings of social preferences with a focus on the role of reward and punishment (Apps et al., 2016; Behrens et al., 2008; Fehr & Camerer, 2007; D. Lee & Seo, 2016; Ruff & Fehr, 2014). Paralleling this line of research, my findings from Chapters 2 and 3 show a substantial overlap between decisions for self and for others, while the degree of overlap varied in par with uncertainty-related parameters.

For instance, the results described in Chapter 2 suggest that choices for others are more attuned to probabilities in the lower ranges than choices for self, which in turn produces enhanced self-other differences in risky choices. Similarly, the results from the harm aversion study presented in Chapter 3 point towards analogous integration of choice parameters (i.e., difference in shocks and difference in money between options) into subjective values across decisions for self and decisions for others, while uncertainty about others' experience of choice-outcomes seemed to affect those parameters' weights differently across decisions for self and decisions for others at least to some extent.

These findings are compatible with existing evidence for individual and social decisions involving overlapping valuation mechanisms (Behrens et al., 2009; Olsson et al., 2020; Piva et al., 2019; Shenhav & Greene, 2010; Zaki & Mitchell, 2011), while they also suggest that uncertainty might have a dissociating effect on the underlying goal-functions: As hypothesised earlier, and as indirectly supported both by the present findings and previous work (Kameda et al., 2016; Kappes et al., 2018), uncertainty about others' experience might translate into a goal-function that represses potential harm to others by maximising the worst possible outcome, akin to a "maximin" strategy (Rawls, 1971). Study 2.3 in the present work indicated that incentivising participants to match their partner's choices slightly reduced self-other differences. However, the manipulation of the motivational goal in Study 2.3 did not eliminate the overall difference whereby choices for others were more risk-averse for low probabilities than choices for self, which might point towards an overarching maximin goal-function in social choices. It thus seems worthwhile to explore the goal-functions underlying decisions for self and for others under uncertainty in a more targeted way and

across different contexts (e.g., competitive versus cooperative). This would also contribute to the grander challenge of characterising social decision-making across different levels of explanation (Lockwood et al., 2020).

Uncertainty in interactions with non-human agents

Besides characterising social decision-making under uncertainty against the florid canvas of individual decision-making under uncertainty, comparisons with decisions concerning "non-social" conspecifics provide a useful perspective (Apps et al., 2013; Lockwood et al., 2020). Past research seizing this viewpoint has often turned to computers as the non-social comparison (see, e.g., V. Lee & Harris, 2013; Mitchell et al., 2005). However, with the emergence of social machines (Cross & Ramsey, 2021; Henschel et al., 2021), the non-social specificity of computer agents is becoming increasingly blurred – up to a degree where people struggle to correctly identify their counterpart as human or artificial (Aharoni & Fridlund, 2007; Alarifi et al., 2016; Crandall et al., 2018). Indeed, my findings from Chapter 4 suggest that the fundamental human need for explaining others' decisions (Lombrozo, 2011; Malle, 2006; Miller, 2019) translates into demand for interpretability in AI (Studies 4.1A/B – Study 4.2) and that it varies along dimensions that also are familiar from psychological research on human-centred decision-making (Studies 4.2-4.4; Brockner & Wiesenfeld, 1996; Lombrozo, 2006; Skitka et al., 2003).

As most people seem to perceive a lack of sentience in present-time machines (Bigman & Gray, 2018; Gray & Wegner, 2012), the locus of uncertainty scrutinised in the work I presented in Chapter 4 sat with machines' inner workings – that is, uncertainty about how they might arrive at a given decision. A considerable body of work has examined the equivalent in human-human

interactions, where uncertainty arises from our limited ability to predict others' decisions (FeldmanHall & Shenhav, 2019; Vives & FeldmanHall, 2018; Weber & Johnson, 2009). It is intriguing to ponder how strategic responses to uncertainty pertaining to others' decisions might further extend to human-machine interactions.

On the one hand, Studies 4.1A/B-4.2 imply that people experience uncertainty about machines' inner workings as aversive, which aligns with previous psychological research considering how people experience uncertainty in human-human interactions (FeldmanHall & Shenhav, 2019; Hsu et al., 2005; Kruglanski et al., 1991; Tversky & Fox, 1995). As suggested by this past work, people might readily embrace opportunities to reduce perceived uncertainty about their counterparts (FeldmanHall et al., 2016; Kagan, 1972). FeldmanHall and Shenhav (2019) have recently proposed three mechanisms which might help to reduce uncertainty about others: automatic inferences grounded in past knowledge and contextual clues pertaining to various characteristics of one's counterpart (e.g., appearance, group-membership); controlled inferences about unobservable internal states of one's counterpart (e.g., their intentional stance), which draw upon effortful perspective-taking; and feedback-based learning which might be informed by observing others or receiving second-hand information about them from another party. As past experiences and opportunities to observe artificial agents remain relatively limited for most people at present, it will be interesting to track how these paths to reducing uncertainty in social settings evolve over time, as human-machine interactions become more common. One conjecture grounded in research on social cognition would be that, as people gain more knowledge and experience in

interaction with artificial social agents, predictions about those agents will become more accurate (Tamir & Thornton, 2018).

Already now, engineers and computer scientists have recognised the need to cater people's need for reducing uncertainty about AI's inner workings. Besides technological solutions to make AI more interpretable (e.g., Doshi-Velez & Kim, 2017; Gilpin et al., 2019; Mittelstadt et al., 2019), strategies that mimic those displayed in human-human interactions, such as AI communicating uncertainty about its underlying model, are explored as avenues to address demand for interpretability and to ultimately enhance engagement with and trust in machines (Bhatt et al., 2021). Not at least in view of Studies 4.3-4.4 indicating an interdependence between AI interpretability and AI accuracy, it will be interesting to see how such displays of metacognitive properties in AI might shape people's willingness to rely on machines' recommendations, advice, or predictions.

On the other hand, it seems worthwhile to explore whether strategic responses to uncertainty in human-machine interactions will also include a readiness to exploit the opacity of AI when this allows to obscure personal responsibility (e.g., Dana et al., 2006; Haisley & Weber, 2010). Anecdotal evidence suggests this might be the case: when allocation decisions for vaccines against Covid-19 went awry, prioritising administrators before frontline healthcare workers, responsible officials readily blamed a "very complex algorithm" for the undesirable outcomes (Diamond, 2020). This algorithm, however, turned out to be a relatively simple, hand-coded rule-based formula (Guo & Hao, 2020). Hence, just as people exploit uncertainty to avoid feeling responsible about their decisions' outcomes for others (e.g., Dana et al., 2006; Haisley & Weber, 2010), they might exploit uncertainty about machines' inner workings to evade

responsibility for machine-assisted outcomes affecting others. As such, this example highlights the potential overlap between people's responses to uncertainty in contexts with other human agents and contexts with machine agents. This in turn reinforces the prospects of considering human-machine interactions as an emerging category of social interactions, as I did in Chapter 4 (Kahn et al., 2004; Rahwan et al., 2019; Severson & Carlson, 2010).

Varieties of uncertainty in social decision-making

Because I consider most varieties of uncertainty not specific to either the social or the individual domain, I refrained from using the term "social uncertainty" throughout this dissertation – at the cost of word count and probably also readability. Across the three empirical chapters, even if I did not directly contrast different varieties of uncertainty, I cut across different ones. Take, for instance, risky uncertainty about outcomes as featured in Chapter 2. This variety of uncertainty is probably the most studied one, both in individual decision-making and in the emerging literature on self-other differences in decision-making (Huettel et al., 2006; Platt & Huettel, 2008). Even though the integration of probability and magnitude for risky outcomes seems to be implemented similarly across social and individual choices (e.g., Ruff & Fehr, 2014; Shenhav & Greene, 2010), Studies 2.1-2.3 highlight that low probabilities seem to receive a different weight in choices for other relative to choices for self. My findings support the interpretation that such divergences are driven by an additional layer of uncertainty, which pertains to others' experience of outcomes and which might amplify varieties of uncertainty that are shared across choices for self and other (i.e., risky uncertainty about outcomes in Studies 2.3). But, as I will discuss along with other limitations further below, if I were to start my

doctoral research again, I would strive for a more incremental empirical approach to test this conjecture – for instance, by investigating how other varieties characterised in individual choices under uncertainty play out in social choices.

The pair of epistemic (knowable) and aleatory (unknowable) uncertainty discussed in the introductory section would be an interesting case, not at least because the inevitable uncertainty about others' minds and experiences might constitute a special case of aleatory uncertainty, or a kind of its own (Kappes et al., 2019). Following the approach from Fox and Ülkümen (2011; Ülkümen et al., 2016), who explored the psychology of the epistemic-aleatory distinction, we might consider cues from natural language to address this question: The authors find that likelihood statements such as "I believe it is fairly likely" signify assessments of aleatory uncertainty, whereas confidence statements such as "I'm fairly confident" signify epistemic uncertainty (Ülkümen et al., 2016). I would argue that communications about others' internal states – their minds and experiences – more naturally and more commonly relate to confidence statements, which might indicate that uncertainty about others' internal states is not merely a special case of aleatory uncertainty but a kind of its own.

Insofar as impact uncertainty, defined as not knowing how an event will impact the well-being of another person (Kappes et al., 2018), directly pertains to the subjectiveness of others' experience, it is specific to social decisions. However, Chapter 3 highlighted that some aspects of the concept remain elusive. For instance, alluding to impact uncertainty as I did in Chapter 3 seemed to not only affect representations about others' subjective experience, but also one's own experience, or the scale on which one's own and others' experiences are represented, or both (Study 3). In further characterising impact uncertainty

and determining its specificity to the social domain, it might be helpful to explore how people respond to allusions to impact uncertainty about social machine's inner states and whether responses vary along the ascription of higher-order mental capacities to artificial agents. As highlighted by this point, my thesis bears several limitations, which in turn highlight some of the avenues for future research.

Limitations and critical reflection

A cardinal problem that struck me while considering varieties of uncertainty is that there could be arbitrarily many – some arising from the format in which uncertainty presents itself (e.g., from actual experience or from description; Hertwig & Erev, 2009), others arising from situational features (e.g., from the environment's volatility; Behrens et al., 2007), and yet more from the characteristics of decision-makers (e.g., their ambiguity-tolerance; Vives & FeldmanHall, 2018). And of course, as the focal point of my thesis, uncertainties arising from our intricate interdependencies with others. In hindsight, I think that direct contrasts between different varieties (as, for instance, in FeldmanHall et al., 2016; Kappes et al., 2018) across social and individual decisions facilitate our ability to further define these varieties and their singular effects on social decisions. With the aim to characterise uncertainty in social decisions more broadly, it might also have been useful to adopt a comparative perspective in Chapter 4 by contrasting people's demand for interpretability in choices made by artificial versus human decision-makers.

My approach to uncertainty in social decisions was strongly influenced by work on impact uncertainty that I, together with my supervisors and colleagues, had started during my master's and first year of doctoral studies (Kappes et al.,

2018). In the excitement of our initial results on dissociable effects of outcome and impact uncertainty on social decisions, exploring impact uncertainty's role in harm aversion for others versus self seemed to be an imminent step. But – not at least in view of the resource-intensity of the lab study I presented in Chapter 3 – it might have been advisable to devise Study 3 at a slightly later point, after developing a more detailed characterisation of impact uncertainty. As elaborated in Chapter 3's discussion section, the results of this empirical work bear some inconclusiveness that might well arise from the operationalisation of impact uncertainty.

To facilitate consistency and ultimately interpretability across empirical chapters, it might have been helpful to employ the same experimental manipulation across projects, for instance by extending the line of studies in Chapter 2 by an experimental manipulation of impact uncertainty in social decisions. My scientific training had not included computational modelling approaches to decision-making, with which I first familiarised myself when I conducted the work reported in Chapter 3. Here, leaving my comfort zone helped me realize the potential of computational models, which I see not only in identifying and quantifying latent variables (such as a fundamental aversion to harm others) but also in their force to commit to testable predictions by means of a universal language. Were I to start my doctoral research again, I would seek to start from a formalisation of uncertainty's role in social decision-making and to apply it throughout the empirical projects (see *Future directions*).

There is only so much one can fit into a doctoral project, and along this reason I neglected repeated interactions with others and more generally social decisions that form a response to others' decisions. As I will discuss below, I think it will be important to address these limitations in future research, not at least

because this could speak to the role of uncertainty in social learning from both human and artificial social agents.

Across the three empirical projects, I gathered first-hand experience about drawbacks and shortcomings of both laboratory, and online experiments. The laboratory study presented in Chapter 3 made me painstakingly realise the importance of meticulous and consistent (co-)experimenter conduct; and that lack thereof results in the exclusion of participants, which implies not only a waste of resources but also a decrease in statistical power. Regarding online studies, the exclusion rates resulting from failed comprehension and attention checks forced me to deviate from my pre-registered sampling plan in Study 2.2. They also limited the number of decision-rounds (Chapters 2 and 4) or scenarios (Chapter 4) I was able to include, which in turn affected statistical power and possibilities to apply computational modelling approaches.

Future directions

In my research, I cut across different varieties of uncertainty and of social settings. A meticulous yet indispensable task will be to further broaden the considered range of varieties in a first step, and to further integrate insights into an overarching framework of social decision-making under uncertainty in a second step. As discussed earlier, the pair of epistemic and aleatory uncertainty makes for an interesting pair to explore in the context of social decisions, for instance in view of the question whether the unresolvable uncertainty about others' minds is perceived akin to aleatory uncertainty, or as a kind of its own. In terms of varieties of social decisions, considering repeated interactions might allow for more direct tests of the adaptive value of precautionary responses to uncertainty in this domain, and hence for addressing the computational level of

explanation following Marr (2010; Lockwood et al., 2020) for explaining social decision-making writ large.

Inferring others' minds, or inner workings in the case of machines, seems to be a potent and even integral response to uncertainty in decisions pertaining to others. Indeed, our understanding of such processes might amount to descriptions of how uncertainty is resolved in social decisions. Linking and integrating literatures on theory of mind, perspective-taking, or mentalising on the one hand, and empirical insights into people's responses to uncertainty in social decisions on the other hand, could bear a meaningful theoretical framework. This, in turn, could provide the basis for exploring whether neural signatures associated with inferences about others' internal states scale with (experimental manipulations of) the degree of uncertainty about others. A further step might be to compare these signatures in decisions under uncertainty involving different social counterfeits, in particular humans versus machines. These examinations would speak to the implementational level of explanation following Marr (2010; Lockwood et al., 2020) and further contribute to understanding social decision-making writ large.

To extend our understanding of the goal-functions governing social decision-making under uncertainty, addressing the computational and thus ultimate level of explanation following Marr (2010; Lockwood et al., 2020), it seems indispensable to consider not only different varieties of uncertainty and social agents, but also different varieties of contexts. If uncertainty in social decisions triggers fundamentally precautionary preferences, I would expect to observe variation across different contexts (e.g., reduced precaution in competitive relative to cooperative settings) but also adherence to the do-no-harm-rule at the overall level; similar to Study 2.3's findings where incentivised

social decisions still were more risk-averse than individual decisions, but to a lesser degree relative to non-incentivised social decisions.

In pursuing these avenues, it would be valuable to describe uncertainty in social decision-making by means of a formal (computational) model. Such a formalized framework would narrow interpretative wiggle room and facilitate falsifiable predictions (Crockett, 2016; Lockwood & Klein-Flügge, 2020). One option might be to start from an existing model (e.g., the entropy model proposed by Shannon, 1948) and to test whether uncertainty about others simply multiplies uncertainty from non-social sources, as suggested by (FeldmanHall & Shenhav, 2019); or whether social decisions under uncertainty are better described by a model where independent terms for social and non-social varieties of uncertainty may receive different weights. Ultimately, a formal modelling approach in combination with neuroimaging, that allows to quantify the influence of latent variables such as an intrinsic need to spare others from harm, might offer an unrivalled avenue to understand social decision-making across implementational, algorithmic, and computational levels.

Implications

Most decisions we make have outcomes not only for ourselves, but also (or exclusively) for others. The unknowns of others' inner states – encompassing their desires, intentions, and experiences – magnify other varieties of uncertainty. The work presented in my thesis suggests that people adopt precautionary preferences in response to uncertainty about others' inner states and that they seek to resolve such uncertainty also in settings involving artificial social agents.

Erring on the side of caution when in doubt about the impact of one's decision on others might have selective value, because such a response safeguards others' and ultimately collective welfare (Crockett et al., 2014; Kappes et al., 2018). Unconditional precautionary preferences in choices for others might also serve as a signal of trustworthiness (Jordan et al., 2016). Considering populations with deficits in social cognition, for instance psychopathic individuals or those with autism spectrum conditions, further highlights the adaptive value of precautionary responses to uncertainty pertaining to others, as well as the burden this uncertainty can impose. I hope that my doctoral research contributes to narrowing the gaps in our knowledge about social decision-making under uncertainty, and to broadening our understanding of its pivotal role for navigating the intricate interdependencies with others.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, *4*(3), 165–178. https://doi.org/10.1038/nrn1056

Aharoni, E., & Fridlund, A. J. (2007). Social reactions toward people vs. computers: How mere lables shape interactions. *Computers in Human Behavior*, *23*(5), 2175–2189. https://doi.org/10.1016/j.chb.2006.02.019

Ahrends, C., Bravo, F., Kringelbach, M. L., Vuust, P., & Rohrmeier, M. A. (2019). Pessimistic outcome expectancy does not explain ambiguity aversion in decision-making under uncertainty. *Scientific Reports*, *9*(1), 12177. https://doi.org/10.1038/s41598-019-48707-y

Alarifi, A., Alsaleh, M., & Al-Salman, A. (2016). Twitter turing test: Identifying social machines. *Information Sciences*, *372*, 332–346. https://doi.org/10.1016/j.ins.2016.08.036

Apps, M. A. J., Green, R., & Ramnani, N. (2013). Reinforcement learning signals in the anterior cingulate cortex code for others' false beliefs. *Neuroimage*, *64*(1), 1–9.

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social Cognition: Tracking the motivation of others. *Neuron*, *90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440.

Arrow, K. J., & Fisher, A. C. (1974). Environmental preservation, uncertainty, and irreversibility. *The Quarterly Journal of Economics*, *88*(2), 312–319. https://doi.org/10.2307/1883074

Atanasov, P. D. (2015). *Risk preferences in choices for self and others: Meta analysis and research directions* (SSRN Scholarly Paper ID 1682569). Social Science Research Network. https://papers.ssrn.com/abstract=1682569

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. *ArXiv [Cs]*. http://arxiv.org/abs/1707.07397

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2019). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*, 134–143. https://doi.org/10.1038/s41562-019-0762-8

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*(2), 290–302.

Batteux, E., Ferguson, E., & Tunney, R. J. (2017). Risk preferences in surrogate decision making. *Experimental Psychology*, *64*(4), 290–297. https://doi.org/10.1027/1618-3169/a000371

Batteux, E., Ferguson, E., & Tunney, R. J. (2020). Do we make decisions for other people based on our predictions of their preferences? Evidence from financial and medical scenarios involving risk. *Thinking & Reasoning*, *26*(2), 188–217. https://doi.org/10.1080/13546783.2019.1592779

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *Science*, *324*(5931), 1160–1164.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Beisswanger, A. H., Stone, E. R., Hupp, J. M., & Allgaier, L. (2003). Risk taking in relationships: Differences in deciding for oneself versus for a friend. *Basic and Applied Social Psychology*, *25*, 121–135.

Berkman, L. F., Glass, T., Brissette, I., & Seeman, T. E. (2000). From social integration to health: Durkheim in the new millennium. *Social Science & Medicine*, *51*(6), 843–857.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *ArXiv:2011.07586 [CS]*. http://arxiv.org/abs/2011.07586

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00085

Bolton, G. E., & Ockenfels, A. (2010). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment. *American Economic Review*, *100*, 628–633. https://doi.org/10.1257/aer.100.1.628

Bolton, G. E., Ockenfels, A., & Stauf, J. (2015). Social responsibility promotes conservative risk behavior. *European Economic Review*, *74*, 109–127.

Brockner, J., & Wiesenfeld, B. (2005). How, when, and why does outcome favorability interact with procedural fairness? In *Handbook of Organizational Justice* (J.

Greenberg, J. A. Colquitt, Eds., pp. 525–553). Lawrence Erlbaum Associates Publishers.

Brockner, J., & Wiesenfeld, B. M. (1996). An integrative framework for explaining reactions to decisions: Interactive effects of outcomes and procedures. *Psychological Bulletin*, *120*, 189–208.

Budescu, D. V., Rapoport, A., & Suleiman, R. (1990). Resource dilemmas with environmental uncertainty and asymmetric players. *European Journal of Social Psychology*, *20*(6), 475–487.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1037/14805-009

Busemeyer, J. R., & Diederich, A. (2014). Estimation and testing of computational psychological models. In *Neuroeconomics* (pp. 49–61). Elsevier.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., & Kunreuther, H. (1989). Decision processes for low probability events: Policy implications. *Journal of Policy Analysis and Management*, *8*, 565–592.

Camerer, C. F., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370.

Campbell, W. K., Bonacci, A. M., Shelton, J., Exline, J. J., & Bushman, B. J. (2004). Psychological entitlement: Interpersonal consequences and validation of a self-report measure. *Journal of Personality Assessment*, *83*(1), 29–45.

Capraro, V., & Kuilder, J. (2016). To know or not to know? Looking at payoffs signals selfish behavior, but it does not actually mean so. *Journal of Behavioral and Experimental Economics*, *65*, 79–84. https://doi.org/10.1016/j.socec.2016.08.005

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society Interface*, *18*(175), 20200880.

Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making, 13*(1), 99-111.

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PloS One*, *8*, e74434. https://doi.org/10.1371/journal.pone.0074434

Chaminade, T., & Okka, M. M. (2013). Comparing the effect of humanoid and human face for the spatial orientation of attention. *Frontiers in Neurorobotics*, *7*, 12.

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*, 103.

Chaminade, T., Zecca, M., Blakemore, S.-J., Takanishi, A., Frith, C. D., Micera, S., Dario, P., Rizzolatti, G., Gallese, V., & Umiltà, M. A. (2010). Brain response to a

humanoid robot in areas implicated in the perception of human emotional gestures. *PLOS ONE*, *5*, e11577. https://doi.org/10.1371/journal.pone.0011577

Charness, G., & Jackson, M. O. (2009). The role of responsibility in strategic risk-taking. *Journal of Economic Behavior & Organization*, *69*, 241–247. https://doi.org/10.1016/j.jebo.2008.10.006

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the limpact on study results. *Social Psychological and Personality Science*, *11*, 464–473. https://doi.org/10.1177/1948550619875149

Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J., & Schultz, W. (2009). Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience*, *29*, 12574–12583.

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

Crockett, M. J. (2016). How formal models can illuminate mechanisms of moral judgment and decision making. *Current Directions in Psychological Science*, *25*(2), 85–90.

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology*, *25*(18), 1852–1859.

Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in Cognitive Sciences*, *25*(3), 200–212. https://doi.org/10.1016/j.tics.2020.11.009

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180034. https://doi.org/10.1098/rstb.2018.0034

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion (Washington, D.C.)*, *12*(1), 2–7. https://doi.org/10.1037/a0025071

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Dana, J., Weber, R. A., & Kuang, J. X. (2006). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113.

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium Series*, 19–26.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*, 105–110. https://doi.org/10.1038/nn.3279

Dequech, D. (2004). Uncertainty: Individuals, institutions and technology. *Cambridge Journal of Economics*, *28*(3), 365–378. https://doi.org/10.1093/cje/28.3.365

Diamond, D. [@ddiamond]. (2020, December 19). *Facing angry doctors, Stanford official tries to explain why vaccine went to others instead. The algorithm "clearly didn't work," he says, as doctors boo + accuse him of lying. "Algorithms suck!" shouts one protester. "Fuck the algorithm," says another. (Video via tipster) [Tweet]*. Twitter. https://twitter.com/ddiamond/status/1340091749595815936

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv:1702.08608*. http://arxiv.org/abs/1702.08608

Edelson, M. G., Polania, R., Ruff, C. C., Fehr, E., & Hare, T. A. (2018). Computational and neurobiological foundations of leadership decisions. *Science*, *361*(6401), eaat0036. https://doi.org/10.1126/science.aat0036

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 643–669.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Epstein, L. G., & Schneider, M. (2008). Ambiguity, information quality, and asset pricing. *The Journal of Finance*, *63*(1), 197–228.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369.

Eriksen, K. W., & Kvaløy, O. (2010). Do financial advisors exhibit myopic loss aversion? *Financial Markets and Portfolio Management*, *24*(2), 159–170. https://doi.org/10.1007/s11408-009-0124-z

European Parliament, & European Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the*

free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

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

Falk, A., Becker, A., Dohmen, T. J., Huffman, D., & Sunde, U. (2016). *The preference survey module: A validated instrument for measuring risk, time, and social preferences* (SSRN Scholarly Paper ID 2725035). Social Science Research Network. https://papers.ssrn.com/abstract=2725035

Faro, D., & Rottenstreich, Y. (2006). Affect, empathy, and regressive mispredictions of others' preferences under risk. *Management Science*, *52*(4), 529–541.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

FeldmanHall, O., Glimcher, P., Baker, A. L., & Phelps, E. A. (2016). Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk. *Journal of Experimental Psychology: General*, *145*(10), 1255–1262. https://doi.org/10.1037/xge0000205

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. https://doi.org/10.1016/j.cognition.2012.02.001

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426. https://doi.org/10.1038/s41562-019-0590-x

Fernandez-Duque, D., & Wifall, T. (2007). Actor/observer asymmetry in risky decision making. *Judgment and Decision Making*, *2*(1), 1–8.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

Fischhoff, B. (1995). Risk perception and communication unplugged: Twenty years of process. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *15*(2), 137–145.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebøen, & H. Montgomery (Eds.), *Perspectives on thinking, judging, and decision making* (p. 14). Universitetsforlaget.

Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Furnham, A., & Marks, J. (2013). Tolerance of ambiguity: A review of the recent literature. *Psychology*, *4*(9), 717–728.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY*, *35*(4), 795–809. https://doi.org/10.1007/s00146-020-00977-1

Garcia-Retamero, R., & Galesic, M. (2012). Doc, what would you do if you were me? On self–other discrepancies in medical decision making. *Journal of Experimental Psychology: Applied*, *18*(1), 38–51. https://doi.org/10.1037/a0026018

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage*, *35*(4), 1674–1684. https://doi.org/10.1016/j.neuroimage.2007.02.003

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *ArXiv:1806.00069*. http://arxiv.org/abs/1806.00069

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gross, J., Faber, N., Kappes, A., Nussberger, A.-M., Cowen, P., Browning, M., Kahane, G., Savulescu, J., Crockett, M., & Dreu, C. D. (2021). *When helping is risky: The behavioral and neurobiological tradeoff of social and risk preferences*. PsyArXiv. https://doi.org/10.31234/osf.io/n4wqd

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable Artificial Intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Guo, E., & Hao, K. (2020, December 21). *This is the Stanford vaccine algorithm that left out frontline doctors*. MIT Technology Review. https://www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388.

Hacking, I. (2006 [1975]). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511817557

Hadar, L., & Fischer, I. (2008). Giving advice under uncertainty: What you do, what you should do, and what others think you do. *Journal of Economic Psychology*, *29*(5), 667–683. https://doi.org/10.1016/j.joep.2007.12.007

Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, *68*(2), 614–625. https://doi.org/10.1016/j.geb.2009.08.002

Hammond, P. J. (1990). Interpersonal comparisons of utility: Why and how they are and should be made. *EUI Working Paper ECO No. 90/3*, 2–65. EUI Working Paper ECO No. 90/3.

Han, P. K. J., Kobrin, S. C., Klein, W. M., Davis, W. W., Stefanek, M., & Taplin, S. H. (2007). Perceived ambiguity about screening mammography recommendations:

Association with future mammography uptake and perceptions. *Cancer Epidemiology and Prevention Biomarkers*, *16*(3), 458–466.

Han, P. K. J., Zikmund-Fisher, B. J., Duarte, C. W., Knaus, M., Black, A., Scherer, A. M., & Fagerlin, A. (2018). Communication of scientific uncertainty about a novel pandemic health threat: Ambiguity aversion and its mechanisms. *Journal of Health Communication*, *23*(5), 435–444. https://doi.org/10.1080/10810730.2018.1461961

Hardesty, L. (2016, October 27). Making computers explain themselves. *MIT News*. https://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, *63*(4), 309–321.

Heider, F. (1958). The naive analysis of action. In F. Heider (Ed.), *The psychology of interpersonal relations* (pp. 79–124). John Wiley & Sons Inc.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*(5972), 1480–1484. https://doi.org/10.1126/science.1182238

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

Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences*, *43*(6), 373–384. https://doi.org/10.1016/j.tins.2020.03.013

Henschel, A., Laban, G., & Cross, E. S. (2021). What makes a robot social? A review of social robots from science fiction to a home or hospital near you. *Current Robotics Reports*, *2*, 9–19. https://doi.org/10.1007/s43154-020-00035-0

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Hogarth, R. M., & Kunreuther, H. (1989). Risk, ambiguity, and insurance. *Journal of Risk and Uncertainty*, *2*(1), 5–35. https://doi.org/10.1007/BF00055709

Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, *10*(2), 227–237.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–257. https://doi.org/10.1006/obhd.1996.0077

Hsee, C. K., & Weber, E. U. (1997). A fundamental prediction error: Self–others discrepancies in risk preference. *Journal of Experimental Psychology: General*, *126*, 45–53.

Hsieh, F. Y., Lavori, P. W., Cohen, H. J., & Feussner, J. R. (2003). An overview of variance inflation factors for sample-size calculation. *Evaluation & the Health Professions*, *26*(3), 239–257. https://doi.org/10.1177/0163278703255230

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, *310*(5754), 1680–1683.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron*, *49*(5), 765–775.

Huppert, F. A., Baylis, N., Keverne, B., Helliwell, J. F., & Putnam, R. D. (2004). The social context of well–being. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1449), 1435–1446. https://doi.org/10.1098/rstb.2004.1522

Johnson, S. G., Bilovich, A., & Tuckett, D. (2020). Conviction narrative theory: A theory of choice under radical uncertainty. *PsyArXiv*. https://doi.org/10.31234/osf.io/urc96

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Jung, D., Sul, S., & Kim, H. (2013). Dissociable neural processes underlying risky decisions for self versus other. *Frontiers in Neuroscience*, *7*, 15.

Kagan, J. (1972). Motives and development. *Journal of Personality and Social Psychology*, *22*(1), 51–66.

Kahn, J. (2018, December 12). Artificial intelligence has some explaining to do. *Bloomberg Businessweek*. https://www.bloomberg.com/news/articles/2018-12-12/artificial-intelligence-has-some-explaining-to-do

Kahn, P. H., Freier, N. G., Friedman, B., Severson, R. L., & Feldman, E. N. (2004). Social and moral relationships with robotic others? *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, 545–550.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Kameda, T., Inukai, K., Higuchi, S., Ogawa, A., Kim, H., Matsuda, T., & Sakagami, M. (2016). Rawlsian maximin rule operates as a common cognitive anchor in

distributive justice and risky decisions. *Proceedings of the National Academy of Sciences*, *113*(42), 11817–11822.

Kappes, A., Nussberger, A.-M., Faber, N. S., Kahane, G., Savulescu, J., & Crockett, M. J. (2018). Uncertainty about the impact of social decisions increases prosocial behaviour. *Nature Human Behaviour*, *2*(8), 573–580. https://doi.org/10.1038/s41562-018-0372-x

Kappes, A., Nussberger, A.-M., Siegel, J. Z., Rutledge, R. B., & Crockett, M. J. (2019). Social uncertainty is heterogeneous and sometimes valuable. *Nature Human Behaviour*, *3*(8), 764–764. https://doi.org/10.1038/s41562-019-0662-y

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Keynes, J. M. (1921). *A treatise on probability*. Macmillan.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Houghton Mifflin.

Knight, W. (2017). *The dark secret at the heart of AI*. MIT Technology Review. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Koller, M. (2016). An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6), 1–24.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS One*, *3*(7), e2597.

Kruglanski, A. W., Peri, N., & Zakai, D. (1991). Interactive effects of need for closure and initial confidence on social information seeking. *Social Cognition*, *9*(2), 127–148.

Langer, E. J., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of 'placebic' information in interpersonal interaction. *Journal of Personality and Social Psychology*, *36*(6), 635–642.

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences*, *39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Lee, V., & Harris, L. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00259

Lerner, M. J., & Lerner, S. C. (2013). *The justice motive in social behavior: Adapting to times of scarcity and change*. Plenum.

Lind, E. A., & Tyler, T. R. (1988). *The Social Psychology of Procedural Justice*. Plenum.

Lipton, Z. C. (2017). The mythos of model interpretability. *ArXiv:1606.03490*. http://arxiv.org/abs/1606.03490

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is there a 'social' brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011

Lockwood, P. L., & Klein-Flügge, M. (2020). Computational modelling of social cognition and behaviour-a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsaa040

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267–286. https://doi.org/10.1037/0033-2909.127.2.267

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551.

Lu, Jingyi, Shang, X., & Li, B. (2018). Self–other differences in decision-making under risk. *Experimental Psychology*, *65*(4), 226–235. https://doi.org/10.1027/1618-3169/a000404

Lu, Joy, Lee, D. (DK), Kim, T. W., & Danks, D. (2019). Good explanation for algorithmic transparency. *ArXiv:10.2139*. https://papers.ssrn.com/abstract=3503603

Machina, M. J., & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica*, *60*(4), 745–780. https://doi.org/10.2307/2951565

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*(1), 130–156.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ArXiv Preprint ArXiv:1908.09635*.

Mengarelli, F., Moretti, L., Faralla, V., Vindras, P., & Sirigu, A. (2014). Economic decisions for others: An exception to loss aversion law. *PLOS ONE*, *9*, e85042.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Tavistock.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(1), 1–10. https://doi.org/10.1038/srep26094

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage*, *26*(1), 251–257. https://doi.org/10.1016/j.neuroimage.2005.01.031

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Conference on Fairness, Accountability, and Transparency*, 279–288. https://papers.ssrn.com/abstract=3278331

Montinari, N., & Rancan, M. (2013). *Social preferences under risk: The role of social distance* (No. 2013/90; RSCAS Working Papers). European University Institute. https://ideas.repec.org/p/rsc/rsceui/2013-90.html

Moorman, S. M., Hauser, R. M., & Carr, D. (2009). Do older adults know their spouses' end-of-life treatment preferences? *Research on Aging*, *31*(4), 463–491. https://doi.org/10.1177/0164027509333683

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised Assessment of Personality–Abbreviated Scale (SAPAS): Preliminary validation of a brief screen for personality disorder. *The British Journal of Psychiatry*, *183*(3), 228–232.

Mukerji, S., & Tallon, J.-M. (2001). Ambiguity aversion and incompleteness of financial markets. *The Review of Economic Studies*, *68*(4), 883–904.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Ogawa, A., Ueshima, A., Inukai, K., & Kameda, T. (2018). Deciding for others as a neutral party recruits risk-neutral perspective-taking: Model-based behavioral and fMRI experiments. *Scientific Reports*, *8*(1), 12857. https://doi.org/10.1038/s41598-018-31308-6

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212. https://doi.org/10.1038/s41583-020-0276-4

Oztop, E., Franklin, D. W., Chaminade, T., & Cheng, G. (2005). Human–humanoid interaction: Is a humanoid robot perceived as a human? *International Journal of Humanoid Robotics*, *2*(04), 537–559.

Pahlke, J., Strasser, S., & Vieider, F. M. (2015). Responsibility effects in decision making under risk. *Journal of Risk and Uncertainty*, *51*(2), 125–146.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Patil, I. (2018). *ggstatsplot: 'ggplot2' based plots with statistical details*.

Patt, A. G., & Weber, E. U. (2014). Perceptions and communication strategies for the many uncertainties relevant for climate policy. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(2), 219–232.

Patterson, S. (2012). *Dark pools: The rise of AI trading machines and the looming threat to Wall Street*. Random House.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Penguin Books.

Pérez-Escudero, A., Friedman, J., & Gore, J. (2016). Preferential interactions promote blind cooperation and informed defection. *Proceedings of the National Academy of Sciences*, *113*(49), 13995–14000.

Pilkington, E. (2019, October 14). Digital dystopia: How algorithms punish the poor. *The Guardian*. https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor

Piva, M., Velnoskey, K., Jia, R., Nair, A., Levy, I., & Chang, S. W. (2019). The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. *ELife*, *8*, e44939. PubMed. https://doi.org/10.7554/eLife.44939

Pizarro, D. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*(4), 355–375. https://doi.org/10.1111/1468-5914.00135

Platt, M. L., & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under uncertainty. *Nature Neuroscience*, *11*, 398–403.

Pollmann, M., Potters, J. J. M., & Trautmann, S. T. (2014). Risk taking by agents: The role of ex-ante and ex-post accountability. *Economics Letters*, *123*(3), 387–390. https://doi.org/10.1016/j.econlet.2014.04.004

Polman, E. (2012). Self–other decision making and loss aversion. *Organizational Behavior and Human Decision Processes*, *119*(2), 141–150.

Polman, E., & Wu, K. (2020). Decision making for others involving risk: A review and meta-analysis. *Decision Making for Others*, *77*, 102184. https://doi.org/10.1016/j.joep.2019.06.007

Press, C., Bird, G., Flach, R., & Heyes, C. (2005). Robotic movement elicits automatic imitation. *Cognitive Brain Research*, *25*(3), 632–640.

Press, C., Gillmeister, H., & Heyes, C. (2007). Sensorimotor experience enhances automatic imitation of robotic action. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1625), 2509–2514.

Price, W. N. (2018). Big data and black-box medical algorithms. *Science Translational Medicine*, *10*(471). https://doi.org/10.1126/scitranslmed.aao5333

Pulford, B. D. (2009). Short article: Is luck on my side? Optimism, pessimism, and ambiguity aversion. *Quarterly Journal of Experimental Psychology*, *62*(6), 1079–1087.

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine theory of mind. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 4218--4227). PMLR. http://proceedings.mlr.press/v80/rabinowitz18a.html

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., & Jackson, M. O. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). Kegan, Paul, Trench, Trubner & Co.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment*, *93*(1), 84–95.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *38*(1), 119–125. https://doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Rich, A. S., Rudin, C., Jacoby, D. M. P., Freeman, R., Wearn, O. R., Shevlin, H., Dihal, K., ÓhÉigeartaigh, S. S., Butcher, J., Lippi, M., Palka, P., Torroni, P., Wongvibulsin, S., Begoli, E., Schneider, G., Cave, S., Sloane, M., Moss, E., Rahwan, I., … Stilgoe, J. (2020, January 17). *AI reflections in 2019* [Collection of Articles]. Nature Machine Intelligence. https://doi.org/10.1038/s42256-019-0141-1

Rigoli, F., Preller, K. H., & Dolan, R. J. (2018). Risk preference and choice stochasticity during decisions for other people. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 331–341.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, *3*(4), 263–277.

Rohrer, J. M., Richter, D., Brümmer, M., Wagner, G. G., & Schmukle, S. C. (2018). Successfully striving for happiness: Socially engaged pursuits predict increases in life satisfaction. *Psychological Science*, *29*(8), 1291–1298. https://doi.org/10.1177/0956797618761660

Rowe, W. D. (1994). Understanding uncertainty. *Risk Analysis*, *14*(5), 743–750. https://doi.org/10.1111/j.1539-6924.1994.tb00284.x

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*, 549–562.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). Dover Publications.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation

of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*(6), 1063–1078.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, *23*(8), 1099–1103. https://doi.org/10.1016/j.neunet.2010.08.014

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125–130. https://doi.org/10.1177/0963721414553264

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy & Public Affairs*, *1*(3), 229–243.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.

Skitka, L. J., Winquist, J., & Hutchinson, S. (2003). Are outcome fairness and outcome favorability distinguishable psychological constructs? A meta-analytic review. *Social Justice Research*, *16*(4), 309–341.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., & Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PLOS ONE*, *1*(1), e39. https://doi.org/10.1371/journal.pone.0000039

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Steffel, M., Williams, E. F., & Perrmann-Graham, J. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes*, *135*, 32–44. https://doi.org/10.1016/j.obhdp.2016.04.006

Stone, E. R., & Allgaier, L. (2008). A social values analysis of self–other differences in decision making involving risk. *Basic and Applied Social Psychology*, *30*(2), 114–129.

Stone, E. R., Choi, Y., de Bruin, W. B., & Mandel, D. R. (2013). I can take the risk, but you should be safe: Self-other differences in situations involving physical safety. *Judgment and Decision Making*, *8*(3), 250–267.

Stoyanovich, J., Van Bavel, J. J., & West, T. V. (2020). The imperative of interpretable machines. *Nature Machine Intelligence*, *2*(4), 197–199. https://doi.org/10.1038/s42256-020-0171-8

Sun, Q., Jiang, T., Zhang, J., Wu, Q., Zhao, L., & Hu, F. (2020). Riskier for me or for others? The role of domain and probability in self-other differences, in risky decision-making. *The Journal of General Psychology*, *147*(2), 169–185. https://doi.org/10.1080/00221309.2019.1664388

Sun, Q., Liu, Y., Zhang, H., & Lu, J. (2017). Increased social distance makes people more risk-neutral. *The Journal of Social Psychology*, *157*(4), 502–512.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Teigen, K. H. (1994). Variants of subjective probabilities: Concepts, norms, and biases. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 211–238). John Wiley & Sons.

Tuckett, D., & Nikolic, M. (2017). The role of conviction and narrative in decision-making under radical uncertainty. *Theory & Psychology*, *27*(4), 501–523. https://doi.org/10.1177/0959354317713158

Tunney, R. J., & Ziegler, F. V. (2015). Toward a psychology of surrogate decision making. *Perspectives on Psychological Science*, *10*(6), 880–885.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*(s), 269–283.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology. General*, *145*, 1280–1297. https://doi.org/10.1037/xge0000202

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. http://www.stats.ox.ac.uk/pub/MASS4/

Vieider, F. M., Villegas-Palacio, C., Martinsson, P., & Mejía, M. (2016). Risk taking for oneself and others: A structural model approach. *Economic Inquiry*, *54*(2), 879–894. https://doi.org/10.1111/ecin.12290

Vives, M.-L., & FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature Communications*, *9*(1), 2156. https://doi.org/10.1038/s41467-018-04631-9

Vlaev, I., Seymour, B., Dolan, R. J., & Chater, N. (2009). The price of pain and the value of suffering. *Psychological Science*, *20*(3), 309–317. https://doi.org/10.1111/j.1467-9280.2009.02304.x

Vlaev, I., Wallace, B., Wright, N., Nicolle, A., Dolan, P., & Dolan, R. (2017). Other people's money: The role of reciprocity and social uncertainty in decisions for others. *Journal of Neuroscience, Psychology, and Economics*, *10*(2–3), 59–80. https://doi.org/10.1037/npe0000063

Volz, K. G., Schubotz, R. I., & Cramon, D. Y. von. (2005). Variants of uncertainty in decision-making and their neural correlates. *Brain Research Bulletin*, *67*(5), 403–412. https://doi.org/10.1016/j.brainresbull.2005.06.011

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, *114*(30), 7963–7968.

von Mises, R. (1939). *Probability, Statistics, and Truth*. Macmillan.

Waldrop, M. M. (2019). What are the limits of deep learning? *Proceedings of the National Academy of Sciences*, *116*(4), 1074–1077. https://doi.org/10.1073/pnas.1821594116

Walters, D. J., Ülkümen, G., Erner, C., Tannenbaum, D., & Fox, C. R. (2020). *Investment Behaviors Under Epistemic and Aleatory Uncertainty* (SSRN Scholarly Paper ID 3695316). Social Science Research Network. https://doi.org/10.2139/ssrn.3695316

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Weber, E. U., & Johnson, E. J. (2009). Chapter 10 - Decisions Under Uncertainty: Psychological, Economic, and Neuroeconomic Explanations of Risk Preference. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics* (pp. 127–144). Academic Press. https://doi.org/10.1016/B978-0-12-374176-9.00010-5

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. https://ggplot2.tidyverse.org

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PloS One*, *7*(9), e45391.

Williams, K. M., & Paulhus, D. L. (2004). Factor structure of the Self-Report Psychopathy scale (SRP-II) in non-forensic samples. *Personality and Individual Differences*, *37*(4), 765–778. https://doi.org/10.1016/j.paid.2003.11.004

Wykowska, A. (2020). Social robots to test flexibility of human social cognition. *International Journal of Social Robotics*, *12*(6), 1203–1211. https://doi.org/10.1007/s12369-020-00674-5

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

Wykowska, A., Chellali, R., Al-Amin, M. M., & Müller, H. J. (2014). Implications of robot actions for human perception. How do we represent actions of the observed robots? *International Journal of Social Robotics*, *6*(3), 357–366.

Xiong, W., Gao, X., He, Z., Yu, H., Liu, H., & Zhou, X. (2020). Affective evaluation of others' altruistic decisions under risk and ambiguity. *NeuroImage*, *218*, 116996. https://doi.org/10.1016/j.neuroimage.2020.116996

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Zaki, J., & Mitchell, J. P. (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences*, *108*(49), 19761–19766. https://doi.org/10.1073/pnas.1112324108

Zaleska, M., & Kogan, N. (1971). Level of risk selected by individuals and groups when deciding for self and for others. *Sociometry*, *34*(2), 198–213. https://doi.org/10.2307/2786410

Zellers, R., Holtzman, A., Clark, E., Qin, L., Farhadi, A., & Choi, Y. (2021). TuringAdvice: A generative and dynamic evaluation of language use. *ArXiv:2004.03607 [CS]*. http://arxiv.org/abs/2004.03607

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends* (SSRN Scholarly Paper ID 3312874). Center for the Governance of AI, Future of Humanity Institute, University of Oxford. https://papers.ssrn.com/abstract=3312874

Zhang, X., Li, S., Liu, Y., Chen, X., Shang, X., Qi, F., Wang, X., Guo, X., & Chen, J. (2019). Gain–loss situation modulates neural responses to self–other decision making under risk. *Scientific Reports*, *9*, 632. https://doi.org/10.1038/s41598-018-37236-9

Zhang, X., Liu, Y., Chen, X., Shang, X., & Liu, Y. (2017). Decisions for others are less risk-averse in the gain frame and less risk-seeking in the loss frame than decisions for the self. *Frontiers in Psychology*, *8*, 1601. https://doi.org/10.3389/fpsyg.2017.01601

Ziegelmeyer, F., & Ziegelmeyer, M. (2016). Parenting is risky business: Parental risk attitudes in small stakes decisions on behalf of their children. *Review of Economics of the Household*, *14*(3), 599–623. https://doi.org/10.1007/s11150-014-9245-x

Zittrain, J. (2019, July 23). *The Hidden Costs of Automated Thinking*. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking

Appendices

Appendix A: List of Figures

Figure 1. Schematic representation of the main task in Study 2.1 27

Figure 2. (**A**) Heatmaps for intentions and judgments across rounds and conditions. (**B**) Distributions of responses for intentions and judgments for individual and social choices aggregated across all rounds 29

Figure 3. Predicted values of intention to choose the risky option for (**A**) probability x condition and (**B**) magnitude x probability interactions of random intercept model from Study 2.1 33

Figure 4. Schematic representation of the main task in Study 2.2 39

Figure 5. Heatmaps for risky choices across conditions from Study 2.2 40

Figure 6. Heatmaps for risky choices across conditions from Study 2.3 51

Figure 7. Predicted values of risky choices for probability x condition interaction in Study 2.3 53

Figure 8. Predicted values for uncertainty ratings of probability x choice x condition interaction in random intercept model for Study 2.3 60

Figure 9. Main experimental task in Study 3 74

Figure 10. Rated similarity of shock experience in Study 3 80

Figure 11. Predicted values for harmful choices of condition x choice-target interaction in generalized regression model of Study 3 83

Figure 12. Distributions of and robust statistical tests for differences in harm aversion for self and other for the respective conditions in Study 3 .. 87

Figure 13. Joyplot visualising demand for interpretability across real-world AI applications 107

Figure 14. Barcharts illustrating main results for Study 4.2 116

Figure 15. Dependent variable and results for Studies 4.3 and 4.4 124

Appendix B: List of Tables

Table 1. Robust random intercept model predicting intentions (Study 2.1) 32

Table 2. Generalized random intercept model predicting choices (Study 2.3) 54

Table 3. Random intercept model predicting uncertainty ratings (Study 2.3) 59

Table 4. Generalized regression model predicting harmful choices (Study 3) 80

Table 5. Regression model predicting response times (Study 3) 85

Table 6. Exemplary AI application with its four versions according to the stakes / scarcity manipulation 114

Table 7. Random intercept model for interpretability ratings (Study 4.2) 115

Table A1. Random intercept model predicting uncertainty ratings (Study 2.3) 182

Table A2. Random intercept model predicting tradeoff ratings (Study 4.3) 186

Appendix C: Additional procedures and materials

Study 3

Set-up of decision rounds:

The set of decision-rounds was based on Crockett and colleagues (2014). As such, it included an initial set of 40 decision-rounds that were calibrated around the indifference points for the range of harm aversion values (κ) observed in earlier work. Both shocks and money were limited to vary between 0 and 20. Shocks always were positive integers, whereas the monetary amounts were positive numbers rounded to the nearest 10^{th}. While in this initial set the number of shocks and amount of money always were larger for the alternative relative to the default option, a "mirrored" set of decision-rounds was created where the alternative option always implied a decrease in shocks and money. With each round being presented twice, once in the "self" and once in the "other" condition, this yielded a total of 160 rounds. For more details, please refer to the Supplementary Information of the original work reported here by Crockett and colleagues (2014)."

Full list of questions:

After completing the harm aversion task, participants were asked about their perception, attitudes, and beliefs concerning various aspects of the decision task. In particular, I first asked in fixed order:

1. "How much would you dislike [1/5/10/20] shock[s] for YOU" (rating scale: not at all --- very much) in conjunction with "How confident are you about this rating?" (rating scale: not at all --- very much)

2. “How much would you dislike [1/5/10/20] shock[s] for THE RECEIVER” (rating scale: not at all --- very much) in conjunction with “How confident are you about this rating?” (rating scale: not at all --- very much)
3. “How much would THE RECEIVER dislike [1/5/10/20] shock[s] for THE RECEIVER” (rating scale: not at all --- very much) in conjunction with “How confident are you about this rating?” (rating scale: not at all --- very much)

Second, I asked in randomized order:

4. “To what extent do you think the Receiver’s experience of the shocks is SIMILAR to your own experience of the shocks? (rating scale: not at all similar --- very similar)
5. “Please consider your feeling during the ‘Receiver’ trials. How much CONTROL did you feel you had over the Receiver’s outcome?” (rating scale: no control --- total control)
6. “Please consider all of your choices during the ‘Receiver’ trials. In your opinion, what proportion of your choice HARMED the Receiver?” (rating scale: none were harmful --- all were harmful)
7. “How difficult did you find the choices when the shocks were for YOU?” (rating scale: not at all difficult --- very difficult)
8. “How difficult did you find the choices when the shocks were for THE RECEIVER?” (rating scale: not at all difficult --- very difficult)
9. “To what extent did you feel you faced a moral dilemma when making decisions that had consequences for the Receiver? (rating scale: not at all --- very much)

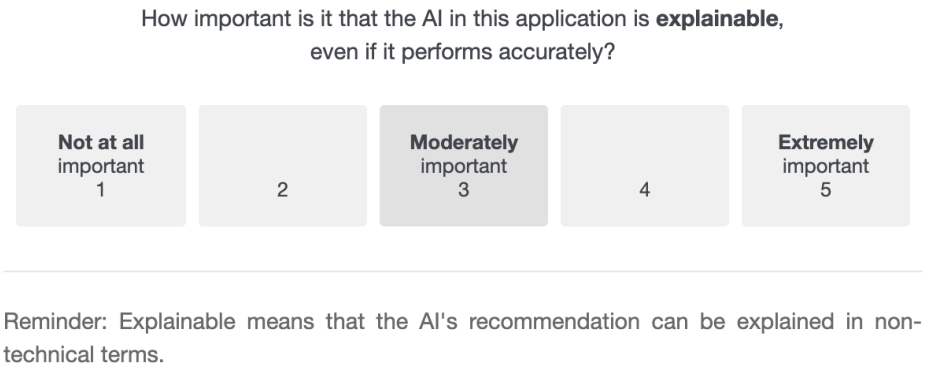
10. “Please consider all of your choices during the ‘Receiver’ trials. In your opinion, how SELFISH were your choices?” (rating scale: very unselfish --- very selfish)
11. “How much EMPATHY did you feel towards the Receiver?” (rating scale: no empathy --- strong empathy)
12. “How much did you IDENTIFY with the Receiver?” (rating scale: no at all --- very much)
13. “Please consider all of your choices during the ‘Receiver’ trials. How RESPONSIBLE did you feel for the Receiver’s outcome?” (rating scale: not responsible --- fully responsible)
14. “Please consider all of your choices during the ‘Receiver’ trials. How GUILTY do you feel about your choices on those trials?” (rating scale: not at all guilty --- very guilty)

And third, I asked again in fixed order:

15. “How unpleasant do you think a level 8 shock would be for YOU?” (rating scale: not at all unpleasant --- very unpleasant)
16. “How unpleasant do you think a level 8 shock would be for THE RECEIVER?” (rating scale: not at all unpleasant --- very unpleasant)
17. Charity donation task: participants were presented with an appeal to their sense of duty to aid the world’s poor by reference to Peter Singer’s (1972) pond analogy and subsequently asked what proportion of their task-based payment they would be willing to donate to ‘Give Directly’, a charity that distributes donations to extremely poor households in Kenya in a very effective way.

Study 4.1A

Descriptions of AI applications with *recommend* and *decide* versions. Note that the 'citizen surveillance' and 'virtual assistants' applications were only presented as *decide* version. For each application, participants submitted their answers on the following dependent variable:



AI Application	Recommend Version	Decide Version
Adaptive Education	AI recommends to a school teacher what learning goals to set for a given student.	AI sets on behalf of a school teacher the learning goals for a given student.
Distribution of Medical Treatment	AI recommends to a healthcare provider which patients should be given expensive and/or scarce medical treatments.	AI selects on behalf of a healthcare provider the patients who will be given expensive and/or scarce medical treatments.
Grading Essays	AI recommends to a teacher what grade to assign for a student's essay.	AI assigns a grade for a student's essay on behalf of a teacher.
Cyber Security	AI recommends to a computer user how to respond to a hacking attempt.	AI responds on behalf of a computer user to a hacking attempt.
Citizen Surveillance		AI surveils suspected criminal offenders through CCTV on behalf of law enforcement authorities.
Organizing Pictures	AI recommends to a photo assistant user how to organize	AI organizes pictures on behalf of a photo assistant user -- e.g., by faces depicted.

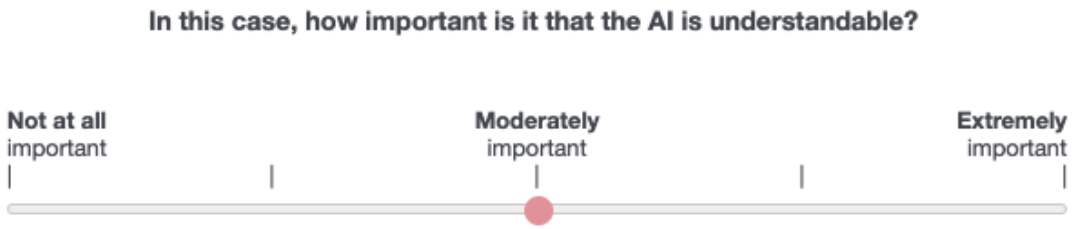
AI Application	Recommend Version	Decide Version
	their pictures -- e.g., by faces depicted.	
Financial Fraud Detection	AI recommends to fiscal authorities what firms might be evading taxes.	AI imposes audits on behalf of fiscal authorities on firms that might be evading taxes.
Processing of Financial Loans	AI recommends to a bank whether to approve or deny applications for financial loans or credit.	AI approves or denies applications for financial loans or credit on behalf of a bank.
Gaming	AI recommends optimized moves to human game-players.	AI executes optimized moves on behalf of human game-players.
Job Hiring	AI recommends to an employer the best candidate to hire for a job vacancy.	AI hires on behalf of an employer the best candidate for a job vacancy.
Historic Photograph Restoration	AI recommends to a photo lab technician how to repair historic photographs that are damaged.	AI repairs on behalf of a photo lab technician historic photographs that are damaged.
Improving Vegetables' Flavor	AI recommends to a greenhouse operator how to optimize the environment of vegetable plants to improve their flavor.	AI optimizes on behalf of a greenhouse operator the environment of vegetable plants to improve their flavor.
Insurance Pricing	AI recommends to an insurance company what insurance rate to set for a customer.	AI sets on behalf of an insurance company the insurance rate for a customer.
Legal Case Research	AI recommends to a lawyer which legal information -- e.g., court decisions or existing laws -- is relevant for a given legal case.	AI determines on behalf of a lawyer which legal information -- e.g., court decisions or existing laws -- is relevant for a given legal case.
Facebook Content Moderation	AI recommends to a Facebook employee what posts -- e.g., those containing nudity or terrorist propaganda -- to take down from the platform	AI takes down social media posts -- e.g., those containing nudity or terrorist propaganda -- on behalf of a Facebook employee.
Medical Diagnosing	AI recommends to a doctor what disease a patient might be suffering from.	AI establishes on behalf of a doctor what disease a patient might be suffering from.

AI Application	Recommend Version	Decide Version
Operation of Military Weapons	AI recommends to a military agent which targets to place under fire.	AI decides on behalf of a military agent which targets to place under fire.
Online Shopping	AI recommends products to a customer that they might need or enjoy.	AI buys products on behalf of a customer that they might need or enjoy.
Operating Private Cars	AI recommends to a human driver how to operate their car -- e.g., how to back their car into a parking spot.	AI operates a car on behalf of a human driver -- e.g., it backs their car into a parking spot.
Parole Review	AI recommends to a judicial officer whether or not to release an inmate from prison.	AI decides on behalf of a judicial officer whether or not to release an inmate from prison.
Political News Reporting	AI recommends to a journalist how data from a political event can be turned into a news report.	AI creates a news report on behalf of a journalist using data from a political event.
Romantic Partner Search	AI recommends compatible romantic partners to a dating app user.	AI selects compatible romantic partners on behalf of a dating app user.
Scientific Discoveries	AI recommends to scientists what experiments might lead to new discoveries -- e.g., the discovery of new materials.	AI runs on behalf of scientists experiments that might lead to new discoveries -- e.g., the discovery of new materials.
Speech Recognition-based Virtual Assistants		AI interacts with a person through speech-recognition-based virtual assistants such as Amazon's Alexa, Apple's Siri, or Google Home -- e.g., by turning on the music when the person tells them to do so.
Sports News Reporting	AI recommends to a journalist how data from a sports event can be turned into a news report.	AI creates a news report on behalf of a journalist using data from a sports event.
Processing of Unemployment Benefits	AI recommends to a civil servant what level of government assistance an unemployed person is eligible to receive.	AI determines on behalf of a civil servant what level of government assistance an unemployed person is eligible to receive.

AI Application	Recommend Version	Decide Version
Immigration Application Processing	AI recommends to a civil servant whether an immigration application should be approved or denied.	AI approves or denies an immigration application on behalf of a civil servant.

Study 4.1B

For each application, participants submitted their answers on a slightly modified dependent variable:



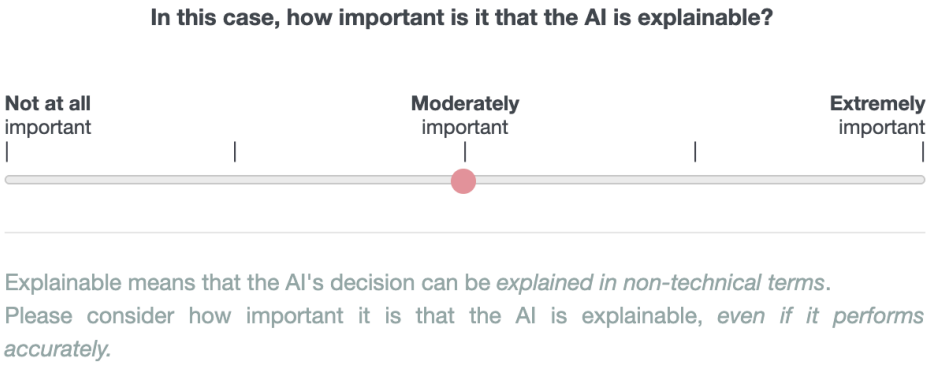
I used the same vignettes as in Study 4.1A, but only the *decide* versions. I added two additional applications:

Allocating standby flight passengers: AI decides which standby passengers are allowed to board an international flight departing in 30 minutes.

Prioritizing first responders for a hurricane: AI decides which locations within a U.S. county will be prioritized to receive aid from first responders -- e.g., police officers and paramedics -- when a hurricane strikes.

Study 4.2

I randomized the order of applications across participants and the order of application-versions both within and across participants. For each application and version, participants submitted their answers on the following dependent variable:



The following scenarios were presented in randomised order. Each version was presented on a separate screen, again in randomised order.

Vaccination allocation

It is flu season. An AI decides whether or not a citizen will get a vaccine.

	Low Stakes	**High Stakes**
Low Scarcity	In this case, the vaccine supply is abundant. Because the vaccine is very cheap, easy to produce, and can be stored at any temperature, the vaccine can be stocked in large quantities. This particular vaccine protects against a mild variant of the flu.	In this case, the vaccine supply is abundant. Because the vaccine is very cheap, easy to produce, and can be stored at any temperature, the vaccine can be stocked in large quantities. This particular vaccine protects against a deadly variant of the flu.
High Scarcity	In this case, the vaccine supply is very limited. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, the vaccine cannot be stocked in large quantities. This particular vaccine protects against a mild variant of the flu.	In this case, the vaccine supply is very limited. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, the vaccine cannot be stocked in large quantities. This particular vaccine protects against a deadly variant of the flu.

Allocating standby flight passengers

An international flight is departing in 30 minutes. An AI decides which standby passengers are allowed to board. Note: Standby passengers are passengers who do have a ticket for a flight to a particular destination, but they do not have a reservation for a specific flight.

	Low Stakes	**High Stakes**
Low Scarcity	In this case, there are 10 standby passengers, and 9 seats are available. The next flight to the same destination leaves one hour later and it has 20 available seats.	In this case, there are 10 standby passengers, and 9 seats are available. The next flight to the same destination does not leave until the following day and it has 20 available seats.
High Scarcity	In this case, there are 10 standby passengers, but only one seat is available. The next flight to the same destination leaves one hour later and it has 20 available seats.	In this case, there are 10 standby passengers, but only one seat is available. The next flight to the same destination does not leave until the following day and it has 20 available seats.

Prioritizing first responders for a hurricane

A U.S. county (population: 30,000) is preparing for a hurricane. An AI decides which locations within the county will be prioritized to receive aid from first responders, such as police officers, and paramedics, when the hurricane strikes.

	Low Stakes	**High Stakes**
Low Scarcity	In this case, the county's first responder services are more than adequately staffed. Therefore, first responders will be able to provide help to most locations within the county. The hurricane is expected to be minor, causing minimal damage across a number of locations.	In this case, the county's first responder services are more than adequately staffed. Therefore, first responders will be able to provide help to most locations within the county. The hurricane is expected to be severe, causing catastrophic damage across a number of locations.
High Scarcity	In this case, the county's first responder services are under-staffed. Therefore, first responders will only be able to provide help to very few locations within the county.	In this case, the county's first responder services are under-staffed. Therefore, first responders will only be able to provide help to very few locations within the county.

The hurricane is expected to be minor, causing minimal damage across a number of locations.	The hurricane is expected to be severe, causing catastrophic damage across a number of locations.

Reviewing insurance claims

An AI decides whether insurance claims get accepted or rejected.

	Low Stakes	**High Stakes**
Low Scarcity	In this case, a claim is being filed with an insurance company that has very lax criteria for accepting or rejecting a given claim. As a result, around 90% of insurance claims are approved. The insurance company reviews claims involving damages between $100 and $200.	In this case, a claim is being filed with an insurance company that has very lax criteria for accepting or rejecting a given claim. As a result, around 90% of insurance claims are approved. The insurance company reviews claims involving damages between $40,000 and $50,000.
High Scarcity	In this case, a claim is being filed with an insurance company that has very strict criteria for accepting or rejecting a given claim. As a result, around 10% of insurance claims are approved. The insurance company reviews claims involving damages between $100 and $200.	In this case, a claim is being filed with an insurance company that has very strict criteria for accepting or rejecting a given claim. As a result, around 10% of insurance claims are approved. The insurance company reviews claims involving damages between $40,000 and $50,000.

Hiring decisions

An AI makes hiring decisions on behalf of a global company.

	Low Stakes	**High Stakes**
Low Scarcity	In this case, the company is thriving and needs to expand, with plans to recruit additional staff in the near future. Hence, various branches of the company are recruiting for numerous positions. The positions are honorary, so a prospective employee would not expect to earn enough from this job to live on.	In this case, the company is thriving and needs to expand, with plans to recruit additional staff in the near future. Hence, various branches of the company are recruiting for numerous positions. The positions are salaried, so a prospective employee would expect to earn enough from this job to live on.

High Scarcity	In this case, the company is thriving and satisfied at its current size, with no plans to recruit additional staff in the near future. However, because a staff member retired, a local branch of the company is recruiting for a single position. The position is honorary, so a prospective employee would not expect to earn enough from this job to live on.	In this case, the company is thriving and satisfied at its current size, with no plans to recruit additional staff in the near future. However, because a staff member retired, a local branch of the company is recruiting for a single position. The position is salaried, so a prospective employee would expect to earn enough from this job to live on.

Study 4.3

I used the same applications descriptions as in Study 4.2. Please refer to *Figure 13A* in the main manuscript for the dependent variable, which differed from Study 4.2.

Study 4.4

The applications' descriptions were essentially the same as in Studies 4.2 and 4.3, although I added sentences that specified the relevant perspective:

Vaccination allocation

It is flu season. An AI decides whether or not a citizen will get a vaccine.

Patient: **You are a citizen seeking to get a vaccine. As a citizen, you are subject to the national healthcare system and thus to the AI making vaccination decisions.**

Agent: **You are the public health director. As public health director, you are responsible for managing the national healthcare system and thus for the AI making vaccination decisions.**

Low Stakes	**High Stakes**
In this case, the vaccine supply is abundant. Because the vaccine is very cheap, easy to produce, and can be stored at any temperature,	In this case, the vaccine supply is abundant. Because the vaccine is very cheap, easy to produce, and can be stored at any

Low Scarcity	the vaccine can be stocked in large quantities. This particular vaccine protects against a mild variant of the flu.	temperature, the vaccine can be stocked in large quantities. This particular vaccine protects against a deadly variant of the flu.
High Scarcity	In this case, the vaccine supply is very limited. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, the vaccine cannot be stocked in large quantities. This particular vaccine protects against a mild variant of the flu.	In this case, the vaccine supply is very limited. Because the vaccine is very expensive, laborious to produce, and must be stored at low temperatures, the vaccine cannot be stocked in large quantities. This particular vaccine protects against a deadly variant of the flu.

Allocating standby flight passengers

An international flight is departing in 30 minutes. An AI decides which standby passengers are allowed to board. Note: Standby passengers are passengers who do have a ticket for a flight to a particular destination, but they do not have a reservation for a specific flight.

Patient: **You are a standby passenger who wants to board the flight departing in 30 minutes. As a standby passenger, you are subject to the airline's management of standby passengers and thus to the AI making boarding decisions.**

Agent: **You are the airline director. As airline director, you are responsible for managing standby passengers and thus for the AI making boarding decisions.**

	Low Stakes	**High Stakes**
Low Scarcity	In this case, there are 10 standby passengers, and 9 seats are available. The next flight to the same destination leaves one hour later and it has 20 available seats.	In this case, there are 10 standby passengers, and 9 seats are available. The next flight to the same destination does not leave until the following day and it has 20 available seats.
High Scarcity	In this case, there are 10 standby passengers, but only one seat is available. The next flight to the same destination leaves one hour later and it has 20 available seats.	In this case, there are 10 standby passengers, but only one seat is available. The next flight to the same destination does not leave until the following day and it has 20 available seats.

Prioritizing first responders for a hurricane

A U.S. county (population: 30,000) is preparing for a hurricane. An AI decides which locations within the county will be prioritized to receive aid from first responders, such as police officers, and paramedics, when the hurricane strikes.

Patient: **You are a local resident hoping to be prioritized for receiving aid from first responders when the hurricane strikes. As a local resident, you are subject to the county-policy for managing first responders and thus to the AI making prioritization decisions.**

Agent: **You are the county's director of emergency services. As director of emergency services, you are responsible for managing first responders when the hurricane strikes and thus for the AI making prioritization decisions.**

	Low Stakes	**High Stakes**
Low Scarcity	In this case, the county's first responder services are more than adequately staffed. Therefore, first responders will be able to provide help to most locations within the county. The hurricane is expected to be minor, causing minimal damage across a number of locations.	In this case, the county's first responder services are more than adequately staffed. Therefore, first responders will be able to provide help to most locations within the county. The hurricane is expected to be severe, causing catastrophic damage across a number of locations.
High Scarcity	In this case, the county's first responder services are under-staffed. Therefore, first responders will only be able to provide help to very few locations within the county. The hurricane is expected to be minor, causing minimal damage across a number of locations.	In this case, the county's first responder services are under-staffed. Therefore, first responders will only be able to provide help to very few locations within the county. The hurricane is expected to be severe, causing catastrophic damage across a number of locations.

Reviewing insurance claims

An AI decides whether insurance claims get accepted or rejected.

Patient: **You are an insurance customer who has submitted a claim. As an insurance customer, you are subject to the insurance's management of submitted claims and thus to the AI making acceptance or rejection decisions.**

Agent: **You are the insurance company director. As insurance company director, you are responsible for managing submitted claims and thus for the AI making acceptance or rejection decisions.**

	Low Stakes	**High Stakes**
Low Scarcity	In this case, a claim is being filed with an insurance company that has very lax criteria for accepting or rejecting a given claim. As a result, around 90% of insurance claims are approved. The insurance company reviews claims involving damages between $100 and $200.	In this case, a claim is being filed with an insurance company that has very lax criteria for accepting or rejecting a given claim. As a result, around 90% of insurance claims are approved. The insurance company reviews claims involving damages between $40,000 and $50,000.
High Scarcity	In this case, a claim is being filed with an insurance company that has very strict criteria for accepting or rejecting a given claim. As a result, around 10% of insurance claims are approved. The insurance company reviews claims involving damages between $100 and $200.	In this case, a claim is being filed with an insurance company that has very strict criteria for accepting or rejecting a given claim. As a result, around 10% of insurance claims are approved. The insurance company reviews claims involving damages between $40,000 and $50,000.

Hiring decisions

An AI makes hiring decisions on behalf of a global company.

Patient: **You are a job candidate seeking to get hired by the company. As a job candidate, you are subject to the company's hiring management and thus to the AI making hiring decisions.**

Agent: **You are the company's recruitment director. As recruitment director, you are responsible for the company's hiring management and thus for the AI making hiring decisions.**

	Low Stakes	**High Stakes**
Low Scarcity	In this case, the company is thriving and needs to expand, with plans to recruit additional staff in the near future. Hence, various branches of the company are recruiting for numerous positions. The positions are honorary, so a prospective employee would not expect to earn enough from this job to live on.	In this case, the company is thriving and needs to expand, with plans to recruit additional staff in the near future. Hence, various branches of the company are recruiting for numerous positions. The positions are salaried, so a prospective employee would expect to earn enough from this job to live on.
High scarcity	In this case, the company is thriving and satisfied at its current size, with no plans to recruit additional staff in the near future. However, because a staff member retired, a local branch of the company is recruiting for a single position. The position is honorary, so a prospective employee would not expect to earn enough from this job to live on.	In this case, the company is thriving and satisfied at its current size, with no plans to recruit additional staff in the near future. However, because a staff member retired, a local branch of the company is recruiting for a single position. The position is salaried, so a prospective employee would expect to earn enough from this job to live on.

Appendix D: Additional results

Study 2.3

Table A1. Random intercept model predicting uncertainty ratings (Study 2.3)

Variables	*b* (*SE*)		*p*	*d*	95% CI_d
(Intercept)	4.99 (0.04)				
Choice	-0.41 (0.02)	<	.001	0.09	[0.07, 0.11]
Medium probability	-0.32 (0.02)	<	.001	0.07	[0.05, 0.09]
High probability	0.53 (0.02)	<	.001	0.10	[0.08, 0.12]
Magnitude	-0.09 (0.01)	<	.001	0.03	[0.01, 0.04]
Social baseline	0.28 (0.06)	<	.001	0.02	[-0.01, 0.05]
Social + match-motivation	0.05 (0.06)		.396	0.00	[-0.03, 0.04]
Social + incentivised match-motivation	0.09 (0.06)		.176	0.01	[-0.03, 0.04]
Medium probability x magnitude	0.08 (0.01)	<	.001	0.03	[0.01, 0.04]
High probability x magnitude	0.19 (0.02)	<	.001	0.05	[0.03, 0.06]
Medium probability x social baseline	0.03 (0.03)		.335	0.00	[-0.02, 0.03]
High probability x social baseline	-0.65 (0.04)	<	.001	0.07	[0.04, 0.09]
Medium probability x social + match-motivation	-0.01 (0.03)		.705	0.00	[-0.02, 0.03]
High probability x social + match-motivation	-0.62 (0.04)	<	.001	0.06	[0.04, 0.09]
Medium probability x social + incentivised match-motivation	-0.12 (0.03)	<	.001	0.02	[-0.01, 0.04]
High probability x social + incentivised match-motivation	-0.76 (0.05)	<	.001	0.07	[0.04, 0.1]
Magnitude x social baseline	-0.02 (0.02)		.246	0.01	[-0.01, 0.02]
Magnitude x social + match-motivation	-0.03 (0.02)		.137	0.01	[-0.01, 0.02]
Magnitude x social + incentivised match-motivation	-0.06 (0.02)		.002	0.01	[0.00, 0.03]
Choice x low probability x magnitude	-0.13 (0.02)	<	.001	0.03	[0.01, 0.05]
Choice x medium probability x magnitude	0.12 (0.02)	<	.001	0.03	[0.01, 0.05]
Choice x high probability x magnitude	0.00 (0.02)		.871	0.00	[-0.02, 0.02]

Choice x low probability x social baseline	-0.68 (0.05)	<	.001	0.07	[0.04, 0.09]
Choice x medium probability x social baseline	-0.40 (0.03)	<	.001	0.05	[0.03, 0.08]
Choice x high probability x social baseline	0.49 (0.04)	<	.001	0.05	[0.02, 0.08]
Choice x low probability x social + match-motivation	-0.49 (0.04)	<	.001	0.05	[0.02, 0.08]
Choice x medium probability x social + match-motivation	-0.12 (0.03)		.001	0.02	[-0.01, 0.04]
Choice x high probability x social + match-motivation	0.71 (0.04)	<	.001	0.07	[0.05, 0.1]
Choice x low probability x social + incentivised match-motivation	-0.58 (0.04)	<	.001	0.06	[0.03, 0.09]
Choice x medium probability x social + incentivised match-motivation	0.04 (0.04)		.303	0.00	[-0.02, 0.03]
Choice x high probability x social + incentivised match-motivation	0.92 (0.05)	<	.001	0.09	[0.06, 0.12]
Choice x magnitude x social baseline	0.00 (0.02)		.920	0.00	[-0.02, 0.02]
Choice x magnitude x social + match-motivation	0.05 (0.02)		.037	0.01	[-0.01, 0.03]
Choice x magnitude x social + incentivised match-motivation	0.15 (0.02)	<	.001	0.03	[0.01, 0.05]

Note. Estimates from random intercept model predicting uncertainty ratings by choice, outcome probability, outcome magnitude, condition, and their three-way interactions. Probability was clustered into three levels: low (5%, 25%), medium (45%, 50%, 55%), and high (75%, 95%).

Study 4.1A

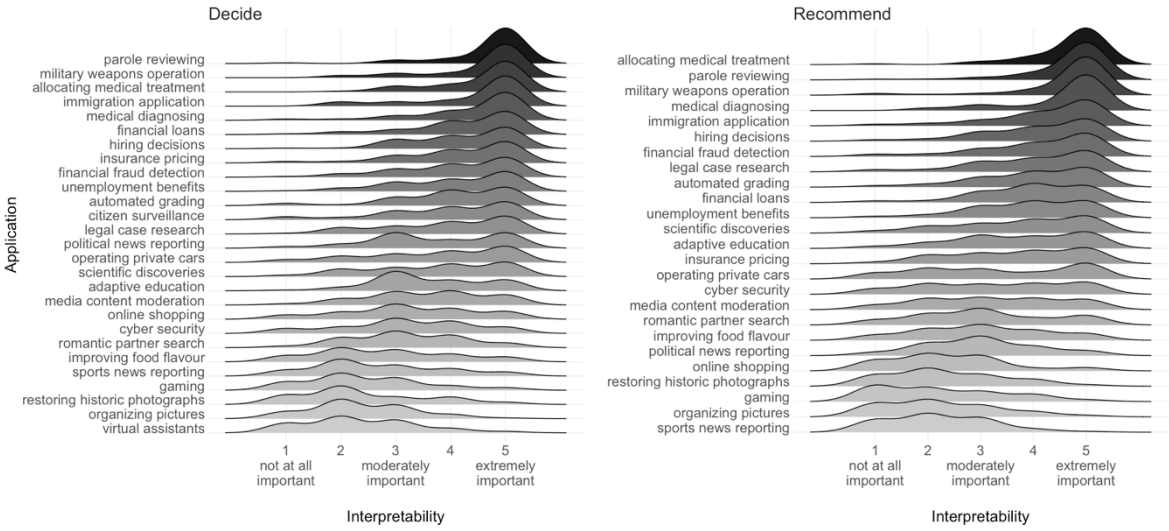


Figure A1. Demand for interpretability across between-subjects manipulation of *AI autonomy* (*decide* versus *recommend*). Joyplot visualizes the distributions of interpretability ratings and accordingly ranks applications from high to low demand for interpretability (top to bottom). Participants responded to the question "how important is it that the AI in this application is explainable, even if it performs accurately?" on a 5-point rating scale (1 = not at all important, 5 = extremely important).

Study 4.1B

US sample. I observed a significant main effect for stakes, $F(2, 7479) = 1139.70$, $p < .001$, $\eta^2 = 0.23$, 95% CI [0.22, 0.25]. Relative to applications involving low stakes, demand for interpretability was stronger for applications involving high ($b = 1.49$, $p < .001$, 95% CI [1.43, 1.55]) or medium stakes ($b = 0.93$, $p < .001$, 95% CI [0.86, 1.00]), and it was stronger amidst high than medium stakes, $b = 0.56$, $p < .001$, 95% CI [0.50, 0.62]. Similarly, a significant main effect for scarcity ($F(1, 7480) = 313.02$, $p < .001$, $\eta^2 = 0.04$, 95% CI [0.03, 0.05]) indicated increased demand for interpretability in applications potentially allocating scarce resources, relative to those that did not, $b = 0.58$, $p < .001$, 95% CI [0.51, 0.64].

UK sample. I observed a significant main effect for stakes, $F(2, 7131) = 1097.00$, $p < .001$, $\eta^2 = 0.24$, 95% CI [0.22, 0.25]. Relative to applications involving low stakes, demand for interpretability was stronger for applications involving high ($b = 1.52$, $p < .001$, 95% CI [1.45, 1.58]) or medium stakes ($b = 0.95$, $p < .001$, 95%

CI [0.88, 1.02]), and it was stronger amidst high than medium stakes, b = 0.57, p < .001, 95% CI [0.50, 0.63]. Similarly, a significant main effect for scarcity (F(1, 7132) = 271.50, p < .001, η^2 = 0.04, 95% CI [0.03, 0.04]) indicated increased demand for interpretability in applications potentially allocating scarce resources, relative to those that did not, b = 0.56, p < .001, 95% CI [0.50, 0.63].

Study 4.3

Model with explanatory controls. In order to explore further explanatory candidates, I asked participants the following additional questions once they had completed the main task; they provided answers for general versions of each application:

> *Reversibility:* "How reversible is a decision performed by the AI?"
>
> *Human expertise:* "What level of expertise would be required for a human to perform the decision instead of the AI?"
>
> *Personal affectedness:* "How likely is it that such a decision would affect you personally?"
>
> *Number of people affected:* "How many people will be affected by the decisions performed by the AI?"

When I added these explanatory variables to the random intercept model predicting preferences over accuracy versus interpretability, main effects for stakes and scarcity remained significant while I also observed a significant main effect for *human expertise,* whereby demand for accuracy over explainability was more pronounced in those applications that participants considered to require a high level of human expertise.

Table A2. Random intercept model predicting tradeoff ratings (Study 4.3)

Variables	*b* (SE)	*p*	*d*	95% CI_d
(Intercept)	0.85 (0.23)			
Stakes	-0.48 (0.11)	< .001	0.09	[0.00, 0.19]
Scarcity	-0.34 (0.11)	.002	0.06	[-0.03, 0.16]
Reversibility	0.04 (0.04)	.301	0.02	[-0.03, 0.08]
Human expertise	-0.28 (0.04)	< .001	0.13	[0.07, 0.19]
Personal affectedness	0.02 (0.04)	.657	0.01	[-0.05, 0.07]
Number of people affected	0.00 (0.03)	.877	0.00	[-0.04, 0.05]
Stakes x scarcity	0.28 (0.15)	.061	0.04	[-0.07, 0.15]

Note. Estimates from random intercept model predicting tradeoff ratings by stakes, scarcity, and explanatory variables.