COMPUTATIONAL MECHANISMS OF MORAL INFERENCE

JENIFER SIEGEL

Magdalen College

Dissertation Submitted for the Degree of

Doctor of Philosophy

of the

University of Oxford

ABSTRACT

Accurately inferring the moral character of others is crucial for avoiding social threats. Putatively "bad" agents command more attention and are identified more quickly and accurately than benign or friendly agents. Such vigilance is adaptive but can also be costly in environments where people sometimes make mistakes, because incorrectly attributing bad character to good people damages existing relationships and discourages forming new ones. Evolutionary models demonstrate that responding to wrongdoers with probabilistic forgiveness can facilitate the evolution of cooperation, but the cognitive mechanisms that enable the implementation of forgiving strategies are unknown. In this dissertation, I explore these mechanisms using novel methods derived from computational science, social cognition, and behavioral economics. Part I of the dissertation demonstrates that moral inference is described by an asymmetric Bayesian updating mechanism, where beliefs about the morality of bad agents are more uncertain (and therefore more amenable to updating) than beliefs about the morality of good agents. The model and data reveal a cognitive mechanism that rapidly discounts prior expectations to permit flexible updating of beliefs about potentially threatening others, a mechanism that could facilitate forgiveness when initial bad impressions turn out to be inaccurate. Part II of the dissertation considers the consequences that ensue when these mechanisms break down. Disruptions in learning and decision-making lie at the heart of many populations characterized by maladaptive social functioning. Consequently, I examine moral inference in two populations associated with interpersonal disturbances: individuals exposed to community violence and patients with Borderline Personality Disorder. The data reveal novel cognitive

processes that may explain the emergence of maladaptive behavior related to chronic exposure to violence and Borderline Personality Disorder. Collectively, the results in this dissertation provide insights into the computational mechanisms of moral inference and their role in adaptive and maladaptive social functioning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

PREFACE

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, mentor, and friend, Molly Crockett. Not only did she spark my interest in research to begin with, but it was Molly who first encouraged me to apply to Oxford to complete a Doctorate degree. When I started on this path, I had never been confident in my scholarly abilities and was certain that the chance that I would be accepted to such a prestigious institute was near impossible. Molly, however, had every faith in me. Her constant support and encouragement pushed me to strive for heights I never knew I could reach, ultimately leading me on a journey that brought me to unimaginable accomplishments. To write that she has taught me hundreds of lessons would undershoot the true number by a lightyear. Everything that I know about science, from good practice, to coding, to computational modelling, to navigating academic politics, to being an encouraging, supporting mentor, I learned from her. I cannot thank Molly enough for teaching me the complex demands of day-to-day research without losing site of the big questions, for allowing me to embark on research that is truly exciting and important to me, and for supporting me in ways that shine a light on her compassion and aptitude as a mentor.

None of the work presented in this dissertation would have been possible without the valuable contributions and support of my astounding colleagues. While I had no formal secondary advisor for my Doctorate, I found my colleagues and mentors Robb Rutledge and Christoph Mathys often taking on that role. I cannot thank you both enough for sharing your time, your advice, and your consummate professionalism over the past seven years.

You have each made both specific and global contributions to the scientist and person I am today. Thank you as well to Kate Saunders for nurturing my interest in studying social functioning in mental illness and guiding me through the ins and outs of clinical research. I also extend my greatest thanks to Arielle Baskin-Sommers whose technical advice and training helped me navigate new research interests throughout my Doctorate. I am very grateful to have such kind and brilliant colleagues to turn to for inspiration, guidance, and wisdom. To acknowledge the immeasurable contributions of my wonderful supervisor and collaborators, I have written the empirical chapters of this dissertation in first-person plural.

Many brilliant colleagues contributed valuable feedback to this work and inspired many of the reported studies. I thank Tim Behrens, Geoff Bird, Matthew Rushworth, Christopher Summerfield, Erie Boorman, Andreas Kappes, and Donald Carlston whose ideas have inspired me immensely.

Thank you to the undergraduate and master students I had the privilege of working with, especially Miriam Gerber, Mary Montgomery, Talia Longhorn, Eloise Copland, and Cassie Popham. I would also like to extend a special thanks to Tatyiana Tyurkina and Lucius Caviola for developing the web application utilized for data collection in several studies reported in this dissertation.

The two thousand or so volunteers who participated in my research deserve much recognition and thanks. I am especially grateful for the volunteers from the Oxford Special Needs Services in Oxford and the Cheshire Correctional Institute in Connecticut. Without you I would not have been able to address the truly unique and important questions that I hope will ultimately inspire future research to help those who suffer from interpersonal difficulties. On that note I would also like to thank those affiliated with the Cheshire

Correctional Institute and the Oxford Special Needs Services for their support of this research.

The research carried out in this dissertation was funded by a Wellcome Trust ISSF award and the Academy of Medical Science. Some of the work was also funded by the John Fell Fund. I am grateful to the Oxford Clarendon Fund and the Wellcome Trust for funding my D.Phil studies.

I thank my lab mates at Oxford and Yale for putting up with my incessant sporadic behavior and long-ended debates. I am particularly grateful to Anne-Marie Nussberger and Hongbo Yu for our evening discussions over pints that were vital to my sanity, and Vlad Chituc and Ryan Carlson for never ever agreeing with me (and therefore nurturing my love for debate). Finally, I thank my family, close friends, and the general community of Magdalen College who provided me emotional support and the occasional much-needed reprieve from science, most especially Irving and Isabel Siegel, Jessica Bird, Clara Colombato, Lindsay Karkheck, and Chay Kraft.

PUBLICATIONS ARISING FROM THE DISSERTATION WORK

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750.

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature communications*, *10*(1), 1942.

Chapter 1

1 INTRODUCTION

1.1 FORMING SUBJECTIVE IMPRESSIONS

Georgeann Hawkings was a female student at the University of Washington in 1974. As she was walking home from a study session with her friends one day, she was approached by a charming man with a cast on his arm. The man asked Georgeann for her assistance loading his briefcase into his car due to his injury; she kindly assisted him. Georgeann was never seen again after that night. The man feigning injury was named Ted Bundy, and Georgeann is one of what is thought to be over 100 women that Ted Bundy brutally murdered, raped and dismembered before his death in the electric chair in 1989.

Fortunately, most people that we encounter in our daily lives are not like Ted Bundy. However, the ability to accurately infer the moral character of others is nonetheless crucial for successful social functioning. Through recurrent encounters we must learn whether someone's intentions are good or bad, that is, whether they represent an opportunity or a threat. Still, as Ted Bundy demonstrated, people can undoubtedly present a public persona that appears to be virtuous, while simultaneously nurturing their darker desires. This chapter will review the faculties designed to optimize detection of these "villains" in

disguise, their consequences for the ability to build trusting relationships with those who are truly good, and potential mechanisms for balancing these two motivations (i.e., to avoid being harmed and build healthy relationships) in moral inference to support adaptive social functioning.

1.1.1 THE PREDOMINANCE OF MORALITY IN PERSON PERCEPTION

Person perception refers to an element of social cognition concerning how we learn about other people. In a way, person perception is not really about perception, per se, but more about the processing of social information to make inferences and form impressions. It is well-documented that morality predominates in person perception, forming the primary basis for the global evaluation of others. Many studies reveal that our evaluations of others depend to a higher degree on moral traits (e.g., trustworthiness, kindness, honesty) than non-moral traits (e.g., intelligence, friendliness, courageousness) (Wojciszke, 2005; Wojciszke, Bazinska, & Jaworski, 1998). When asked to judge a familiar or novel other, individuals' impressions are more strongly predicted by the target's moral traits than traits indicative of competence or sociability (Goodwin, Piazza, & Rozin, 2014). Across different types of relationships (e.g., family members, employees) trustworthiness is rated as the most desirable trait for an ideal person to possess, whereas other traits, such as intelligence and conscientiousness, are only valued to the extent that they are relevant to the nature of the relationship (Cottrell, Neuberg, & Li, 2007). Thus, it is no surprise that when forming social impressions, we are especially motivated to learn moral trait information. Moral traits are processed more rapidly than traits indicative of competence (Willis & Todorov, 2006; Ybarra, Chan, & Park, 2001), and when seeking out information about others, people

are more interested in acquiring information indicative of the target's morality than their competence (Brambilla, Rusconi, Sacchi, & Cherubini, 2011; De Bruin & van Lange, 2000; Wojciszke et al., 1998). This work substantiates a disproportionate role for moral traits in forming global impressions of others and suggests that people preferentially seek out moral information in their endeavors to make such impressions.

Given that one primary function of impression formation is to identify and avoid potential threats (Wojciszke et al., 1998), it is not surprising that moral traits forms the basis for our global evaluations of others. Moral traits signal the extent to which an individual cares about another's wellbeing and therefore have particular value for inferring whether someone intends us harm (Fiske, Cuddy, & Glick, 2007). From an evolutionary perspective, the primacy of moral traits is therefore apt because learning whether someone has good or bad intentions (i.e., whether they mean to help or harm us) is more important for survival than whether they can carry out those intentions. Accordingly, Fiske et al. (Fiske et al., 2007) have argued that moral information is sought out first to anticipate whether someone represents an opportunity or a threat, and only then is competence evaluated to determine *how* valuable or threatening they may be.

A growing body of evidence suggests that we infer moral character rapidly and effortlessly (Engell, Haxby, & Todorov, 2007; Goodwin et al., 2014; Todorov, Said, Engell, & Oosterhof, 2008). Adults make stable trustworthiness appraisals after merely 100 milliseconds exposure to novel faces (Willis & Todorov, 2006), and face trustworthiness influences subsequent moral evaluations even when faces preclude conscious awareness (Todorov, Pakrashi, & Oosterhof, 2009). People readily attribute good and bad moral character to others based on a single piece of behavioral information (Inbar, Pizarro, &

Cushman, 2012; Mende-Siedlecki, Cai, & Todorov, 2012; Todorov & Uleman, 2003). For example, financially benefitting from others' suffering is not only judged as a blameworthy act, but those who engage in these behaviors are also thought to possess a bad moral character (Inbar et al., 2012). Such rapid assessments of moral character are likely to be especially important for identifying the immediate goals of an actor and predicting what they will do in the future.

Aside from helping us predict others' behaviour, moral impressions also guide our own behavior; people will approach those whom they infer to have a good moral character and avoid those whom they infer to have a bad moral character (Brambilla, Sacchi, Pagliaro, & Ellemers, 2013; Pagliaro, Brambilla, Sacchi, D'Angelo, & Ellemers, 2013). Even implicit social cues, such as the perceived trustworthiness of a face, impact subsequent social decisions in behavioral economic games. For instance, people invest significantly more money with strangers represented by faces that have previously been rated as more trustworthy, despite no objective relationship between appearance and the return on investment (van 't Wout & Sanfey, 2008). This suggests that we depend on our ability to assess others' morality from social cues to help reap benefits from social interactions and avoid exploitation. Together, the research highlights the importance of building accurate representations of others' morality; inaccurately inferring malintent in kind individuals marks missed opportunities, while inaccurately inferring good intent in self-serving individuals exposes us to manipulation and harm.

1.1.2 A NEGATIVITY BIAS IN IMPRESSION UPDATING

Human's ability to automatically judge others' moral character from minimal information is undeniable (Inbar et al., 2012; Mende-Siedlecki et al., 2012; Todorov &

Uleman, 2003), but to adaptively navigate our social environment we must continually and flexibly update our impressions of others. Consider a time when you may have drastically changed your mind about someone. Perhaps this is a romantic partner who cheated on you, or a student who you once believed had great research potential but soon came to realize that they lacked the drive, skill, or loyalty to remain in your good graces. Changing your impression not only served the purpose of updating your expectations about their future actions but it allowed you to adapt your own decisions given those new expectations. For example, learning that your student has gossiped about fellow principal investigators in the department might not only motivate you to update your impression of their trustworthiness but also teach you to "play your cards closer to your chest", so to speak.

A growing body of research on impression updating reveals that our moral impressions of others are certainly amenable to updating in light of new evidence (Cone & Ferguson, 2015; Mende-Siedlecki et al., 2012; Reeder & Coovert, 1986; Skowronski & Carlston, 1992). To study impression updating, researchers have typically employed lists of traits (e.g., energetic, persuasive, cold) or phrases describing people's social behaviours (e.g., "volunteered to stay late to help a coworker" and "kicked a stray cat to get it to leave his yard"). Participants are asked to provide an impression of the target's character after being presented with a subset of the list, and then asked to provide another impression rating after being presented with another subset that is either congruent or incongruent in valence with the early behaviours. Updating is then quantified as the difference between the initial impression rating and subsequent impression rating. Work using these paradigms has made great strides towards our understanding of how impressions are updated. For example, Asch (Asch, 1946) raised the idea that the overall impression of an individual is

not equal to the algebraic sum of its components. That is, each piece of information does not hold an equal weight towards updating an impression (N. H. Anderson, 1965).

Consistent with this notion, one of the most robust phenomena in psychological science is the tendency for immoral actions to carry greater weight towards updating an impression than moral actions (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). This is commonly known as the *negativity bias* in impression formation. In an early demonstration (Reeder & Coovert, 1986), participants read a subset of either highly moral or highly immoral behaviours describing a target individual (e.g., "Donated time to take blind children to the park" and "Hit a child for no reason"). Next, participants rated their impression of the target from 1 (*highly immoral*) to 9 (*highly moral*). Lastly, participants read a final behaviour in which the valence was incongruent with the initial subset of behaviours and provided a new impression based on the combined information. The authors found that impressions were more heavily updated when immoral information followed moral information than the other way around. This study, along with many others (Briscoe, Woodyard, & Shaw, 1967; Cone & Ferguson, 2015; Dibbets, Pauline et al., 2012; Freedman & Steinbruner, 1964; Martijn, Spears, Van Der Pligt, & Jakobs, 1992; Riskey & Birnbaum, 1974), suggest that initial positive character impressions are more amenable to updating when presented with disconfirming evidence than negative character impressions. What might explain the emergence of a negative bias in updating moral impressions?

One of the most influential frameworks posits that people may consider negative moral information to be more diagnostic about an individual's character than positive moral information (Skowronski & Carlston, 1987, 1989). Skowronski & Calston's

hypothesis stems from the fact that not all actions are equally informative about a person's character (Fiske, 1980). That is, there is a probabilistic relationship between actions and their causes; bad people often act morally and immorally, but good people rarely act immorally. Hence, we define others by their worst actions because their best actions are less useful for determining someone's character and predicting their behaviour. This heuristic may be useful for building expectations about the probability and magnitude of possible harms. Knowing someone's "limit" – i.e., the line they *will not cross*, morally speaking – enables us to more easily and confidently predict how they will behave in a larger number of scenarios. This is consistent with research showing that people are more confident that someone who behaves selfishly is bad and will continue to behave immorally than someone who behaves friendly is good and will continue to behave morally (Martijn et al., 1992). Indeed, if the Ted Bundy case taught us anything, it would be that bad people can occasionally (or even often) appear good.

1.1.3 REVISING NEGATIVE MORAL IMPRESSIONS

Biasing immoral information in social evaluations is theorized to hold adaptive motivational value (Cacioppo, Gardner, & Berntson, 1997; Vaish, Grossmann, & Woodward, 2008): those that were more attuned to cues indicative of bad character would be more likely to survive due to their advantage over avoiding harmful outcomes. It is no doubt that evolution has endowed us with a strong motivation to avoid being harmed by others. However, the tendency to form inflexible negative impressions runs some risks of its own. Defining others morality from minor transgressions leaves little room for people to make mistakes and can hinder the development of secure trusting relationships. For instance, erroneously inferring bad character can lead people to prematurely terminate

valuable relationships and thereby miss out on the potential benefits of future cooperative interactions (Axelrod, 2006; Johnson, Blumstein, Fowler, & Haselton, 2013; McCullough, 2008; Molander, 1985). Thus, successfully navigating social life requires strategies for maintaining social relationships even when others behave inconsistently and sometimes commit immoral acts.

One possible strategy is to respond to defection with probabilistic cooperation (Nowak & Sigmund, 1992). Evolutionary models show such "generous" strategies outcompete strategies that summarily end cooperative relationships in the face of a single betrayal (Fudenberg, Rand, & Dreber, 2012; Wu & Axelrod, 1995). Generous strategies are also observed in humans playing repeated prisoner's dilemmas where others' intended actions are implemented with noise (Fudenberg et al., 2012). Thus, it appears that the most successful strategies for repeated interactions in a world of uncertainty and noise are ones that have an element of forgiveness and leniency (Fudenberg et al., 2012; Rand, Ohtsuki, & Nowak, 2009).

Forgiveness can be understood as the positive amendment of thoughts, behaviour, or emotions towards a transgressor (McCullough, 2000; McCullough, Pargament, & Thoresen, 2000; Snyder & Lopez, 2001). Because the negativity bias posits that negative impression are especially resistant to updating, a negativity bias naturally leads to the intuition that people should be highly unlikely to forgive wrongdoers. Yet it turns out that forgiveness is not only abundant in social interactions but also important for an individual's mental and physical health (T. W. Baskin & Enright, 2004; Worthington & Scherer, 2004). Research indicates that the propensity to forgive others is related to fewer symptoms of depression, decreased anxiety, and lower blood pressure (Brown, 2003; Krause & Ellison,

2003; Maltby et al., 2001; Witvliet, Ludwig, & Laan, 2001). Difficulty forgiving others' minor transgressions is commonly observed in mental illnesses characterized by interpersonal dysfunction (Barnow et al., 2009; Unoka, Seres, Áspán, Bódi, & Kéri, 2009). Patients with Borderline Personality Disorder often hold grudges and write people off following seemingly insignificant slights (Sansone, Kelley, & Forbis, 2013; Thielmann, Hilbig, & Niedtfeld, 2014). Chronic unforgiveness is therefore believed to be a key component leading to abnormal social cognition and behaviour in patients with Borderline Personality Disorder (Gartner, 1988; Holm, Berg, & Severinsson, 2009; Sansone et al., 2013), leading to the integration of forgiveness skills in standard treatments (Sandage et al., 2015). Together, the findings suggest that the ability to update negative beliefs about others may be a necessary component for healthy social functioning. But how do we make sense of a strong propensity to forgive alongside the negativity bias?

The bulk of evidence for the negativity bias in impression formation has employed very similar paradigms, with each study holding its own set of methodological limitations. In many cases, impression formation was examined using narrative descriptions of extreme and rare behaviours, such as theft and violence. For instance, Reeder and Coovert (Reeder & Coovert, 1986) used phrased such as "Donated time to take blind children to the park" and "Hit a child for no reason". Not only is it unlikely for a person to encounter one of these behaviours in a real social interaction partner, but to encounter both behaviours *within the same individual* is almost unbelievable. No wonder why prosocial actions are met with certain skepticism when alongside such disgusting cruelty. Interestingly, when impression formation is examined using more minor transgressions and everyday acts of kindness (e.g., "He gave out toys at the children's hospital" and "He told a colleague in public that

she should lose weight"), reports of the negativity bias are inconsistent (Carr & Walther, 2014). Skowronski and Carlston (Skowronski & Carlston, 1992) found that the difficulty in revising an initial bad impression was directly related to the extremity of behaviours; less extreme behaviours were easier to update in light of disconfirming evidence than more extreme behaviours. Consistently, Wojcizske et al. found that the negativity bias was either weak or non-existent when only moderately good and bad behaviours were used (Wojciszke, Brycz, & Borkenau, 1993). In a similar vein, recent work suggests that the negativity bias in impression updating can be explained by perceptions of how rare immoral behaviours are, relative to moral ones (Mende-Siedlecki, Baron, & Todorov, 2013), and report no evidence for a negativity bias after controlling for the perceived frequency of behaviours. These findings call into question the pervasiveness of a negativity bias in impression formation, and consequently, the extent to which our everyday bad impressions can be revised.

Another matter to addressed is that most studies employ relatively few behaviours and only a single 'update'. That is, participants often rate their impression at only two time points; first after being presented with one set of behaviour, and again after being presented with another set of behaviours that are incongruent in valence. However, learning about other's moral character is a continuous, dynamic process, and the restricted methods employed by many studies do not afford the precision to examine momentary fluctuations in moral evaluations. The subsequent chapters will address these limitations by integrating methods from social psychology, behavioural economics, and computational science. With the development of a precise, quantitative measure of impression updating, I hope to fine-

tune the role morality plays in impressions formation and elucidate the mechanisms through which negative beliefs are updated.

To summarize, evolution has endowed us with a strong motivation to avoid being harmed by others. This may lead us to be especially sensitive to signs of immorality leading to a negativity bias in impression formation. However, a tendency to form inflexible negative impressions would hinder the development and maintenance of stable, supportive social relationships. While these are common outcomes observed in psychopathology (Barnow et al., 2009; Unoka et al., 2009) we know that most people are highly adept at building and maintaining social relationships. Responding to immoral acts with leniency and forgiveness may therefore enable the evolution of cooperation despite occasional transgressions. Evolutionary and economic models provide descriptive accounts of these behaviours (Fudenberg et al., 2012; Grim, 1995; Nowak & Sigmund, 1992; Rand et al., 2009), however the cognitive mechanisms that enable them are not well understood. Using novel methods derived from computational science, social cognition, and behavioural economics, I aim to elucidate the mechanisms through which humans maintain healthy relationships in the wake of minor transgressions despite a strong motivation to avoid being harmed by others.

1.2 COMPUTATIONAL MECHANISMS OF IMPRESSION FORMATION

1.2.1 A BAYESIAN ACCOUNT OF IMPRESSION FORMATION

Evaluations of others rely on probabilistic inferences made about people's actions, intentions and motivations, which unfold over time and are influenced by multidimensional

factors including past experiences. Standard experimental designs employed by much of the work on impression formation were limited in their ability to capture the complex dynamics of this inference process. A separate line of research sidesteps these limitations by considering cognitive processes of impression formation from a computational perspective. Computational approaches move beyond observational, passive forms of learning to capture the active, feedback-based nature of dynamic impression updating. These approaches have advanced social cognition by allowing researchers to quantify trial by trial dynamics of learning and decision making to make precise predictions about behaviour.

Recent studies in cognitive science have begun to use computational learning models to investigate the mechanisms through which we infer other's hidden preferences, intentions, and desires over time (Aksoy & Weesie, 2014; Diaconescu et al., 2014; Jern & Kemp, 2015). Bayesian inference models, for instance, provide a normative framework for building models of our environment, using accumulating evidence to optimally update beliefs about hidden states (e.g., preferences, desires, intentions) in the form of posterior probability distributions. The peak of the distribution is analogous to our current belief about the hidden state. In social cognition, beliefs about others' character traits are used to make predictions about their future behaviour. To illustrate, imagine you are standing behind your co-worker, George, in a coffee shop. George hands the cashier a 10 pound bill to pay for his coffee and you notice that the cashier accidentally hands him back a 50 pound bill instead of a 5 pound bill for change. Because you believe George to be honest, you might then expect that George will take a series of actions: speak to the cashier, inform of the mistake, and return the change. The prior character attribution – honesty – thus serves

as the basis for making mental state attributions that are consistent with the trait in question – namely, altruistic preferences and desires.

Now, imagine that instead of observing George follow your predicted response, you watch him fold the bill and place it in his pocket. Consider a scenario where you have known George for a long time and are highly confident about your prior character attribution – honesty. At this point you may begin to search for alternative explanations for George's behaviour -- for example, you may hypothesize that George didn't notice the mistaken change -- and maintain your positive impression. Now consider a scenario where George had just started working with you. Because your experience with George is limited, in this instance you may instead consider that your initial impression was incorrect and begin to believe him to have more self-interested desires. These divergent behaviours illustrate the importance of uncertainty in our prior character attribution in impression updating, a measure which has more often than not been overlooked in the impression formation literature. Yet because our beliefs are limited by both the number and quality of our past experience, we are constantly challenged by the question of how confident we are in our representation of others.

In Bayesian inference, the *variance* of the posterior probability distribution captures how uncertain we are about our beliefs. Bayesian inference suggests how we might optimally incorporate this type of uncertainty, often called estimation uncertainty, towards updating our impressions of others. According to this framework, uncertainty should *suppress* top-down expectation guided processes (prior expectations), but *boost* bottom-up sensory induced signals to promote learning about the uncertain state (Yu & Dayan, 2003). To this end, estimation uncertainty is related to the rate of updating our beliefs in the face

of new information (i.e., sensory signals), where high levels of uncertainty call for faster updating. This relationship between estimation uncertainty and the learning rate is normative: the more uncertain our beliefs, the more we should incorporate new evidence into our beliefs. Accordingly, a weak prior belief that George has honest preferences should motivate belief updating when presented with evidence that does not support this character attribution. In this way, even if initial beliefs are unreliable, they might still serve as a basis for learning about others by motivating more accurate models of a person's character via enhanced information seeking and updating from stimulus inputs.

1.2.2 THE COMPUTATIONS OF MORAL INFERENCE

Seminal work in social cognition has shown that Bayesian reinforcement learning models can accurately capture how humans update beliefs about the intentions and trustworthiness of others (Behrens, Hunt, Woolrich, & Rushworth, 2008; Diaconescu et al., 2014, 2017). In a probabilistic reinforcement learning task, Diaconescu et al. (Diaconescu et al., 2014, 2017) assigned pairs of participants to the role of the 'player' and the 'adviser'. The player received information about the reward probabilities of binary outcomes and predicted which option was associated with reward. The adviser, however, received more complete (though not entirely complete) information about the reward probabilities and used this information to recommend to the player which option to choose. Crucially, the player was aware that the adviser's intentions to provide harmful or helpful advice would vary across trials. Because the outcomes were only probabilistic, an unexpected outcome could either result from chance or a mistake inferring the adviser's intentions (e.g., thinking the advisor was providing helpful advice when the advisor actually meant to mislead the player). Thus, in order to optimally decide whether to update

beliefs about the adviser's intentions from the unexpected outcome participants must also infer the rate at which the adviser's intention change over time. Using a Hierarchical Bayesian reinforcement learning model, the authors found that beliefs about the adviser's intentions were *dynamically* updated as a function of the precision (i.e., inverse uncertainty) of those beliefs and beliefs about how rapidly the adviser's intentions changed across trials. Specifically, the less the adviser's intentions changed across trials, the more certain participants beliefs became about the advisor's intentions. As a consequence, beliefs about the adviser's intentions were less likely to be updated following an unexpected outcome. The results highlight how complex situational factors that influence the stability of other's moral preferences can impact computations of belief updating. Importantly, the results suggest that when other's moral preferences become increasingly deterministic, the extent to which beliefs about others are updated from feedback is dependent on the precision of those beliefs.

The present Dissertation builds on previous computational accounts of social inference by considering how inferences about moral character affect subsequent belief updating. Past studies have demonstrated that prior beliefs about others' character influence the way people behave towards them (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2012; King-Casas et al., 2005), but only a few have sought to understand how character inference affects computations of the learning process itself (Delgado et al., 2005; Fouragnan et al., 2013; King-Casas et al., 2005; Lee & Harris, 2014; Mende-Siedlecki et al., 2012). In reinforcement learning models, people update their beliefs when actual outcomes differ from predicted outcomes, resulting in a *prediction error*. Caudate nucleus activity reflects prediction errors, responding strongest when

feedback deviates most from the predicted outcome, and diminishing over time as feedback becomes less informative. Following from this, Delgado and colleagues (Delgado et al., 2005) had participants make a series of risky decisions about whether to trust agents to return an investment after having read descriptions of their behaviours that indicated good, neutral, or bad moral character. Although the agents were described as having different moral preferences, all agents were selfish 50% of the time. One prediction is that prior information about moral character will set expectations about the agent's behaviour and thereby *increase* prediction error signals when the good and bad agents deviate from those expectations to update beliefs. Instead, they found that prior information about an agent's moral character as either good or bad subsequently *diminished* prediction error signals in the caudate when learning about the agent's trustworthiness. Notably, the reduction in caudate response was most pronounced when the agent was depicted as possessing a good, rather than bad, moral character.

In a subsequent study, Fouragnan et al. (Fouragnan et al., 2013) used a similar investment paradigm to further specify the role that prior moral character information plays in repeated social interactions. They manipulated whether prior moral character information was provided about the agents and whether the agents behaviour was mostly selfish or cooperative. Using a computational learning model, the study revealed that participants were faster to learn about an agent's choices when no prior information was given about their moral character, relative to when good or bad priors were provided. Furthermore, prediction error signals in the caudate were blunted when prior information was given about the agent's moral character. In this case, the results did not indicate whether blunted caudate responses differed for good and bad moral priors. Together these

findings support a dynamic relationship between moral inference and belief updating that is consistent with Bayesian inference; prior character information may disrupt belief updating by suppressing bottom-up sensory-driven processing in favour of top-down expectation-guided processing (Yu & Dayan, 2003). Whether the nature of the prior – i.e., whether the inferred character is good or bad -- impacts the balance between stimulus-driven and expectation-guided processing, however, remains unresolved.

1.3 A ROLE FOR AFFECT IN IMPRESSION FORMATION

1.3.1 VALENCE IN ATTENTION AND LEARNING

Prior experience being equal, the possibility that belief updating is modulated by moral character inference stands in sharp contrast to an optimal Bayesian hypothesis, which dictates that our predictions of others' choices should only depend on the probabilities of possible outcomes based on past experience. However, emotions and biases flood human learning, often impacting how information is processed in systematic ways (Phelps, 2006). For example, seminal work in decision-making and learning demonstrate that people disproportionately weight aversive cues (events, stimuli, and traits) and identify them with greater ease relative to comparable appetitive cues (Baumeister et al., 2001; Rozin & Royzman, 2001). This can explain why people are more affected by a loss than an equivalent gain even when the probability of either outcome is at chance (Kahneman & Tversky, 1979). People also learn what stimuli to avoid faster than what stimuli to approach (Fazio, Eiser, & Shook, 2004), suggesting that people are more interested in avoiding negative outcomes than seeking positive ones. Moreover, negative reinforcement leads to

faster learning than positive reinforcement, and is more resistant to extinction (Ohman & Mineka, 2001). Such findings reflect the longstanding theory that the motivation to avoid a loss is greater than the motivation to achieve a gain (Kahneman & Tversky, 1979, 2000), because the former pose a greater threat to survival. Together, these studies substantiate a disproportionate role for aversive cues in learning and information processing and raise the possibility that affect may influence Bayesian updating in systematic ways.

With a constant stream of incoming information, people cannot afford to process all information to an equal extent. Because information processing is costly, cognitive resources must prioritize information that is most important to the individual. In a similar vein, with the large number of *people* that we encounter in our daily interactions, it is computationally intractable to process all social information to an equal extent. This raises the question; how do we learn efficiently in complex social environments?

One possibility is that people selectively attend to certain aspects of their social environment, prioritizing a subset of individuals for more in-depth learning while generalizing over others. Thus, people may rely on simple heuristics to rapidly categorize individuals from minimal information and use this categorical assessment to determine who they should build more accurate predictive models for. If people selectively attend to a subset of individuals in their environment, this set of individuals should be one that is important to achieving an individual's goals. Chapter 1.1.1 highlighted that people are especially motivated to categorize others into "good" and "bad" moral agents from minimal information, while Chapter 1.1.2 illustrated that people use others' immoral actions to fine-tune this discrimination. What is not yet clear is whether people process information about good and bad agents differently after an assessment of morality is made.

Research in social cognition suggests that people may be especially motivated to process information relevant to immoral character attributes. Early evidence comes from Pratto and John (Pratto & John, 1991) who investigated whether attentional resources would automatically be directed towards task-irrelevant immoral traits. Using a modified Stroop task, participants were presented with a stream of social traits (e.g., *wicked*, *sadistic*, *honest*) and instructed to name the color of the ink that the trait was printed. The extent to which attention was automatically directed away from the task was represented by the latency of the color response. The latency for negative traits was longer than for positive traits and people remembered more of the negative traits in a subsequent memory task. The findings specify a range of cognitive resources that are preferentially seized by the meaning of traits indicative of bad character.

Pratto and John's findings align with emerging research advocating a role for morality in shaping perceptual processes (Gantman, Bavel, & J, 2015). For example, information about whether an actor is morally good or bad predicts visual gaze preferences to amoral outcomes (Callan, Ferguson, & Bindemann, 2013). Other work demonstrates that inferring bad moral character (as opposed to good or neutral character) enhanced the detection of neutral faces during binocular rivalry (Eric Anderson, Siegel, Bliss-Moreau, & Barrett, 2011). Consistently, signs of immorality enhance the detection of perceptually ambiguous moral words while signs of morality hinder detection (Gantman & Van Bavel, 2016). Together, these findings suggest that moral concerns exert top-down influences on what we remember, attend to and see, and that immoral character information *specifically* may enhance cognitive processes, though it is not clear how.

1.3.2 MECHANISMS FOR ASYMMETRIC CHARACTER LEARNING

Due to the importance of rapidly and efficiently responding to motivationally relevant information, humans are likely endowed with faculties designed to optimize their detection and subsequent processing. Arousal is expected to be one such internal state that signals a motivationally relevant event or cue. The relationship between arousal and cognitive processing is well-documented (Eysenck, 2012; Phelps, 2006; Storbeck & Clore, 2008). There is evidence that arousal enhances early visual processing (Keil et al., 2003; Phelps, Ling, & Carrasco, 2006) and facilitates the retention of episodic memory (Heuer & Reisberg, 1992). Threatening stimuli are believed to induce an enhanced state of arousal to optimize the ability to detect and process environmental stimuli when one's wellbeing is at risk (Davis & Whalen, 2001; Kapp, Whalen, Supple, & Pascoe, 1992). Therefore, one possibility is that signs of bad moral character enhance cognitive processing because threatening social stimuli are arousing (Öhman, 1986). This hypothesis is consistent with the finding that participants who were more emotionally aroused when observing the outcomes of an untrustworthy partner were more accurate at recognizing the individual who betrayed them in a subsequent memory task (Fouragnan, 2013). The results suggest that signs of immorality induce an enhanced state of arousal which may increase attention and information processing.

Research in non-social perceptual learning suggest another role for arousal that is especially relevant for Bayesian inference. Computational models of learning suggest that how certain we are about a belief depends on the history of information available to us: the less ambiguous the information, the more certain we should be. However, recent evidence suggests that this is not the whole story. Instead, internal signals such as heart rate and

alertness can influence how certain we are about our beliefs independently from whether they are correct or not (D. L. Neumann, Fitzgerald, Furedy, & Boyle, 2007). Nassar et al. (Nassar et al., 2012) directly examined the relationships between arousal and non-social perceptual learning in a predictive-inference task where participants inferred the mean of a gaussian distribution that changed stochastically over time. The authors found that pupil diameter, which is believed to be a reliable indicator of central arousal (Eldar, Cohen, & Niv, 2013; McGinley et al., 2015; Reimer et al., 2016), was positively related to uncertainty about the inferred mean and subsequent belief updating, as estimated from a Bayesian learning model. Importantly, the authors found that a task-independent manipulation of arousal influenced subsequent learning rates, suggesting that momentary arousal can impact the computations of uncertainty in perceptual inference.

The relationship between pupil-linked arousal and uncertainty is also supported by research in non-social perceptual decision-making. For instance, larger pupil diameter was not only linked to greater subjective decision uncertainty in a forced choice motion discrimination task but also predicted an increased tendency to switch responses on the following trial (Urai, Braun, & Donner, 2017). Likewise, greater pupil-linked arousal was associated with an increased tendency to explore alternative choice options in a four-armed bandit task (Jepma & Nieuwenhuis, 2010). Thus, pupil-linked arousal systems encode uncertainty that increase both learning rates and information-seeking behaviour, consistent with the idea that uncertainty signals the need for rapid learning.

The research in non-social perceptual learning and decision-making raise the intriguing possibility that arousal linked to uncertainty regulates the influence of new information on existing beliefs. Consistent with this notion, brain areas that regulate the

extent to which new information is incorporated into existing beliefs are strongly linked to arousal and autonomic functions (Behrens et al., 2008; Elliott & Dolan, 1998; Jepma & Nieuwenhuis, 2010; Yu & Dayan, 2003). These areas include the locus coeruleus and anterior cingulate cortex, which have also been linked to uncertainty (Elliott & Dolan, 1998; Yu & Dayan, 2003). Recent work using a threat-of-shock manipulation suggests how this process might unfold. Threat-induced arousal was shown to increase sensitivity to new information by augmenting the influence of prediction errors via amplified cortical excitability while dampening feedback from prefrontal cortices to reduce reliance on prior expectations (Cornwell, Garrido, Overstreet, Pine, & Grillon, 2017). Such an arousal-linked learning system may be especially advantageous in the face of impending threats, where the ability to rapidly detect and respond to one's environment is essential for avoiding harm.

One question is whether this system extends to learning about *socially* threatening agents as well as non-social cues. Because social threats are also emotionally and physiologically arousing (Fouragnan, 2013; Noordewier, Scheepers, & Hilbert, 2019; Öhman, 1986; Roelofs, Hagenaars, & Stins, 2010), people may be especially uncertain about immoral agents and consequently augment the relative influence of newly arriving over historical information on existing beliefs. In other words, beliefs about bad agents may be updated more rapidly because uncertainty associated with potential threats promotes learning from new information. Conversely, diminished vigilance may suppress updating from new information to favour rapidly developed (prior) assessments when evaluating the choices of good agents. This postulates a theoretical model where inferences about moral character impact the relative influence of historical and new information on

existing beliefs in impression formation. More uncertain, flexible beliefs may help maintain relationships in the wake of immoral behaviours by enabling them to update rapidly in light of new information.

To summarize, computational models may afford the necessary precision to study moment-by-moment fluctuations in our evaluations of others moral character. Bayesian learning models provide a normative framework to capture the dynamics of moral inference, where beliefs about others morality are updated as a function of the certainty of prior beliefs. To date, little direct evidence exists to discriminate the computations underlying how people form impression about agents with different moral traits. However, the literature suggests that arousal related to the moral character of the agent may impact critical mechanisms of belief updating by shaping the extent to which people rely on new information versus prior knowledge when learning about others. Consequently, examining the computations underlying moral inference may elucidate the mechanisms through which people can update negative impressions despite the potency of immoral information in impression formation. Given the importance of cooperation and forgiveness in maintaining healthy relationships, these processes may be crucial for understanding populations displaying dysfunctional social cognition and behaviour.

1.4 OVERVIEW OF CHAPTERS

The present dissertation presents four empirical chapters that shed new light on the computational mechanisms underlying how moral impressions are formed and evolve over time. The chapters raise important insights into how the identified mechanisms support adaptive social functioning, and how maladaptive behaviours arise when these processes

go awry. To address the research questions, each of the studies in this dissertation apply computational modelling to behavioural data from a Moral Inference Task. In the following section, I outline the theoretical framework and results for each of the studies presented in this dissertation.

1.4.1 PART I

Part I of this dissertation examines the cognitive mechanisms of moral inference in humans. A large body of work in social psychology suggests that negative moral impressions are especially difficult to change because inaccurately attributing good character to bad people increases one's chances of being harmed (Baumeister et al., 2001; Skowronski & Carlston, 1989). However, the tendency to rigidly form negative moral impressions can be costly as well: attributing bad character from rare and minor transgressions leaves little room for people to make mistakes and can impede the development of stable relationships necessary for healthy social functioning. Responding to immoral acts with leniency and forgiveness may enable the development of successful relationships despite occasional harms. Although evolutionary and economic models provide descriptive accounts of these behaviours (Fudenberg et al., 2012; Grim, 1995; Nowak & Sigmund, 1992; Rand et al., 2009), the cognitive mechanisms that enable them are not well understood. Advances in non-social perceptual learning may hold a clue. This work suggests that a pupil-linked arousal system may encode uncertainty signals that facilitate learning from future outcomes (Allen et al., 2016; Nassar et al., 2012). If this mechanism extend to learning about socially threatening others, who are also believed to evoke arousal, this may enable negative moral impressions to be flexibly updated from new information (Öhman, 1986). The experiments described in Chapters 3-4 set out to examine

the cognitive mechanisms of moral inference to elucidate the processes that support adaptive functioning in social relationships.

Chapter 3 employs a novel behavioural task designed to assess two separate measures of moral inference: the ability to learn and predict others' objective moral preferences and the formation of explicit moral character impressions. In the Moral Inference Task, participants predicted whether agents chose to profit from harming an anonymous stranger. Intermittently participants rated their impression of the agent's moral character and how confident they were about their impression. One agent (the 'good' agent) required significantly more money to harm the stranger than the other (the 'bad' agent). No information about the agents was provided to participants prior to observing their choices. Therefore, to optimally predict the agent's choices participants had to gather information across trials to infer the agent's exchange rate between money and harm. A Bayesian reinforcement learning model was fit to participants' trial-wise predictions to capture the influence of historical information and newly arriving information on existing beliefs about the agents' moral preferences.

Moral inference was described by an asymmetric Bayesian updating mechanisms where beliefs about the bad agent were more volatile than beliefs about the good agent, relying more heavily on new over historical information (Study 1-3). Participants also expressed greater subjective uncertainty about their impressions of the bad relative to the good agent, suggesting that meta-cognitive processes play a role in moral inference. Predicting sports performance did not elicit a similar asymmetry when learning about agents who significantly differed in their skill level in the absence of moral information (Study 4). However, beliefs about skill-level were more uncertain and volatile when the

agent also presented signs of immorality (Study 5), suggesting that properties of an agent's moral character influence social inference more generally. The vulnerability of negative moral beliefs to new information, over historical information, contributed to the ability to revise negative impressions in light of new evidence (Study 6).

The findings from Chapter 3 raise two plausible hypotheses for why beliefs about bad agents are especially uncertain and volatile. One hypothesis is that the bad agent violated participants' expectations to a greater degree than the good agent. Beliefs about the bad agent would therefore be especially amenable to Bayesian updating, by which belief updates are optimized to minimize surprise (Mathys, Daunizeau, Friston, & Stephan, 2011). An alternative hypothesis is that participants were especially motivated to learn about socially threatening agents. Because both threat and unexpected outcomes evoke arousal (Allen et al., 2016; Nassar et al., 2012; Storbeck & Clore, 2008) either one may drive uncertainty signals to facilitate learning about putatively bad agents.

If participants hold optimistic moral expectations, asymmetric updating should be related to the extent to which participants believe that others are good and trustworthy. Additionally, if in the absence of additional information people expect others to have preferences similar to their own (Brañas-Garza, Rodríguez-Lara, & Sánchez, 2017; Hsee & Weber, 1997; Yamagishi et al., 2013), then participants who are more averse to harming a stranger should show a larger asymmetry in updating beliefs about good and bad agents. Each of these predictions were tested in Chapter 4 with no supporting evidence found. Incentivizing a separate group of participants to predict, in the context of decisions to profit from others' pain, how "most people" would choose also found no evidence for optimistic expectations.

The final study described in Chapter 4 further explored the role that moral expectations play in moral inference. The study was designed to arbitrate between alternative explanations for asymmetric updating by dissociating moral expectations using facial cues (high versus low threatening faces) from the agent's morality (bad versus good). The data revealed that beliefs about bad agents were more volatile and uncertain than beliefs about good agents, and these effects did not covary with manipulated prior moral expectations. Together, the results support a theoretical model for moral inference whereby moral expectations are rapidly discounted to promote cognitive flexibility in the service of building richer models of potentially threatening others.

To summarize, Part I (Chapters 3-4) presents a series of experiments designed to elucidate the computational mechanisms supporting moral inference. Chapter 3 demonstrates that moral inference is explained by an asymmetric Bayesian updating mechanism where beliefs about the morality of putatively bad agents are more uncertain (and thus more unstable) than beliefs about the morality of good agents. Chapter 4 clarifies why such a mechanism presents by examining the role that prior expectations play towards asymmetric belief updating of good and bad agents. The data presented in Chapter 4 support the hypothesis that harmful expectations, specifically, motivate flexible belief updating.

1.4.2 PART II

Any discussion about the mechanisms for adaptive social inference would be incomplete without considering the potential consequences that ensue when these mechanisms break down. Aberrant learning and decision-making lie at the heart of many populations characterized by maladaptive social functioning. Computational modelling

may be particularly useful for understanding the mechanisms underlying dysfunction. Consequently, Part II of this dissertation applied the Moral Inference Task to examine learning in two populations associated with abnormal interpersonal functioning: individuals exposed to violence and patients with Borderline Personality Disorder (BPD).

Exposure to community violence is a prominent risk factor for aggression, antisocial behaviour and incarceration. How and why exposure to violence leads to maladaptive social behaviours is not well understood. Perhaps one of the most fundamental ingredients for healthy social functioning is learning who is good and trustworthy versus who is not. Consequently, Chapter 5 applied the Moral Inference Task to investigate how chronic exposure to violence affects the ability to learn others' morality and use this information to adaptively modulate trust behaviour in a sample of incarcerated males. The data suggest that exposure to violence adversely impacted some components of moral inference, but not all. Although all participants were able to learn the agents' moral preferences and predict their decisions, only those with little exposure to violence were able to use this information to make adaptive trust decisions, placing more trust in the good agent than the bad. Meanwhile, those with the highest exposure to violence trusted the good and bad agents equally. In particular, participants with high exposure to violence extended less trust than optimal when interacting with the good agent, which resulted in missed opportunities.

The relationship between exposure to violence and maladaptive trusting behaviour was mediated by disturbances in impression formation. Consistent with evidence that violence normalizes beliefs about harm (Ng-Mak, Stueve, Salzinger, & Feldman, 2002), exposure to violence predicted more lenient impressions of the bad agent. Meanwhile, those with the highest exposure made harsher evaluations of the good agent, consistent

with evidence that individuals who have experienced violence themselves interpret the behaviour of neutral actors as hostile (Dodge, Bates, & Pettit, 1990). Disturbances in impression formation and trust behaviour in turn predicted real antisocial behaviour in prison. The findings suggest that chronic exposure to violence leaves resounding effects on the ability to use other's actions to distinguish those we should avoid from those we should befriend. In turn, this may lead individuals to respond inappropriately in social interactions.

BPD is a serious mental illness similarly characterized by marked disturbances in interpersonal relationships, with symptoms often persisting after years of resource intensive treatment. Part I of this dissertation indicates that healthy adults hold more flexible beliefs about those who they infer a bad moral character relative to a good moral character. This is hypothesized to reflect an adaptive mechanism for sustaining relationships when others sometime behave badly. In BPD, relationships are characterized as intense and unstable (American Psychiatric Association, 2013). Relative to healthy adults, BPD patients' social networks have a greater number of relationships that are terminated (Clifton, Pilkonis, & McCarty, 2007), they often hold grudges and present difficulty forgiving others (Sansone et al., 2013; Thielmann et al., 2014). This suggests that patients with BPD may lack an adaptive mechanism for maintaining relationships. Consequently, Chapter 6 set out to test the hypothesis that BPD would be associated with slower updating of initially bad moral impressions, and that this deficit would be restored following treatment in a Democratic Therapeutic Community (DTC), which has shown some promise in ameliorating interpersonal disturbances in BPD.

In line with this prediction, the results indicate that the effect of BPD on moral inference is intrinsically related to the morality of the agent. BPD was associated with less uncertain and flexible beliefs about putatively bad agents, relative to a sample of matched healthy control participants. This may explain findings that patients exhibit less coaxing following a rupture of trust (King-Casas et al., 2008), slower learning rates when inferring other's trustworthiness (Fineberg et al., 2018), and difficulty forgiving others (Thielmann et al., 2014). Conversely, BPD was associated with more uncertain and flexible beliefs about putatively good agents. This may explain the ease patients have in terminating relationships or clinical observations that social evaluations shift rapidly from a period of admiration to dislike in the wake of minor slights (Bender & Skodol, 2007). The effects of DTC treatment were also related to the morality of the agent. While there were no effects of DTC on moral inference for good agents, DTC-treated BPD patients were faster to update beliefs about bad agents relative to untreated patients. The results provide a mechanistic explanation for social deficits in BPD and suggest that DTC treatment may shape social functioning by increasing patients' openness to learning about adverse social interaction partners.

To summarize, Part II (Chapter 5-6) applied the Moral Inference Task to populations characterized by maladaptive social functioning to investigate the role of moral inference in dysfunctional cognition and behaviour. Chapter 5 demonstrates that chronic exposure to community violence impairs subjective impression formation and, consequently, adaptive trust behaviour in a sample of incarcerated males. Chapter 6 demonstrates that asymmetric Bayesian updating for good and bad agents is absent in a sample of patients diagnosed with Borderline Personality Disorder (BPD) but is present in a sample of BPD patients who

received treatment at a therapeutic community. The data presented in Part II show that the processes underlying moral inference are crucial for understanding social dysfunction in both clinical and non-clinical populations.

Chapter 2

2 EXPERIMENTAL MATERIALS

Inferences of moral character feed into everyday decisions that we make, from deciding who to date, electing our nation's leader, choosing our children's caretaker, or admitting students to our lab. To make these decisions we must infer an agent's character based on their actions, by considering the likely end consequences of their actions (Malle, 2011). Recent work has shown that people readily infer people's mental states (intentions, feelings, desires) by observing their decisions, deploying a "naïve utility calculus" that assumes people's choices are aimed at maximizing desirable consequences and minimizing undesirable consequences, where desirability is evaluated with respect to the agent's preferences (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). This means that in situations where agents make deterministic choices, intentions can be inferred by observing their actions and learning their preferences. Evaluations of moral character are intimately linked to inferences about preferences, where accumulated evidence of self-serving preferences, indicative of bad intent, leads to a judgment of bad character (Leifer, 1971; Leslie, Knobe, & Cohen, 2006; Uhlmann, Pizarro, & Diermeier, 2015). These inferences are used to adaptively learn and decide whom to trust in social interactions (Haidt & Joseph, 2004; Stanley, Sokol-Hessner, Banaji, & Phelps, 2011).

This highlights two distinct components related to learning about others' morality. On the one hand, people use social cues (such as observing people's choices) to *objectively* update their beliefs about others' preferences by gradually accumulating information over time to predict future outcomes. For the purposes of this Dissertation, I refer to this process as learning about *moral preferences*. On the other hand, people use learned information to form *subjective,* global impressions about character, which I refer to as forming *character impressions*. To clarify the distinction, beliefs about moral preferences refer to specific knowledge about moral behaviours, for example, '*John will not sacrifice his own time to help his friend move*'. While character impressions refer to global estimates about a trait derived from historical information and can be used to infer preferences, for example, '*John is selfish*'. Note that while different people may have the same objective knowledge about an agents' behaviour (e.g., that John will not sacrifice his time to help his friend move), they may form distinguishable character impression (e.g., some may form a highly negative impressions of John's selfishness but another may form a mildly negative impression). We developed a novel Moral Inference Task to measure each of these components of moral learning.

In this chapter, I outline the Moral Inference Task along with other experimental methods common to the studies described in Chapters 3-6. Methods unique to individual chapters will be described in those chapters' methods sections. I will first describe the Moral Inference Task developed to investigate the mechanisms of moral learning and belief updating. I will follow by describing the computational model that was fit to participant behaviour, and finish with a description of a trust game that was used to assess adaptive trust behaviour in relevant chapters.

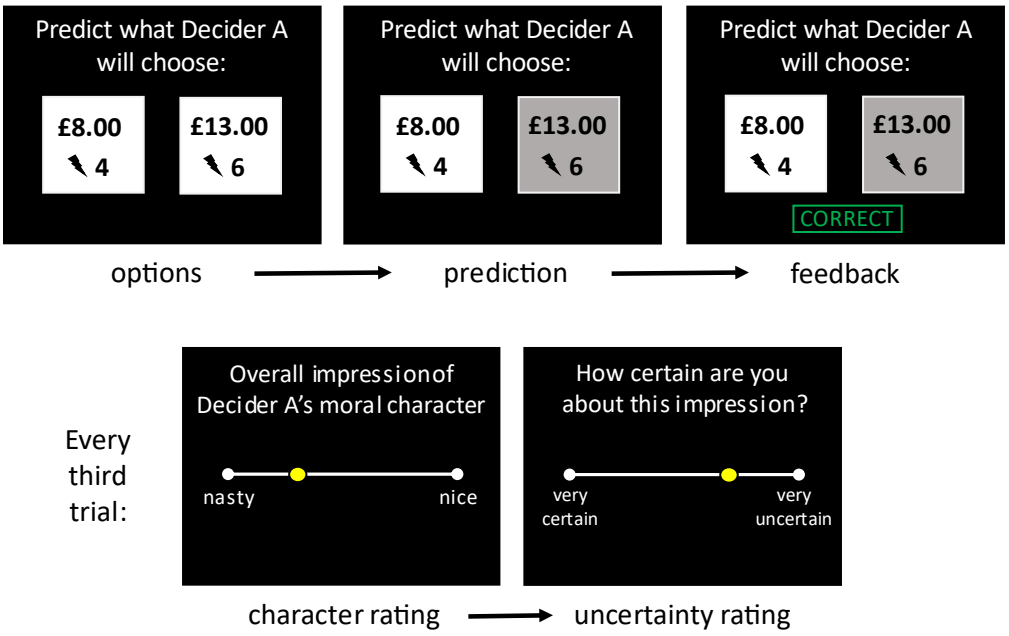
2.1 MORAL INFERENCE TASK

Paradigm overview

We developed a novel Moral Inference Task to investigate the computational basis of moral inference and its temporal dynamics. In the task, participants predict the choices of two "agents" who repeatedly decide whether to inflict painful electric shocks on another person in a different room in exchange for money. For a series of 50 choices each, agents chose between two options: more money for themselves plus more shocks for the anonymous victim, or less money for themselves plus fewer shocks for the victim (see **Figure 2.1**). No *a priori* information is given about the agents. Instead, participants are required to learn about the agents' moral preferences through trial and error. Participants complete the whole sequence for one agent before beginning with the second, and the order of agents is randomized across participants. Additionally, on every third trial participants rate their subjective impression of the agent's character on a continuous visual analogue scale ranging from 0 (*nasty*) to 100 (*nice*). After making subjective character ratings, participants indicate how uncertain they are about this characterization on a scale ranging from 0 (*very certain*) to 100 (*very uncertain*) (see **Figure 2.1**). These rating measures are additionally collected before participants observe either of the agents' choices, in order to obtain an estimate of participants' prior beliefs about how agents will behave in this setting. Together, this provided us with a trajectory of participants' explicit subjective ratings of each agent's moral character, and how uncertain participants are about their characterization, which can be used to assess the consistency between subjective experiences and behaviour as measured by our computational model (described in Chapter 2.2).

Figure 2.1 Moral Inference Task

Participants predict sequences of choices for two agents, "Decider A" and "Decider B". On each trial, the agent chose between a more harmful (more shocks inflicted on another person for more money) and a less harmful (fewer shocks/money) option. After every third trial, participants rate their impression of the agent's moral character and the uncertainty of their impression.



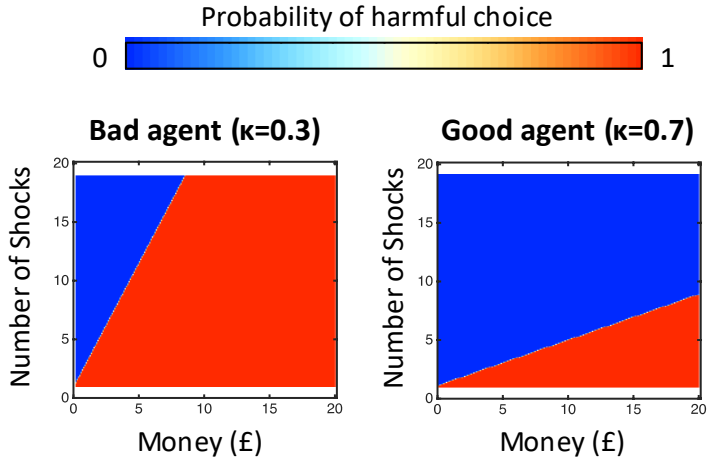
Simulating agents' behaviour and generating trial sequences

We generated agent behaviour using a model that accurately captures individual differences in moral preferences in this choice setting and correlates with several traits related to prosocial behaviour, including empathy and psychopathy (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Crockett et al., 2015). The model includes a "harm aversion" parameter, κ, which quantifies the subjective cost of harming the victim as an exchange rate between money and pain. Because ethical systems universally judge harming others for personal gain as morally wrong (Gert, 2004), we operationalized moral preferences as harm aversion in our paradigm. When $\kappa = 0$, agents are minimally harm averse and will accept any number of shocks to the victim to increase their profits; as κ approaches 1, agents become maximally

harm averse and will forgo infinitely increasing amounts of money to avoid delivering a single shock. For the inference task, we created two agents who differed substantially in their harm aversion, with the "good" agent requiring more compensation per shock to inflict pain on others than the "bad" agent (bad: $\kappa = 0.3$ or £0.43 per shock; good: $\kappa = 0.7$ or £2.40 per shock; **Figure 2.2**). A small pilot study indicated that the preferences of the good and bad agent were symmetric around participants' expectations of "average" behaviour, which was not significantly different from $\kappa = 0.5$ (reported in Chapter 4, page 122). For each study, the inference task used local currency.

Figure 2.2 Probability of harmful choice as a function of money and shocks

Heat maps summarize the bad and good agents' probability of choosing the more harmful option as a function of money gained and shocks delivered.



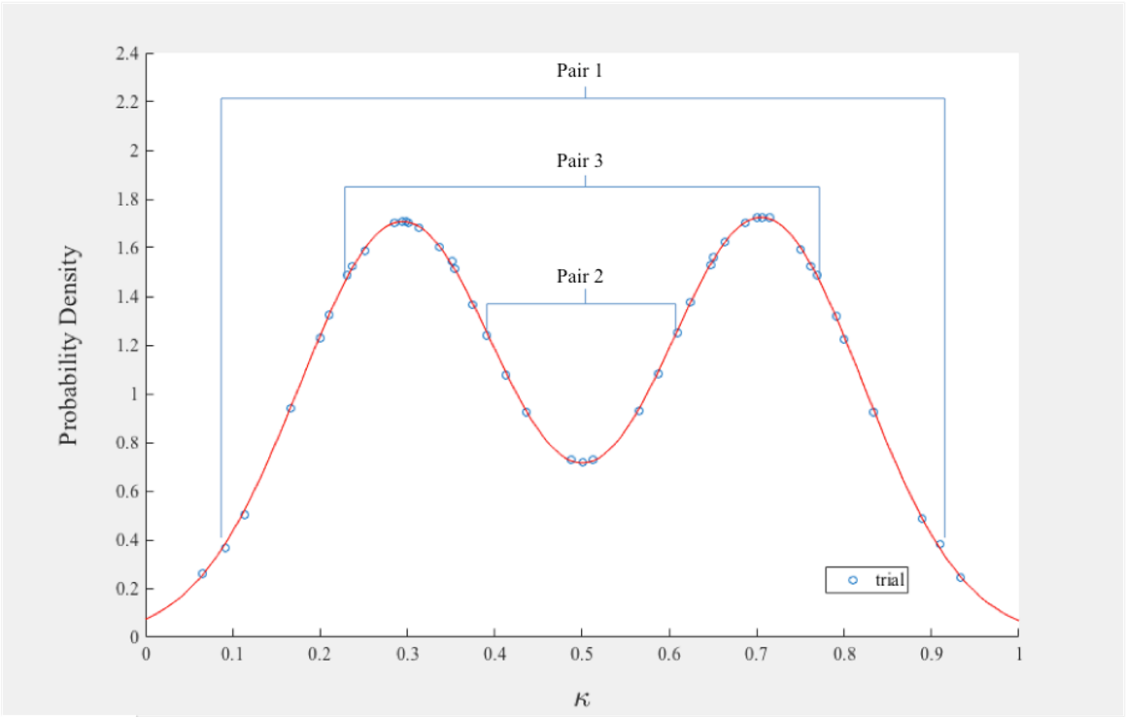
Participants observe the two agents make choices for identical trial sequences. This means that on every trial the agents face the same two options, but because the agents have different preferences towards harming the victim, they often choose differently.

Each trial contains a pair of choices [s-, m-] and [s+, m+] that matches the indifference point of a specific κ value. We first created a set of 24 trials where values of κ were randomly drawn from a normal distribution around the good agent's indifference

point (M = 0.7, s.d. = 0.15), and constrained such that $\kappa < 0.95$. Next, we created a set of 24 matched trials around the bad agent's indifference point by subtracting each κ value from 1. We wanted participants to observe identical trial sequences for the two agents, but also minimize any potential differences in learning about the agents that could be explained by discrepancies in the informational value of the trial sequence. Note that a trial with high informational value for the bad agent will have relatively low informational value for the good agent, and vice versa. That is, if someone is willing to commit an extremely harmful act in exchange for some personal gain, it is highly probable that they would also be willing to commit a lesser harm in exchange for the same gain. Within the context of this task, this means that knowing that an agent accepted 5 pounds to increase the number of shocks by 5, one can confidently predict the agent will accept an offer of 6 pounds to increase the number of shocks by 5 but less confidently predict whether the agent will accept an offer of 4 pounds to increase the number of shocks by 5. Consequently, we created pairs of trials $[\kappa, 1-\kappa]$ where the members of each pair were matched in informational value for the good and bad agent. Effectively, this meant that a trial that was highly informative about one agent's indifference point was paired with a trial that was equally informative about the other agent's indifference point (**Figure 2.3**). We then randomized the order of presentation of each member of the pair. The pairs comprised trials 2-49 of the sequence, while the initial and final trials were fixed to $\kappa = 0.5$.

Figure 2.3 Graphical depiction of the optimized trial sequence

To minimize the possibility that differences between agents could be attributed to the order of observations we created pairs of trials that were matched in informational value for the good and bad agent. Each pair comprised of trials with mirrored κ values: one member of the pair was randomly drawn from a normal distribution around the good agent's indifference point, and the other member was the mirrored deviation from the bad agent's indifference point (1 – κ). This resulted in a bimodal distribution of trials and ensured that a trial that was highly informative about one agent was sequentially paired with a trial that was equally informative about the other agent.



Given a sequence of κ values, we then generated shock and money options for each κ value by generating 10,000 random pairs of positive shock movements Δs ($1 < \Delta s < 20$), and positive money movements Δm ($0.10 < \Delta m < 19.90$), and selected the pair closest to the indifference point of that κ value [Δs, Δm]. Next, these pairs were transformed into choices containing smaller amounts of shocks and money (s- and m-) and greater amounts of shocks and money (s+ and m+) as follows: s- was a positive integer between 0 and 20, randomly drawn from a uniform discrete distribution with the constraint that $0 < s- + \Delta s < 20$. Similarly, m- was a positive number between 0 and 20, randomly drawn from a uniform

discrete distribution, rounded to the nearest 10th and constrained such that $0 < m- + \Delta m < 20$. s+ and m+ were then set by adding Δs and Δm to s- and m-, respectively.

We simulated the agents' decisions by computing the utility for choosing the more harmful option (V_{harm}) as a function of the agent's κ ($\kappa_{bad} = 0.3$, $\kappa_{good} = 0.7$). This model is identical to the model that best predicts human choices in the same setting (Crockett et al., 2014, 2015).

Equation 2.1

$$V_{harm} = (1 - \kappa_n)\Delta m - \kappa_n \Delta s$$

Where κ_n is the κ for agent n. A softmax function was used to transform V_{harm} into a probability of choosing the more harmful option, P_{harm} :

Equation 2.2

$$P_{harm} = \frac{1}{1 + e^{-\beta \times V_{harm}}}$$

Where β defines the steepness of the slope in the sigmoid function. As β approaches 0 the slope become increasingly horizontal, signifying a large amount of noise in the agent's choices. As β approaches infinity the sigmoid approximates a step function, and indicates increasingly deterministic choice preferences. Except where otherwise indicated, β was fixed to 100 to simulate agents that were completely deterministic in their choices.

Equation 2.3

$$u = [x_{rand} < P_{harm}]$$

Equation 2.3 converts the probability of choosing the more harmful option into a binary choice, *u*. x_{rand} is a random number between 0 and 1.

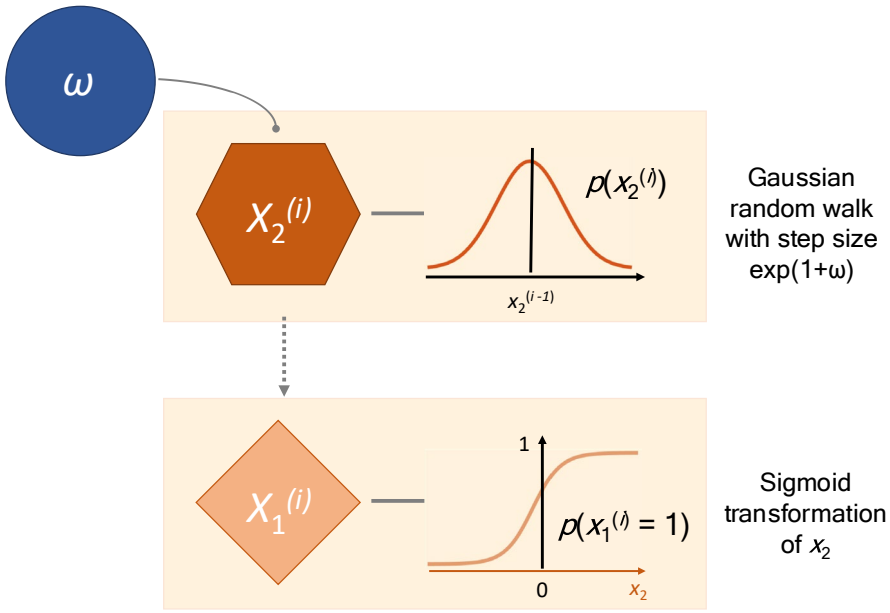
2.2 COMPUTATIONAL MODELLING

Perceptual Model

Our main goal was to assess how participants updated their beliefs about an agent's harm preference (κ) when that agent was either more harmful ('bad') or less harmful ('good'). For this purpose we applied one primary perceptual model to participants' trial-by-trial responses: a reduced version of the Hierarchical Gaussian Filter (HGF), a Bayesian model for learning hidden states with informational uncertainty (due to a lack of knowledge) and without environmental volatility (See Mathys and colleagues (Mathys et al., 2011) for a theoretical background on the full model). The HGF draws on the belief that the brain has evolved to process information in a manner that approximates statistical optimality given individually varying priors about the nature of the process being predicted; effectively maintaining and updating a generative model of its inputs (u) to infer on hierarchically organized hidden states (see **Figure 2.4**).

Figure 2.4 Graphical representation of the Hierarchical Gaussian Filter

The model describes how participants infer the moral character of the agent (x_2) in order to predict whether the agent will harm on a given trial, i. Beliefs about moral character (x_2) are represented by probability distributions characterized by a mean, μ, and a variance, σ. x_2 evolves over time as a Gaussian random walk whose step-size is governed by ω, a participant-specific parameter which captures individual differences in belief updating. Beliefs about whether the agent will choose the more (x_1=1) or less (x_1=0) harmful option are represented by a sigmoid transformation of x_2.



For the purpose of this study, our model comprises only two hidden states x_1^i and x_2^i, where *i* signifies the trial index. The first state, x_1, is time-varying and denotes the agent's upcoming choice. x_1 is binary because there are only two options that the agent can choose: the more harmful option (greater profit for the self and more shocks for the victim) or the less harmful option (less profit for the self and fewer shocks for the victim). The probability that an agent will choose the more harmful option ($x_1^i = 1$) versus the less harmful option ($x_1^i = 0$) is governed by the next state in the hierarchy, x_2. x_2 is a continuous state evolving over time as a Gaussian random walk, and signifies the belief about the agent's (logit-transformed) κ -- the exchange rate between money and pain. The hierarchical coupling between x_1^i and x_2^i explains that a participant's prediction about an agent's choice on trial

i is dependent on their current belief about that agent's κ, defined as a probability density. The conditional probability of x_1 given x_2 is described in **Equation 2.4**.

Equation 2.4

$$p(x_1|x_2) = s(x_2)^{x_1}\big(1 - s(x_2)\big)^{1-x_1} = \text{Bernoulli}(x_1; s(x_2))$$

Where $s(\cdot)$ is a logistic sigmoid (softmax) function:

Equation 2.5

$$s(x) \stackrel{\text{def}}{=} \frac{1}{1 + \exp(-x)}$$

The temporal evolution of x_2 is governed by a participant-specific parameter ω, which allows for inter-individual differences in belief updating. Thus, ω represents a *global* measure of the tonic volatility – i.e., the extent of trial-wise changes in x_2. In other words, ω captures inter-individual variability in the rate at which beliefs evolve over time, and consequently how rapidly people update their beliefs about the agent's harm aversion across all trials. As ω approaches ∞ beliefs become increasingly unstable and new information is favoured over prior beliefs. Conversely, as ω approaches -∞ beliefs become increasingly stable, so greater weight is instead placed on prior beliefs. Given ω and the previous value (with time index $i - 1$) of x_2, we now have the generative model for the current values (with time index i) of x_1 and x_2 in **Equation 2.6** (graphically represented in **Figure 2.4**; for details see (Mathys et al., 2011)).

Equation 2.6

$$p\big(x_1^i, x_2^i, |\omega, x_2^{i-1}\big) = p\big(x_1^i \big| x_2^i\big) p\big(x_2^i \big| x_2^{i-1}, \omega\big)$$

with

Equation 2.7

$$p\left(x_2^i \middle| x_2^{i-1}, \omega\right) = \mathcal{N}\left(x_2^i; x_2^{i-1}, \exp(\omega)\right)$$

Model inversion was used to optimize the posterior densities over hidden states, x_1 and x_2, and parameter ω. Participants' posterior beliefs were represented by probability distributions (**Figure 2.4**) with mean μ and variance σ. Variational Bayesian inversion yields a simple update equation under a mean-field approximation, where beliefs are updated as a function of precision-weighted prediction errors. For the present study we focus on the update at level 2 of the hierarchy (Mathys et al., 2011).

Equation 2.8

$$\Delta\mu \propto \sigma_2 \delta_1^i$$

with

Equation 2.9

$$\delta_1^i = \mu_1^i - \hat{\mu}_1^i$$

and

Equation 2.10

$$\sigma_2 = \frac{\hat{\pi}_1^i}{\hat{\pi}_2^i \hat{\pi}_1^i + 1}$$

Where π is the precision (i.e., the inverse variance) in participants' posterior belief $\frac{1}{\sigma}$, and δ_1^i is the prediction error on the trial outcome. Caret symbols (^) are used to denote predictions *prior* to observing the outcome at trial *i*. Thus, $\hat{\pi}_1^i$ is the precision of the prediction at the first hierarchical level and $\hat{\pi}_2^i$ is the precision of the prediction of the

posterior belief. It can be shown from **Equation 2.10** that prediction errors are given a larger weight when the precision of the prediction of the agent's choice is high, or when the precision of the belief about the agent's κ is low. In summary, these equations describe trial-wise updating of beliefs about an agent's preference towards harming the victim, which approximates Bayes optimality (in an individualized sense given differences in ω) and determines the participant's estimate of the probability that an agent will harm. Crucially, our model provides a trial-by-trial estimate of the subject's uncertainty about the agent's preference towards harming the victim as measured by the variance of beliefs, σ. The variance weights predictions errors on a trial-by-trial basis and thus represents a *dynamic* learning rate because it accounts for the precision of the belief at any given time.

Decision Model

The decision model describes how the participant's posterior belief about the agent's κ maps onto their predictions of the agent's decisions (y). In the HGF, this belief $\hat{\mu}_1^i$ corresponds to the logistic sigmoid transformation of the predicted preference μ_2^{i-1} of the agent towards harming the victim.

Equation 2.11

$$\hat{\mu}_1^i = s(\mu_2^{i-1})$$

For the present study, we assumed that participants would predict others' decisions using a similar rationale to how they make decisions themselves. In other words, we assumed that people's preferences are described by a utility model, and that people think others' preferences are described by the same model. Consequently, we applied a decision model that accurately describes human choices in the same choice setting (Crockett et al.,

2014, 2017, 2015) (and that was also used to simulate the agent's actual choices, **Equation 2.1**).

Equation 2.12

$$V^i_{\text{harm}} = \left(1 - \hat{\mu}^i_1\right)\Delta \text{m}^i - \hat{\mu}^i_1 \Delta \text{s}^i$$

This model replaces a participant-specific parameter κ with the predicted belief derived from the perceptual model $\hat{\mu}^i_1$ to compute the value that the agent will choose the more harmful option on trial *i*. The probability that the participant predicts the more harmful option ($y = 1$) as opposed to the more helpful option ($y = 0$) is described by the softmax function in **Equation 2.13**.

Equation 2.13

$$P^i_{\text{harm}} = s(\beta V^i_{\text{harm}})$$

Where β is a free parameter (individually estimated like ω) that describes how sensitive predictions are to the relative utility of different outcomes, or the prediction noise.

Estimation of Model Parameters

A crucial aspect of Bayesian inference is the specification of a prior distribution for the belief (listed in **Table 2.1**). We defined the priors based on our experimental design. Specifically, for the present studies we wanted to compare learning parameters between the two agents. In keeping with our experimental design, which did not give participants any basis for assumptions about the agent's tendency to harm, we chose to initialize the prior mean over μ_2 and σ_2. such that it amounted to a neutral prior belief about κ which was equidistant from the true value of the agents' preferences. For the free parameters ω

and β, we chose a prior mean that was relatively uninformative (with large variance) to allow for substantial individual differences in learning both between participants and within participants (i.e. between agents). Notably, the prior means on ω and β were equally unconstrained with a variance of 1. This ensured that adjustments in parameter estimates were not biased towards favouring one parameter over the other.

Table 2.1.

Prior mean and variance of the perceptual and response model parameters.

Parameter	Notes		mean	variance
ω	Constant component of the tonic volatility at the second level. Represents the temporal evolution of x_2. *Estimated in native space.*		-4	1
Predictions (x_1)	Predictions are a sigmoid transformation of x_2 , and so do not have prior values.	$\mu_{1:}$ $\sigma_{1:}$	none none	none none
Probabilities (x_2)	The prior mean on x_2 (prior belief about agent's harm-aversion, κ) was fixed to a neutral point that was equidistant from the true κ value of both agents. Estimated in logit space. The prior variance on x_2 was fixed to ensure that any differences in learning about good and bad agents derived from the model could not result from differences in the prior estimates. Estimated in log-space.	$\mu_{2:}$ $\sigma_{2:}$	0.5 0.35	0 0
β	Constant component that describes how sensitive prior beliefs are to the relative utility of different outcomes, or the prediction noise. Estimated in log-space.		1	1

The perceptual model parameter ω and decision model parameter β were estimated from the trial-wise predictions using the Broyden Fletcher Goldfarb Shanno optimization algorithm as implemented in the HGF Toolbox (https://tnu.ethz.ch/tapas). This allowed us to obtain the maximum-a-posteriori estimates of the model parameters and provided us with state trajectories and parameters representing an ideal Bayesian observer given the individually estimated parameter ω.

We fit the model separately for participant's predictions of the bad and good agent. This produced for each agent a sequence of trial-wise beliefs about the agent's κ ($\hat{\mu}_1^i$), as well as variance on those beliefs (σ^i), and two participant-specific parameters, ω and β. Where possible, we validated findings from the HGF model with raw behavioural data not derived from a model.

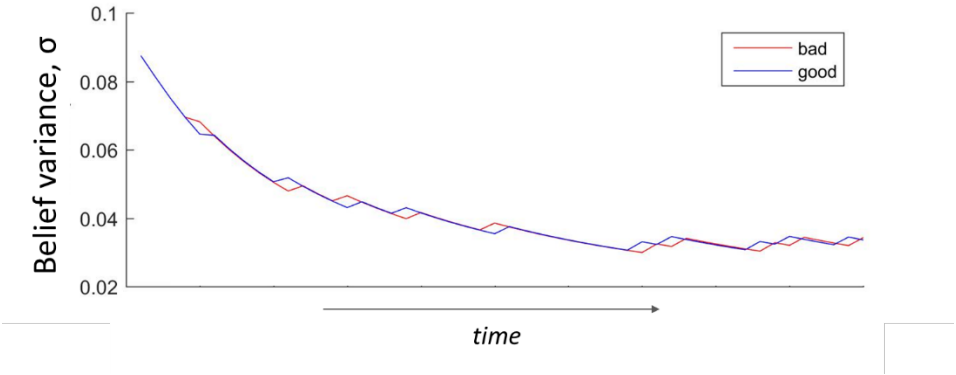
Matching optimal Bayesian trajectories

A main goal of the present research is to investigate whether potential differences in learning about good and bad agents reflects systematic deviations from the performance of a task-local definition of optimal Bayesian learning. Thus, the crucial test is whether human learning differs for good and bad agents in a setting where an ideal Bayesian observer learns identically about these agents. Although we took great efforts to minimize differences in learning that might stem from discrepancies in informational value of trials, it was not possible to eliminate such discrepancies completely. Indeed, even for an optimal observer, the only way two learning trajectories (one for each of two actors) could be identical is if both the trials and the choices made were identical. Thus, small residual discrepancies in informational value about good and bad agents across trials could potentially create learning differences that do not reflect a true asymmetry in learning between agents. Consequently, we generated 100 permutations of trial sequences and simulated behaviour for an ideal Bayesian observer for each (see tapas_fitModel.m and tapas_bayes_optimal_binary_config.m in the HGF toolbox). Two sequences were selected that best minimized differences in our main dependent variables (ω and σ; **Figure 2.5**) for an ideal observer. Each participant in the study was randomly assigned to complete 1 of the 2 trial sequences. With this process in place, we minimized the possibility that any

differences between agents observed could be explained by the order of observations, and instead reflect a systematic deviation from optimal learning.

Figure 2.5 Graphical depiction of belief variance for an ideal Bayesian learner

We ensured that the history of information provided for the good and bad agent did not bias learning for either agent. The graph illustrates that estimates of belief variance decrease over time as information is acquired and residual differences between agents are minimal.



Model comparison

To demonstrate that the HGF model offers a reasonable description of behaviour compared to simpler models, for each study in the subsequent chapters I compare our HGF model to two alternative models: (a) a Rescorla Wagner (RW) model, in which beliefs are updated by prediction errors with a single fixed learning rate (1 learning rate RW), and (b) a Rescorla Wagner model, in which beliefs are updated by prediction errors with separate fixed learning rates for positive and negative outcomes (2 learning rate RW). For details about the alternative models, see **Table 2.2**. For each study, I verify that the log-model evidence (LME) indicates that our model outperforms both a simple single learning rate RW model and a RW model with separate learning rates for positive and negative outcomes. I validate these findings using formal Bayesian Model Selection, which is a random-effects procedure that takes into account inter-subject heterogeneity (Rigoux,

Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). This analysis yields a protected exceedance probability, indicating effectively the probability that the model better explains the data than other models included in the comparison. A protected exceedance probability indistinguishable from 1 indicates a 100% probability that the specified model better explains the data than the other models. For each study, I report the protected exceedance probability from the Bayesian Model Selection.

Table 2.2

Details of alternative models for model comparison

Model	Notes	Estimated parameters
1 Learning rate Rescorla Wagner	Beliefs are symmetrically updated, with a single learning rate for each participant.	α = Learning rate β = Prediction noise
2 Learning rate Rescorla Wagner	Beliefs are asymmetrically updated, with separate learning rates for positive versus negative outcomes, for each participant.	α_{pos} = Learning rate positive outcomes α_{pos} = Learning rate negative outcomes β = Prediction noise
HGF	A two level model, with one estimated parameter governing the volatility of beliefs at the second level, and a second estimated parameter governing the prediction noise.	ω = Tonic volatility β = Prediction noise

2.3 TRUST GAME

In many cases throughout my D.Phil research I was interested in how people use moral preference information to make adaptive social decisions. That is, how do our inferences about other people's moral preferences influence subsequent social interactions? To address these questions, I utilized a standard behavioural economic task, the trust game. In the trust game, an 'investor' is given some amount of money that they can entrust with a 'trustee'. Any amount that is entrusted with the trustee is tripled, and the trustee chooses how much of the tripled amount to return to the investor. Thus, if no money

is entrusted with the trustee, investors keep the initial amount they were given. However, if investors choose to entrust some amount of money then they might receive a higher amount in the end, depending on how much the trustee gives back to them.

For relevant studies, after completing the Moral Inference Task we had participants play the role of the investor in the trust game with each of the agents acting in the role of the trustee. We instructed participants that the percent returned by each agent has been predetermined, and thus the agents are not playing actively. We set the returned amount to correspond to the agents' actual moral preferences, such that the bad agent behaved more selfishly than the good agent (the bad agent returned 20% and the good agent returned 50% of the tripled amount), see **Equation 2.14**. The final amount was paid out to participants as a bonus.

Equation 2.14

Bad agent bonus = (initial amount - amount entrusted) + (amount entrusted* 3 * 0.2)

Good agent bonus = (initial amount - amount entrusted) + (amount entrusted* 3 * 0.5)

2.4 DATA ANALYSIS

In general, data analysis was completed in Matlab (Mathworks) and PASW Statistics 24 (SPSS/IBM). All statistical tests were two-sided. To test whether group mean parameter estimates and mean ratings differed significantly between good and bad agents, I used nonparametric within-subject statistical tests that do not make any assumptions about their underlying distributions (e.g., Wilcoxon signed rank test). Effect sizes were computed for significant results using Rosenthal's formula: $r = Z/\sqrt{n}$, which has been proposed as a viable alternative calculation when the general assumptions of Cohen's formula have been violated (Rosenthal & DiMatteo, 2001). For time-series analyses, I used robust linear

regression models with a bisquare weighting function, while controlling for the effects of time. For other analyses I used nonparametric statistical tests that do not make any assumptions about the underlying distributions of variables (e.g., Spearman's ρ and Mann-Whitney). I report means and standard error of the mean (sem) as mean±sem.

Chapter 3

3 COMPUTATIONAL MECHANISMS OF MORAL INFERENCE

This chapter is an extended version of a paper published as:

Siegel, J.Z., Mathys, C., Rutledge, R.B., & Crockett, M. J. Beliefs about bad people are volatile. *Nature Human Behaviour*, 2.10 (2018): 750.

Doi: https://doi.org/10.1038/s41562-018-0425-1

People form moral impressions rapidly, effortlessly, and from a remarkably young age. Putatively "bad" agents command more attention and are identified more quickly and accurately than benign or friendly agents. Such vigilance is adaptive, but can also be costly in environments where people sometimes make mistakes, because incorrectly attributing bad character to good people damages existing relationships and discourages forming new ones. The ability to accurately infer others' moral character is critical for healthy social functioning, but the computational processes that support this ability are not well understood. Here we show that moral inference is explained by an asymmetric Bayesian updating mechanism where beliefs about the morality of bad agents are more uncertain (and thus more volatile) than beliefs about the morality of good agents. This asymmetry

appears to be a property of learning about immoral agents in general, as we also find greater uncertainty for beliefs about bad agents' non-moral traits. Our model and data reveal a cognitive mechanism that permits flexible updating of beliefs about potentially threatening others, a mechanism that could facilitate forgiveness when initial bad impressions turn out to be inaccurate. Our findings suggest that negative moral impressions destabilize beliefs about others, promoting cognitive flexibility in the service of cooperative but cautious behaviour.

3.1 INTRODUCTION

Signs of bad character capture attention (Baumeister et al., 2001; Fiske, 1980; Pratto & John, 1991; Skowronski & Carlston, 1989) because people are strongly motivated to avoid being exploited by others (Cosmides & Tooby, 1992; Johnson et al., 2013). However, erroneously inferring bad character can lead people to prematurely terminate valuable relationships and thereby miss out on the potential benefits of future cooperative interactions (Axelrod, 2006; Johnson et al., 2013; McCullough, 2008; Molander, 1985). Thus, successfully navigating social life requires strategies for maintaining social relationships even when others behave inconsistently and sometimes commit immoral acts.

One possible strategy is to respond to defection with probabilistic cooperation (Nowak & Sigmund, 1992). Evolutionary models show such "generous" strategies outcompete strategies that summarily end cooperative relationships in the face of a single betrayal (Fudenberg et al., 2012; Wu & Axelrod, 1995). Generous strategies are also observed in humans playing repeated prisoner's dilemmas where others' intended actions are implemented with noise (Fudenberg et al., 2012). Although evolutionary and economic

mods provide descriptive accounts of these behaviours, the cognitive mechanisms that enable them are not well understood. In particular, the computational processes that support adaptive moral inference in humans are unknown.

We propose that when people form beliefs about others' moral character, their impressions about bad agents are more uncertain than their impressions about good agents. This makes impressions about bad agents more amenable to Bayesian updating, by which belief updates are proportional to the uncertainty of beliefs in accordance with Bayes' rule (Mathys et al., 2011). Our hypothesis is based on evidence that threatening social stimuli are arousing (Öhman, 1986), and that arousal increases belief uncertainty in non-social perceptual learning (Nassar et al., 2012). This evidence suggests that threatening social stimuli (such as agents with inferred bad character) might induce belief uncertainty. Our proposal provides a possible solution for maintaining social relationships when others sometimes act immorally by enabling negative impressions to be more easily revised: if beliefs about putatively "bad" agents are volatile, such beliefs could be readily updated if the initial impression turned out to be mistaken.

At first blush, our hypothesis may appear inconsistent with decades of research in social psychology, much of which has examined impression formation from narrative descriptions of extreme and rare behaviours, such as theft or violence. This work provides evidence for a negativity bias in impression formation, where people update their moral impressions to a greater degree from negative relative to positive information (Baumeister et al., 2001; Reeder & Coovert, 1986; Skowronski & Carlston, 1989). The primary explanation for this valence asymmetry is that it reflects a differential diagnosticity of immoral vs. moral behaviours: bad people often behave morally, but good people rarely

behave immorally (Skowronski & Carlston, 1989). Indeed, recent work has suggested that valence asymmetries in impression updating can be explained by perceptions of how rare immoral behaviours are, relative to moral ones (Mende-Siedlecki et al., 2013). This leaves open the question of whether people actually learn differently about agents inferred to be more vs. less moral when their actions are equally diagnostic of their underlying character. This is the central question we addressed in the current studies. We focused on moral inference from behaviours that are not extreme or definitive of character. Such behaviours comprise the vast majority of our daily social interactions: we most often judge others based on behaviours that are nasty or nice, not evil or saintly. Inferring character from minor slights or small favours is considerably more difficult than doing so from criminal deeds or heroic actions, but our success as a social species suggests we are nevertheless able to do this effectively.

Across several studies we applied the Moral Inference Task to investigate the computational basis of moral inference and its temporal dynamics. Our approach extends previous methods employed to probe moral impression formation in several ways. First, because our paradigm used a computational model of moral preferences rather than narrative descriptions of behaviours (as in past social psychology research), we were able to very tightly control how informative agents' behaviours were with regard to their underlying preferences. We precisely matched the trial sequences with respect to how much information was provided about each agent's character over the course of learning (see Chapter 2.1-2.2 for details). In this way, we ensured that the statistics of the environment did not advantage learning about either the good or bad agent, and this symmetry was confirmed by the fact that an ideal Bayesian observer learned similarly

about the good and bad agents across trial sequences (see Chapter 2.2.4, page 60). Because of this design feature, we can confidently infer that any belief asymmetries we observed in our studies cannot be attributed to asymmetries in the information we provided to participants (in contrast to past studies using narrative descriptions of behaviours, where moral information was evaluated as less diagnostic than immoral information (Mende-Siedlecki et al., 2013). Second, in contrast to past work, which focused on descriptive measures over relatively few trials, our methods allowed us to measure the dynamics of impression formation over time. Finally, our paradigm allowed us to measure the *uncertainty* and *volatility* of people's impressions in addition to the valence of those impressions, which has been the primary focus of past work. By doing so, we are able to bridge our investigation of moral inference with foundational work on perceptual and reinforcement learning (Mathys et al., 2011; Nassar et al., 2012) and show that similar computational principles underlie learning across these diverse domains (Behrens et al., 2008; Diaconescu et al., 2014; Hackel, Doll, & Amodio, 2015).

3.2 STUDY 1: BELIEFS ABOUT BAD PEOPLE ARE VOLATILE

3.2.1 METHODS

Participants

Study 1 took place at the Department of Experimental Psychology, University of Oxford and was approved by the Oxford research ethics committee (MS-IDREC-C1-2015-001). Thirty-nine participants were recruited from the Oxford Psychology Research recruitment scheme. Participants with a history of systemic or neurological disorders,

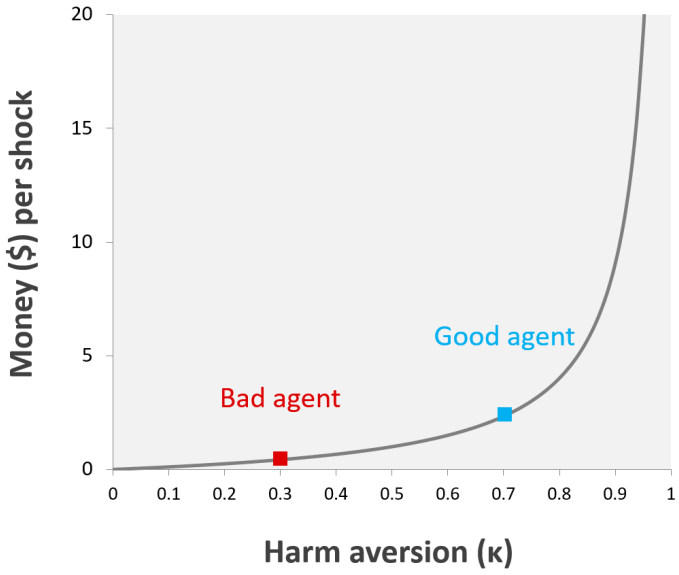
psychiatric disorders, medication/drug use, pregnant women, and more than a years' study of psychology were excluded from participation. All participants provided informed consent prior to initiation of the study and were compensated for their time. One participant was excluded from the analysis as their performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 38 participants. We confirm the pattern of results is similar when we include all participants in **Appendix A-E**.

Experimental procedure

Descriptions of the Moral Inference Task and the computational model are available in Chapter 2. As an overview, participants predicted and observed the choices of two "agents" who repeatedly decided whether to inflict painful electric shocks on another person in a different room in exchange for money (**Figure 2.1**). We generated agent behaviour using a model that accurately captures typical preferences in this choice setting (Crockett et al., 2014, 2017). The model includes a "harm aversion" parameter, κ, which quantifies the subjective cost of harming the victim as an exchange rate between money and pain and ranges from 0 (profit maximizing) to 1 (pain minimizing) (**Figure 3.1**). The two agents differed substantially in their harm aversion, with the "good" agent requiring more compensation per shock to inflict pain on others than the "bad" agent (bad: κ =0.3 or \$0.43 per shock; good: κ =0.7 or \$2.40 per shock; **Figure 2.2**).

Figure 3.1 Trade-off between money and shocks as a function of harm aversion

Money per shock is plotted against the harm aversion parameter, κ, which can range from 0 to 1. The bad agent requires less money per shock than the good agent.



On each trial, participants predicted the choice made by the agent and received immediate feedback on their accuracy. After every third trial, participants rated their subjective impressions of the agent's morality on a scale ranging from "nasty" to "nice" and rated how uncertain they were about their impression on a scale ranging from "very certain" to "very uncertain" (**Figure 2.1**) .

We modeled participants' predictions for each agent separately with a Bayesian reinforcement learning model (Mathys et al., 2011) that generated a global estimate of belief volatility (parameter, ω) that describes the rate at which beliefs evolve over time (**Figure 2.4**). We use the term volatility here to be consistent with previous work using a similar model (Diaconescu et al., 2014; Mathys et al., 2011; Vossel et al., 2014) and because the volatility parameter in our model captures how rapidly beliefs change. Belief volatility is set in log space and is monotonically related to belief uncertainty as measured

by belief variance, σ (i.e., more uncertain beliefs are more volatile (Mathys et al., 2011); for example, a change in ω from -3.5 to -4.0 corresponds to a 20% decrease in the average variance of posterior beliefs, σ).

3.2.2 RESULTS

Model validation and manipulation checks

We verified that the log-model evidence (LME) indicated that our model outperforms both a simple single learning rate RW model and a RW model with separate learning rates for positive and negative outcomes. See **Table 3.1** for LME for individual models for all studies in Chapter 3. We validated these findings using formal Bayesian Model Selection. To this end, we used LME data across all studies in Chapter 3 (N = 1419) to compare between the HGF and our two RW models. This analysis yielded a protected exceedance probability indistinguishable from 1 for the HGF model for both agents, indicating effectively a 100% probability that the HGF model better explains the data than the other models included in the comparison. Thus, the HGF outperformed simpler Rescorla-Wagner models that do not allow dynamic learning rates that account for the uncertainty of beliefs.

Table 3.1

Model comparison, Chapter 3. Sum log-model evidence (LME) for each study

	HGF	1 learning rate RW	2 learning rate RW
Study 1	-1662.70	-1855.68	-1905.96
Study 2	-7599.38	-7849.20	-7759.33
Study 2s	-4375.62	-4859.85	-5250.32
Study 3	-7291.61	-7543.39	-7099.73
Study 4 morality	-4050.18	-4820.50	-5230.21
Study 4 competence	-3983.84	-5315.96	-5791.98
Study 5 morality	-8149.24	-9252.11	-9916.40
Study 5 competence	-6021.07	-7167.02	-8085.64
Study 6	-5361.11	-5567.04	-5568.75
Total	-48494.75	-54230.75	-56608.30

We investigated whether we were able to recover participant's choices using the estimates derived from the model. To this end, we simulated 100 sequences of choices using each participant's parameter estimates and compared the simulated choices to participants' actual predictions. The model fit participants' predictions well, explaining behaviour with a mean 87.0% accuracy for the bad agent and a mean 86.3% accuracy for the good agent in Study 1 (see **Table 3.2** for details of model goodness-of-fit for all studies in Chapter 3).

Table 3.2

Model accuracy (%)

		bad	bad s.d.	good	good s.d.	average
Study 1*		87.00	4.842	86.29	5.375	86.65
Study 2		75.00	7.996	68.70	9.561	71.85
Study 2s		77.31	7.726	77.76	9.760	77.54
Study 3		69.40	7.521	64.50	9.257	66.95
Study 4	morality	77.59	7.374	78.92	8.722	78.26
	competence	80.68	9.659	77.94	6.877	79.31
Study 5	morality	77.70	6.976	79.10	9.014	78.40
	competence	76.30	9.496	76.50	9.412	76.40
Study 6		73.44	10.971	66.59	12.309	70.02
Average		77.16		75.14		76.15

*Study 1 was conducted in the laboratory and included participants recruited from Oxford's Psychology Research recruitment scheme. All subsequent studies were conducted online, with participants recruited from Amazon's Mechanical Turk (MTurk). Because MTurk studies typically feature a larger amount of noise, the model is more accurate for Study 1 than subsequent studies.

Next, we investigated whether participants indeed learned about the agents' moral preferences in the task. We analyzed the model's final estimates of participants' beliefs about each agent's κ ($\hat{\mu}_1^{50}$), and verified that participants formed beliefs that closely resembled the agent's true κ and significantly differed from one another (mean±SEM bad: 0.322±0.004; good: 0.681±0.004; Z = -5.373, $p < 0.001$; **Appendix A:**). That is, by the end of the task, participants accurately learned that the bad agent required approximately £0.43 per shock to the victim and the good agent required approximately £2.40 per shock to the victim. Subjective character ratings confirmed that participants inferred distinct moral character for agents with different moral preferences in the task; final ratings indicated the good agent was generally characterized as nice and the bad agent as nasty (bad: 0.427±0.040; good: 0.789±0.029; Z = -5.303, $p < 0.001$; **Appendix D**:), where ratings above 0.5 are classified as 'nice' and ratings below 0.5 are classified as 'nasty'.

Subjective uncertainty and learning rates

In line with our predictions, participants exhibited faster updating for the bad agent than the good agent, as demonstrated by a larger ω (bad: -3.779±0.102; good: -4.212±0.104; $Z = 3.212$, $p = 0.001$; **Appendix B:**). Subjective uncertainty ratings corroborate the findings from the model, as participants reported greater uncertainty about their characterizations for the bad agent (bad: 28.62 ±2.428; good: 20.612±1.371; $Z = 3.444$, $p < 0.001$; **Appendix E**:). Taken together Study 1 suggests that moral inferences modulate the computational and cognitive mechanisms for updating beliefs, which may stem from a reduced reliance on priors for more harmful agents.

In a supplementary analysis, we investigated whether observed differences in ω could be explained by differences in β or an overlap between β and ω. Across studies, we found no consistent relationship between ω and β (see **Table 3.3**). In addition, we found no consistent differences in β between good and bad agents across studies (see **Appendix C:**).

Table 3.3

Correlation between model free parameters ω and β. Analysis investigating (a) the relationship between ω and β for the bad agent, (b) the relationship between ω and β for the good agent, (c) the relationship between Δω and Δβ between good and bad agents

	Bad agent		**Good agent**		**Bad - Good**	
Study	ρ	P	ρ	P	ρ	P
Study 1	0.002	0.992	-0.008	0.960	0.315	0.055
Study 2	0.272	0.000	0.316	0.000	0.472	0.000
Study 2s	-0.175	0.061	-0.507	0.000	-0.039	0.673
Study 3	0.288	0.001	0.156	0.072	0.338	0.000
Study 4, morality	-0.292	0.002	-0.434	0.000	-0.229	0.017
Study 4, competence	-0.331	0.000	-0.408	0.000	-0.171	0.071
Study 5, morality	-0.131	0.073	-0.039	0.596	0.098	0.180
Study 5, competence	-0.288	0.000	-0.253	0.000	0.016	0.830
Study 6*	0.363	0.000	0.118	0.117	n/a	n/a

*Because Study 6 is between-subject and we did not have a within-subject manipulation of agent, we cannot compute the difference between agents (Δω and Δβ).

3.3 STUDY 2: REPLICATION AND EXTENSION

An advantage of Bayesian models is that they provide an *optimal* calculus for combining uncertain prior beliefs with new information. Using Bayesian decision theory, we can then simulate an ideal Bayesian learner that uses this information to choose actions that maximize task performance by minimizing uncertainty. Owing to these properties, ideal Bayesian learners perform a task as well as the task can be performed, meaning that the performance of an ideal Bayesian learner on a task defines optimal behaviour in that task. In Study 2, we aimed to replicate our findings in a larger sample and compare participant behaviour with an ideal Bayesian learner.

3.3.1 METHODS

Participants

Two-hundred and fifty-three U.S. residents were recruited from Amazon's Mechanical Turk (AMT (Horton, Rand, & Zeckhauser, 2011)). The study utilized the web

application framework Ruby on Rails and was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Conducting the study online had the advantage of engaging a large number of diverse respondents outside of the University's limited subject pool. All participants provided informed consent and were compensated for their time. Eighty-seven participants were excluded from the analysis as their behavioural performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 163 participants. We confirm the pattern of results is similar when we include all participants in **Appendix A-E**.

Experimental procedure

We constructed the learning task for Study 2 using identical procedures to those outlined in Chapter 2 and applied in Study 1. In order to motivate accurate predictions for the online task, participants in Study 2 were explicitly instructed to pay attention and learn about the behaviour of the agents, as they would later have to decide whether to trust the agents in a one-shot investment game (Berg, Dickhaut, & McCabe, 1995) that could earn them additional money. A description of the trust game is available in **Chapter 2.3**.

3.3.2 RESULTS

First, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences in the task. We analyzed the model's final estimates about each agent's κ ($\hat{\mu}_1^{50}$), and verified that participants formed beliefs that closely resembled the agent's true κ (bad: 0.301±0.004; good: 0.707±0.003; **Appendix A**; See **Figure 3.2** for a graphical depiction of the temporal evolution of $\hat{\mu}_1$). Subjective character

ratings indicated participants inferred the good agent generally had a nice character and the bad agent a nasty character (bad: 0.422±0.020; good: 0.807±0.014; **Figure 3.3a** and **Appendix D**).

Figure 3.2 Graphical depiction of temporal evolution of belief estimates

Trajectory of model estimates of inferred beliefs ($\hat{\mu}$) about each agent's κ for each trial, averaged across all participants (solid lines) and for an optimal Bayesian learner (dotted lines).

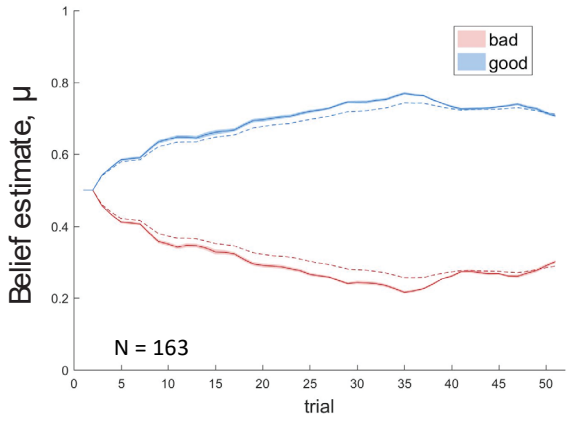
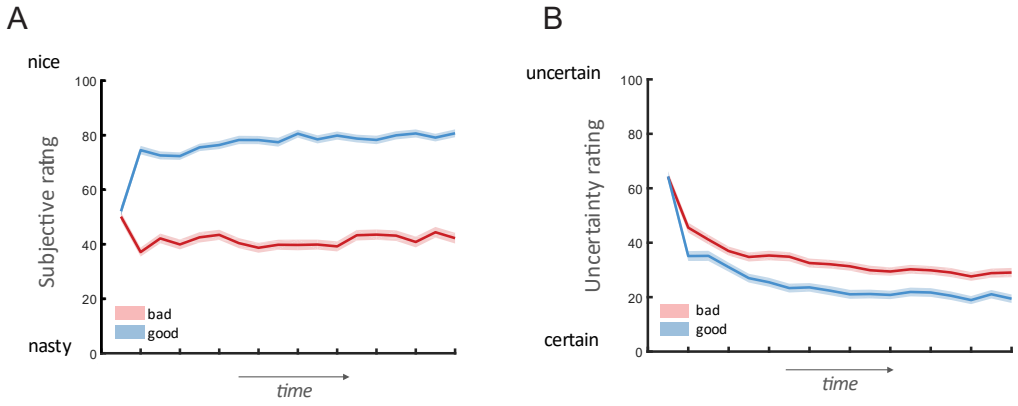


Figure 3.3 Trial-by-trial subjective ratings of the good and bad agent

*(**A**) Trajectory of subjective character ratings over time in Study 2, averaged across participants (N=163 for both panels). (**B**) Trajectory of subjective uncertainty ratings over time, averaged across participants. Subjects reported greater uncertainty about bad agents. Shaded bounds in trajectories represent SEM.*



Next, we examined whether participants trusted the good agent to a greater extent than the bad agent by comparing the amount participants entrusted to each agent in the trust game. As expected, participants entrusted significantly more with the good agent (good: 7.74±0.24) than the bad agent (bad: 3.82±0.27; $Z = -8.522$, $p < 0.001$; **Table 3.4** and **Figure 3.4**). Notably, the extent to which participants adapted their trust behaviour between the good and bad agent was predicted by subjective character ratings; increased character sensitivity (Δjudgment, calculated as final impression of the good agent – final impression of the bad agent) was associated with increasingly differentiated trust behaviour (Δentrust, calculated as amount entrusted with good agent – amount entrusted with bad agent; $\rho = 0.410$, $p < .001$).

Table 3.4

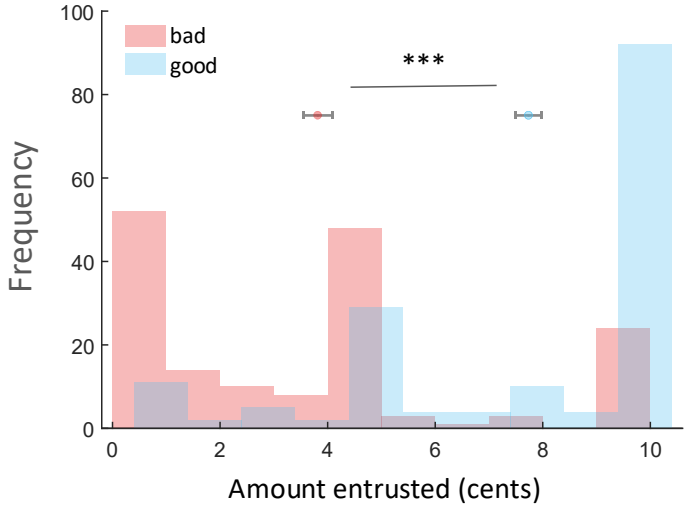
Trust game & agent comparison

		Amount entrusted (mean ± SEM)	*Z*-statistic	p-value	effect size (*r*)
Study 2	bad:	3.82 ± 0.265	-8.522	<0.001	0.667
	good:	7.74 ± 0.243			
Study 2s	bad	3.79 ± 0.354	-5.957	<0.001	0.520
	good	6.97 ± 0.327			
Study 3	bad:	3.36 ± 0.303	-7.589	<0.001	0.653
	good:	7.15 ± 0.295			
Study 4	bad:	2.80 ± 0.337	-7.034	<0.001	0.674
	good:	7.72 ± 0.326			
	low-skill:	6.02 ± 0.380	2.967	0.003	0.280
	high skill:	5.19 ± 0.381			
Study 5	bad:	2.70 ± 0.241	-10.112	<0.001	0.736
	good:	7.90 ± 0.234			
Study 6*	bad:	6.59 ± 0.259	-2.055	0.040	0.161
	good:	7.32 ± 0.248			

* between-subjects sign test

Figure 3.4 Amount entrusted in the trust game

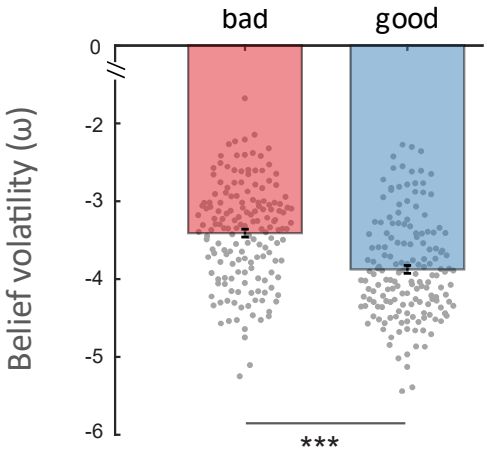
In a one-shot trust game, participants entrusted the good agent with twice as much money as they did the bad agent.



Subjective uncertainty ratings indicated that participants reported greater uncertainty about their characterizations of the bad agent on average (bad: 33.078±1.330; good: 24.087±1.371; $Z = 7.213$, $p < 0.001$; **Appendix E:**). Greater uncertainty about the bad agent effectively translated into faster updating as demonstrated by a larger tonic volatility ω (bad: -3.411±0.050; good: -3.877±0.051; $Z = 6.830$, $p < 0.001$; **Figure 3.5** and **Appendix B:**). Effectively, this means that prediction errors were ascribed greater weights when updating beliefs about the bad agents' harm preferences than the good agents' harm preferences.

Figure 3.5 Belief volatility (ω) model estimates, Study 2

*Volatility of beliefs (ω in the model) was higher for the moral character of the bad compared to the good agent. Error bars represent SEM. ***P < 0.001*



To investigate whether participants similarly exhibited larger updating in their subjective character ratings for the bad agent, we computed for each participant the absolute change in rating from trial to trial (Δrating), and compared the average Δrating between the two agents. As expected, the average Δrating was larger for the bad agent (bad: 9 .780±0.602; good: 7.909±0.452; Z = 2.787, p = 0.005; **Table 3.5**), indicating a greater tendency to adjust moral impressions in response to new information.

Table 3.5

Trial-wise updating of character ratings (Δrating) and agent comparison. While we did not find larger impression updating for the bad agent relative to the good agent in studies 1 and 3, we see the same pattern of effects. A parametric meta-analysis, including the results from all within-subject studies (thus, excluding Study 6 and the competence condition from study 4) yielded significant results (Z = 7.461, p < 0.001).

Study	Bad agent		Good agent		test-statistic	p-value
	mean	SEM	mean	SEM		
Study 1	7.369	0.655	6.880	0.759	0.558	0.577
Study 2	9.780	0.602	7.909	0.452	2.787	0.005
Study 2s	7.821	0.478	5.539	0.472	5.677	0.000
Study 3	10.103	0.610	9.577	0.522	1.140	0.254
Study 4, morality	8.933	0.679	6.840	0.453	2.579	0.010
Study 4, competence	5.818	0.477	6.707	0.553	-2.299	0.021
Study 5, morality	7.183	0.367	6.477	0.358	2.433	0.015
Study 5, competence	6.461	0.229	5.348	0.240	5.407	0.000
Study 6*	12.445	0.628	10.615	0.602	2.460	0.014

*Between-subjects sign test

Our data showed that beliefs were more uncertain when observing a bad agent relative to a good agent, and this was accompanied by a faster learning rate. An additional goal of Study 2 was to investigate whether participant's behaviour deviated from that of an ideal Bayesian observer. Although we took great efforts to minimize differences in optimal learning parameters between agents in our trial sequences, we additionally pursued post-hoc testing as validation. To this end, we computed for each participant the difference in parameter estimate, ω, between the good and bad agent for our main dependent measure (Δω) and compared this to the difference in ω for an ideal learner observing the same trials. We found that human learning in this setting significantly differed from Bayes-optimal learning, as the effective difference between agents was significantly greater in our sample for estimates extracted from the model (Δω: Z = -7.383, $p < 0.001$; **Table 3.6**). To graphically illustrate the differences between human and ideal Bayesian learning in this setting, we extracted precision-weighted updates for an exemplary participant and

compared this to estimated precision-weighted updates for an ideal Bayesian learner in **Figure 3.6**.

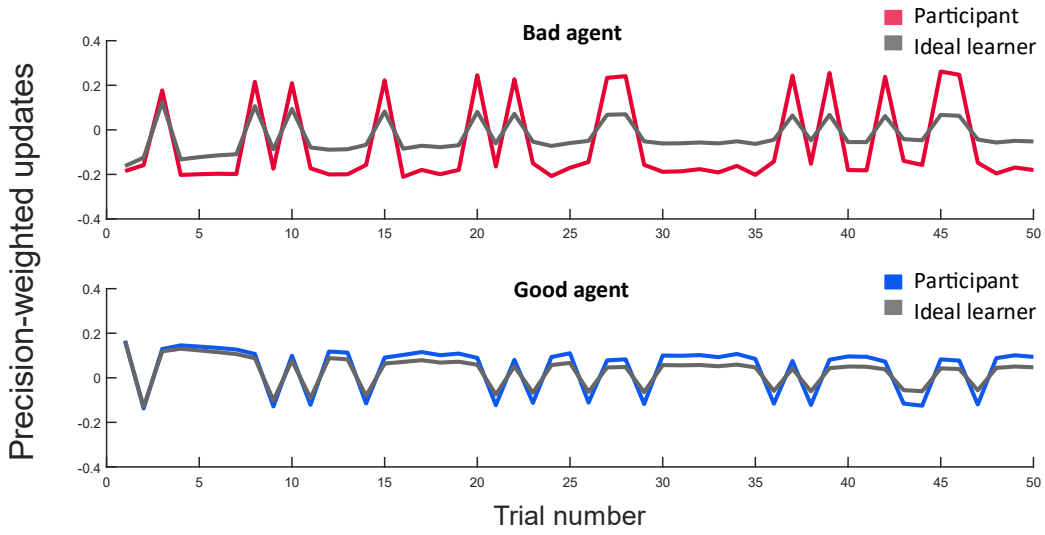
Table 3.6

Δω statistical comparison between human learning and ideal Bayesian

Study	Subject or Bayesian	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	Subject	0.433 ± 0.121	3.200	0.001	0.519
	Bayesian	0.015 ± 0.011			
Study 2	Subject	0.446 ± 0.061	7.382	<0.001	0.578
	Bayesian	0.021 ± 0.012			
Study 2s	Subject	0.304± 0.067	5.090	<0.001	0.473
	Bayesian	-0.037 ± 0.015			
Study 3	Subject	0.506 ± 0.056	7.659	<0.001	0.659
	Bayesian	-0.007 ± 0.002			
Study 4	Subject morality	0.316 ± 0.069	4.579	<0.001	0.439
	Bayesian morality	-0.049 ± 0.015			
	Subject competence	-0.060 ± 0.069	-1.266	0.206	N.S.
	Bayesian competence	0.045 ± 0.015			
Study 5	Subject morality	0.313± 0.059	4.558	<0.001	0.332
	Bayesian morality	0.042 ± 0.002			
	Subject competence	0.103 ± 0.042	3.277	0.001	0.238
	Bayesian competence	0.001 ± 0.003			

Figure 3.6 Graphical depiction comparing human learning to ideal Bayesian

Prediction errors were ascribed greater weights when updating beliefs about the bad agents' harm preferences than the good agents' harm preferences. Asymmetric belief updating was significantly greater in human participants compared to an ideal Bayesian learner.



In a subsequent analysis, we asked whether behaviour using a simpler reinforcement learning mechanism would lead to a similar pattern of results identified in our HGF model. That is, we aimed to determine whether larger ω for bad agents could be recovered from behaviour based on simpler Rescorla Wagner models. To investigate, we first extracted the average parameter estimates from the two Rescorla Wagner models outlined in our model comparison, Chapter 2.2, Model Comparison. The first model included a single fixed learning rate and a noise parameter, while the second model included separate learning rates for positive and negative outcomes and a noise parameter.

We next simulated behaviour for 1000 fake participants on the set of trials from Study 2, whose parameter estimates were drawn from distributions with means equal to the extracted average parameter estimates and standard deviations equal to the standard

deviation of those estimates. Next, we fit our HGF model to the simulated data and checked whether our observed effect (larger ω for bad relative to good agent), could be recovered by fitting the behaviour of 'participants' who were actually behaving in a manner consistent with either of these Rescorla-Wagner models.

When we simulated behaviour based on either Rescorla-Wagner updating process, this did not lead to the same parameter differences observed in our data. While our data showed larger volatility estimates for the bad agent than the good agents, the data simulated using the Rescorla Wagner models led to larger volatility estimates for the good agent than the bad agent (Single learning rate RW ω: bad = -4.250±0.005; good = -4.152±0.006; 2 learning rate RW ω: bad = -4.284±0.001; good = -4.173±0.004). Thus, in addition to the fact that the Rescorla-Wagner models do not fit participants' behaviour as well as the HGF model, these alternative learning models do not lead to the same volatility parameter differences we observed in our data.

To ensure that our findings in Study 1 and 2 were not an artifact of the scale participants used to rate the agents' morality (ranging from *nasty* to *nice*), we replicated all findings in a supplementary study (Study 2s; N = 125) using an alternative scale (ranging from *bad* to *good*). Participants indicated greater subjective uncertainty in their impression of the bad, relative to good, agent (bad: 29.209±1.485; good: 24.602±1.474; Z = 3.207, p = 0.001; **Appendix E:**). This translated into faster updating for the bad agent, as demonstrated by a larger ω (bad: -4.303 ±0.064; good: -4.608 ±0.060; $Z = 4.16$, $p < 0.001$; **Appendix B:**). Thus, we can be certain that our findings are not dependent on the specific labels that were used on the scale to rate the agents' moral character.

3.4 STUDY 3: ADDING NOISE TO AGENTS' CHOICE BEHAVIOUR

For Study 1 and 2, agents were simulated to behave deterministically, never deviating from their preference towards harming the victim. In other words, agents deterministically chose the more harmful option when V_{harm} in **Equation 2.1** was greater than zero, given the agent's harm aversion, κ, and deterministically chose the less harmful option when V_{harm} was smaller than zero. However, human behaviour is not always consistent, especially when choices become increasingly difficult (i.e., when V_{harm} is close to zero). Consequently, in a third study we increased the stochasticity of agent choices to test whether the differences we observed for learning about bad compared to good agents are robust to noisy environments.

3.4.1 METHODS

Participants

One-hundred and sixty-two U.S. residents were recruited from AMT. All participants provided informed consent and were compensated for their time. Study 3 was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Twenty-seven participants were excluded from the analysis as their behavioural performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 135 participants. We confirm the pattern of results are similar when we include all participants in **Appendix A-E**.

Experimental procedure

In general, the experimental procedure for Study 3 was very similar to Study 1 and 2 (including the instructions). However, in order to simulate agents that did not behave deterministically, we fixed β in **Equation 2.2** to 1.5 instead of 100. As was discussed in Chapter 2.1.2, β defines the linearity of the sigmoid function in **Equation 2.2**. Decreasing β thus increased the linearity of the slope, meaning that agents often behaved in a manner that was not consistent with their harm aversion preference, κ.

Again, we minimized the possibility that any differences between agents observed in Study 3 could be explained by the order of observations using the methods described in Chapter 2.1 and 2.2.

3.4.2 RESULTS

Study 3 replicated all the findings from Study 1 and 2. Again, participants indicated greater subjective uncertainty in their impression of the bad, relative to good, agent (bad: 35.609±1.432; good: 30.858±1.665; $Z = 3.896$, $p < 0.001$; **Appendix E:**). This translated into faster updating for the bad agent, as demonstrated by a larger ω (bad: -3.468±0.042; good: -3.974±0.043; $Z = 7.296$, $p < 0.001$; **Appendix B:**).

3.5 STUDY 4: INFERRING MORALITY VERSUS COMPETENCE

Study 3 demonstrated that the observed asymmetry in learning between good and bad agents is robust to a setting where agents' choices are noisy rather than deterministic. A remaining question is whether the effect is triggered by negative traits more generally, or specifically by negative *moral* traits. We tested this in a fourth study where we examined

whether the asymmetry in learning about bad compared to good agents extend to learning about a trait unrelated to morality. If the asymmetry is specific to *moral* impressions, then it should be larger when learning about moral character than when learning about a non-moral trait such as competence.

3.5.1 METHODS

Participants

Two-hundred and eighty U.S. residents were recruited from AMT and randomized to either a morality condition or a competence condition. All participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Thirty-one participants from the morality condition and twenty-nine participants from the competence condition were excluded from the analysis as their behavioural performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 109 participants in the morality condition and 111 participants in the competence condition. We confirm the pattern of results is similar when we include all participants in **Appendix A-E**.

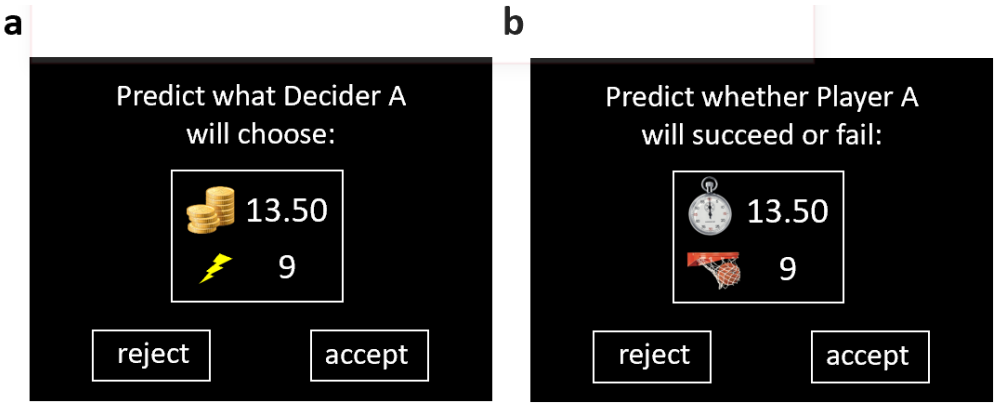
An *a priori* power analysis indicated that the study required 104 participants in each condition to have 80 percent power to detect a moderate effect (0.4) in a nonparametric between-groups analysis. Thus, our study was sufficiently powered to observe an effect in our between-groups design.

Experimental Procedure

Like our previous studies, participants randomized to the morality condition predicted the moral choices of a bad agent ($\kappa = 0.3$) and a good agent ($\kappa = 0.7$). Instead of predicting which of two options was chosen by the agents, in Study 4 participants predicted whether the agents would accept or reject a sequence of offers of a certain amount of money at the expense of a certain number of shocks to the victim (**Figure 3.7a**). Thus, if the offer is accepted, the agent receives the indicated amount of money and the victim receives the indicated number of shocks. However, if the offer is rejected, the agent receives no money and the victim receives no shocks. Trial sequences were created in a similar manner to Study 2, where we created pairs of trials that were matched in their informational value for the good and bad agent (**Figure 2.3**) and presented minimal differences in learning trajectories for an ideal Bayesian observer. As in Studies 1 and 2, β in **Equation 2.2** was fixed to 100 to simulate agents that were completely deterministic in their behaviour.

Figure 3.7 Experimental design, morality vs. competence learning

*In the morality condition (**A**), participants predicted whether 'agents' ('Decider A' and 'Decider B') accepted or rejected offers of a certain amount of money in exchange for a certain number of shocks to an anonymous victim. In the competence condition (**B**), participants predicted whether 'agents' ('Player A' and 'Player B') succeeded or failed at scoring a certain number of points in a certain amount of time in a series of basketball games.*



In the competence condition, participants predicted whether agents would succeed or fail at scoring a certain number of points in a certain amount of time in a series of basketball games (**Figure 3.7b**). To manipulate competence, we created agents who differed in their ability to score points. This was parameterized as tau (τ) and represents the agent's skill level. When $\tau = 0$, agents are extremely skilled and can effortlessly score points in minimal amounts of time; as τ approaches 1, agents become weakly skilled and require increasing amounts of time to score a single point. We created the agents to behave identically to the agents in the morality condition, such that one agent had a low basketball skill ($\tau = 0.7$) and the other agent had a high basketball skill ($\tau = 0.3$). Effectively, this meant that the low-skill agent required more time to score each point than the high-skill agent. The trial sequences for the competence condition were identical to the trial sequences for the morality condition. Thus, accepting an offer of \$15.50 in exchange for 4 shocks in the morality condition was analogous to successfully scoring 4 points in 15.50 minutes in the competence condition. This means that the structure of the task was identical between morality and competence conditions. This ensured that any differences observed between the two conditions could only be driven by differences in the *framing*.

As in the previous studies, on every third trial, participants indicated their general impression of the agent's morality (or basketball skill) on a scale ranging from 0 = nasty (or beginner) to 1 = nice (or expert), and how uncertain they were about each impression.

For Studies 2 and 3, participants were motivated to learn about the moral character of the agents because they were instructed that they would later have to decide whether to trust the agents in a one-shot trust game that could earn them additional money. Because this motivation was less relevant to participants randomized to the competence condition,

we chose to omit this instruction entirely for Study 4. Instead, participants were instructed to learn well about the behaviour/performance of the agents because the more accurate their predictions, the more money they could gain. This money was paid out to participants as a bonus after completing the task. The trust game was additionally included at the end of the task to be used as a manipulation check (see **Table 3.4** for results).

We chose to examine learning about basketball ability rather than other traits related to competence, such as intelligence or social ability, because previous work has shown that the latter are not independent of impressions of moral character (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; Rosenberg, Nelson, & S, 1968). In contrast, we expected inferences about basketball ability to be independent from inferences about moral character. This was supported in a supplementary pilot study (N=97), where participants rated the moral character and athleticism of two agents who decided how much money they were willing to pay to prevent an anonymous victim from receiving 10 painful electric shocks. One agent (the *bad* agent), indicated that they would require $4.30, and the other agent (the *good* agent), indicated that they would require $23.40. The order that the agents were presented was randomized across participants. As expected, participants rated the bad agent as significantly less moral than the good agent (bad: 37.07±2.226; good: 81.47±2.300; $Z = -8.027$, $p < 0.001$), but rated the two agents similarly on athleticism (bad: 43.66±1.792; good: 44.91±1.801; $Z = -0.260$, $p = 0.795$). Thus, our design allowed us to directly test the specificity of our observed effect for moral inference because it is unlikely that impressions of morality and impressions of basketball ability are associated.

Statistical analysis

The main goal in Study 4 was to investigate whether the observed asymmetry in learning about bad versus good agents was specific to *moral* inference. Consequently, we computed for each participant the difference between good and bad agents for each of our main dependent measures ($\Delta\omega$, Δ uncertainty rating) and compared this to the difference between low-skill and high-skill agents using non-parametric statistical tests for independent samples.

3.5.2 RESULTS

As a manipulation check, we examined whether participants trusted the good agent to a greater extent than the bad agent by comparing the amount participants entrusted to each agent in the trust game. As expected, participants in the morality condition entrusted significantly more with the good agent (good: 7.62±0.326) than with the bad agent (bad: 2.80±0.337; Z = -7.034, $p < 0.001$; **Table 3.4**). Participants in the competence condition entrusted slightly more to the low-skill agent (low-skill: 7.65±0.507) than the high-skill agent (high-skill: 6.03±0.380; Z = 2.967, $p = 0.003$), but the difference in trust for high- versus low-skill agents was significantly smaller than the difference in trust for the good versus bad agents (morality, good - bad: 3.368±0.790; competence, low-skill - high-skill: 0.826±0.291; Z = -3.608, $p < 0.001$).

Next, we analyzed the model's final estimates about each agent's κ ($\hat{\mu}_1^{50}$) for the morality condition, and verified that participants formed beliefs that closely resembled the agent's true κ (bad: 0.287±0.002; good: 0.715±0.002; Z = -9.062, $p < 0.001$; **Appendix A:**). Final estimates about each agent's τ ($\hat{\mu}_1^{50}$) in the competence condition also verified

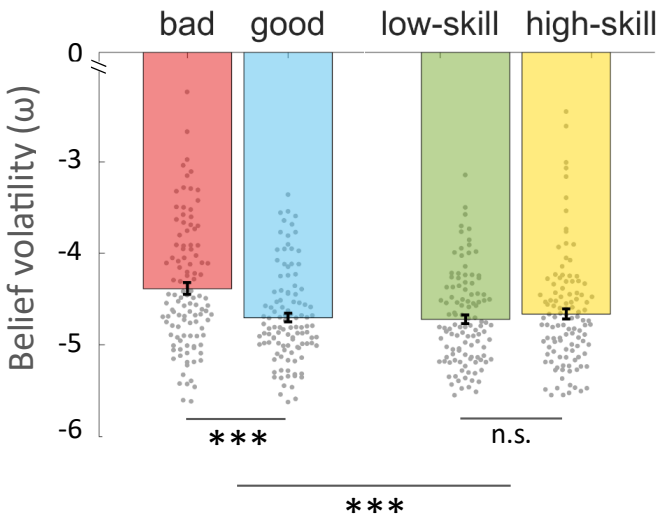
that participants formed beliefs that closely resembled the agent's true τ (low-skill: 0.714±0.002; high-skill: 0.288±0.002; $Z = 9.145$, $p < 0.001$;). Final subjective ratings indicated the good agent was generally characterized as nicer than the bad agent (bad: 0.362±0.024; good: 0.770±0.020; $Z = -8.091$, $p < 0.001$; **Appendix D:**), and the high-skill agent was characterized as more experienced in basketball than the low-skill agent (low-skill: 0.153±0.017; high-skill: 0.787±0.014; $Z = -9.113$, $p < 0.001$;).

A rank sum test found a significant difference in Δ uncertainty ratings between the morality condition and the competence condition. Specifically, the magnitude of the difference in subjective uncertainty ratings between agents was significantly greater in the morality condition compared to the competence condition (morality condition: 6.082±1.802; competence condition: -2.129±1.872; $Z = 4.118$, $p < 0.001$). Simple effects analysis demonstrated that subjective uncertainty was significantly greater for the bad agent, relative to the good agent in the morality condition (bad: 29.335±1.598; good: 24.165±1.607; $Z = 3.649$, $p < 0.001$; **Appendix E:**). No significant differences in subjective uncertainty were observed between agents in the competence condition (low: 18.457±1.227; high: 20.653±1.274; $Z = -1.775$, $p = 0.076$;).

The magnitude of Δω was also greater in the morality condition, relative to the competence condition (morality condition: 0.324±0.069; competence condition: 0.060±0.069; $Z = 3.392$, $p < 0.001$, **Figure 3.8**). Simple effects analysis demonstrated a higher ω for the bad agent relative to the good agent (bad: -4.390 ±0.064; good: -4.714±0.048; $Z = 4.219$, $p < 0.001$; **Appendix B:**), however there was no significant difference in ω between low- and high-skill agents (low: -4.726±0.047; high: -4.665±0.057; $Z = -0.574$, $p = 0.566$).

Figure 3.8 Inferring morality versus competence

*Interaction between agent (bad/low-skill vs. good/high-skill) and condition (morality vs. competence) for the volatility of beliefs (ω in the model). Error bars represent SEM. ***P < 0.001; n.s. = not significant*



3.6 STUDY 5: INFERRING MORAL CHARACTER INFLUENCES COMPETENCE LEARNING

Previous work has shown bad behaviours carry more weight than good behaviours in moral impression formation (Baumeister et al., 2001; Fiske, 1980; Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1989). In our studies, the bad agent by definition makes more immoral choices than the good agent, and so we cannot be sure that the observed asymmetry in learning is driven by inferences about the moral *character* of the good and bad agents rather than responses to the *choices* that good and bad agents make. We predicted that the threatening nature of bad agents would increase the uncertainty and volatility of beliefs, thereby destabilizing beliefs in a non-specific manner. This prediction is consistent with past literature showing that task-irrelevant threatening cues increase attention and information processing (Lojowska, Mulckhuyse, Hermans, & Roelofs, 2019;

Robinson, Vytal, Cornwell, & Grillon, 2013). If inferring bad moral character exerts a global effect on social impression formation, then beliefs about other traits, such as competence, should also be more volatile for agents that are believed to be immoral. Such a mechanism would be advantageous because it is useful to attend to and learn about all aspects of bad people, in order to build a richer model of those who pose a threat. We tested this hypothesis in a fifth study where participants simultaneously inferred the morality and competence of a good and bad agent with similar levels of competence. Study 4 showed that people do not learn differently about two agents who significantly differed in basketball ability in the competence task. Thus, in Study 5 we implemented this same task to test whether we could manipulate learning and uncertainty about competence as a function of inferences about the agent's moral character.

3.6.1 METHODS

Participants

Two-hundred and fifty-nine U.S. residents were recruited from AMT. Participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Seventy participants were excluded from the analysis as their behavioural performance was below chance for at least one agent (<50% accuracy). Final analysis was carried out on the remaining 189 participants. We confirm the pattern of results is similar when we include all participants in **Appendix A-E**.

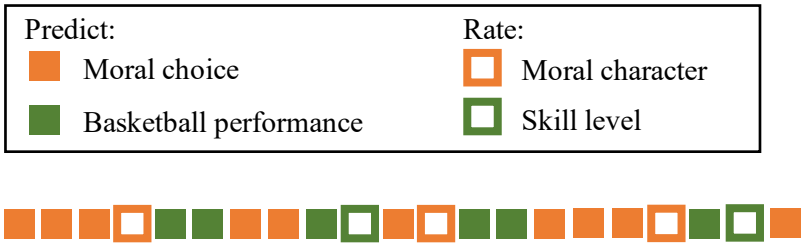
Experimental Procedure

Participants predicted both the moral choices and basketball performance of two agents in Study 5. One agent was characteristically low in morality (bad) and the other was high in morality (good), however both agents were similarly competent in their basketball ability. Trial sequences were made up of 60 'morality' trials and 40 'competence' trials. On morality trials, participants predicted whether the agent would accept or reject an offer of a certain amount of money at the expense of a certain number of shocks to an anonymous victim (**Figure 3.7a**). On competence trials, participants predicted whether the agent would succeed or fail at scoring a certain number of points in a certain amount of time during a basketball game (**Figure 3.7b**).

Trial sequences were created by interleaving morality trials with competence trials such that (a) participants initially predicted three of the agent's moral choices, and (b) every second or third morality trial would be followed by either one or two competence trials (**Figure 3.9**). Across trials we randomized whether competence trials were presented after 2 or 3 morality trials, and whether morality trials were presented after 1 or 2 competence trials. Subjective character and uncertainty ratings were collected following every third morality trial, while subjective competence and uncertainty ratings were collected following every third competence trial.

Figure 3.9 Experimental design, competence learning with moral information

In study 5, participants experienced trial sequences with interleaved morality (Figure 3.7a) and competence trials (Figure 3.7b). Participants rated their impressions of and uncertainty about the agents' moral character and skill level after every third morality and competence trial, respectively.



The morality trial sequences were created using the same procedure as referred to in Chapter 2.1, where one agent was significantly more averse to harming the victim ($\kappa = 0.7$) than the other ($\kappa = 0.3$) with minimal differences in learning trajectories for an optimal Bayesian observer.

Although we wanted the good and bad agent to behave similarly in their basketball performance, we sought to ensure that behaviour was not identical in the event that participants could recall the previous agent's performance and thus more easily predict that of the second agent observed. Consequently, we simulated one agent to be slightly less competent ($\tau = 0.45$) than the other ($\tau = 0.55$), and randomized across participants which competence simulation was paired with which agent (bad versus good). In other words, for half of the participants, the good agent was slightly less competent, while the bad agent was slightly more competent; for the other half, the good agent was slightly more competent, while the bad agent was slightly less competent.

Competence trial sequences were created in a similar manner to morality trial sequences. We first created a set of 19 trials where the values of τ were randomly drawn

from a normal distribution around one agent's indifference point (M = 0.55, s.d. = 0.15). Next, we created a set of 19 matched trials around the other agent's indifference point by subtracting each τ value from 1. Again, we sequentially paired trials that were matched in their informational value for each of the two agents (as in **Figure 2.3**), and randomized the order of presentation of each member of the pair. The pairs comprised trials 2-39 of the sequence, while the initial and final trials were fixed to $\tau = 0.5$.

As in Study 4, participants were instructed to learn well about the behaviour/performance of the agents because they would receive a financial bonus in proportion to the accuracy of their predictions. The trust game was additionally included at the end of the task to be used as a manipulation check.

Statistical analysis

The primary goal for Study 5 was to investigate whether asymmetries in learning are driven by inferences about moral character, or by asymmetries in the choices that good and bad agents make. If the observed learning differences are driven by asymmetries in the choices that morally good and bad agents make, then the effects should be restricted to analysis of the morality trials where agents behave differently. However, if the effects are driven by immoral agents, rather than immoral choices, then learning differences should span across morality and competence trials. Consequently, we performed all analyses separately for morality trials and competence trials. We used two-tailed signed tests to confirm group mean parameter estimates differed significantly between good and bad agents on morality trials (replicating findings from Studies 1-4). A similar analysis restricted to competence trials allowed us to investigate whether we could independently

manipulate how people form impressions about an agents' competence as a function of their moral character.

3.6.2 RESULTS

First, we investigated whether participants indeed learned through trial-and-error about the agents' moral preferences and skill level in the task. We analyzed the model's final estimates about each agent's κ and τ ($\hat{\mu}_1^{50}$), and verified that participants formed beliefs that closely resembled the agent's true κ and τ (mean±SD bad morality: 0.290±0.028; good morality: 0.707 ± 0.003; bad competence: 0.500 ± 0.061; good competence: 0.499 ± 0.061; **Appendix A:**). Specifically, participants inferred that the bad agent required less money to increase shocks to the victim than the good agent ($Z = -11.922$, $p < 0.001$), but both agents would spend similar amounts of time to score additional points in basketball ($Z = 0.011$, $p = 0.991$). Participants' beliefs about the agents' character also affected their social behaviour, as they entrusted the good agent with more money than the bad agent in the trust game (bad: 2.70±0.241; good: 7.90±0.234; $Z = -10.112$, $p < 0.001$; **Table 3.4**).

Subjective character ratings confirmed that the bad agent was characterized as nastier than the good agent (mean±SD; bad: 0.346 ± 0.202; good: 0.741 ± 0.173; $Z = -11.755$, $p < 0.001$; **Appendix D:**). However, despite the agents' similar basketball performance, participants rated the bad agent as less skilled than the good agent (bad: 0.426 ± 0.214; good: 0.510 ± 0.237; $Z = -3.061$, $p = 0.002$). This is consistent with previous work on the halo effect, where impressions created about one trait spread to other traits (Dion, Berscheid, & Walster, 1972). Study 4 found that the computational mechanisms of competence inference was not influenced by impressions of basketball skill. Given that

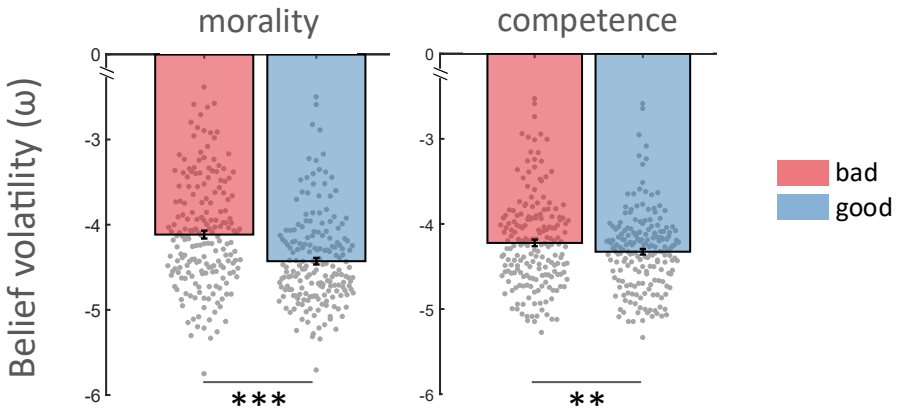
Study 4 demonstrated that people do not learn differently about agents who drastically differ in their basketball competence, the observed difference in competence ratings here was not a concern for subsequent analyses. Thus, we can be confident that any observed between-agent differences in competence inference are unlikely to be attributed to asymmetries in impressions of good and bad agents' basketball ability.

Again, participants were more uncertain about their characterization of the bad agent's morality, as demonstrated by their subjective uncertainty ratings (bad: 27.880±1.019; good: 24.209±1.027; $Z = 4.127$, $p < 0.001$; **Appendix E:**). Additionally, beliefs about the bad agent's morality were more volatile than beliefs about the good agent's morality, as demonstrated by a higher ω for the bad agent (bad: -4.116±0.046; good: -4.428±0.039; $Z = 5.079$, $p < 0.001$; **Appendix B:**).

Confirming our hypothesis, participants expressed greater uncertainty in their impression of the bad agent's basketball skill (bad: 28.875 ± 0.955; good: 27.277 ± 0.992; $Z = 2.323$, $p = 0.020$; **Appendix E:**). Thus, it is not surprising that participants also formed more volatile beliefs about the bad agent's competence (bad: -4.224 ± 0.039; good: -4.327 ± 0.034; $Z = 3.030$, $p = 0.002$; **Appendix B:**), relative to the good agent, as indicated by a higher ω (**Figure 3.10**). These findings suggest that our observation of more uncertain and volatile beliefs about the bad agent cannot be attributed to asymmetries in the choices that good and bad agents make.

Figure 3.10 Moral character information shapes competence inference

*Comparison of volatility of beliefs about the good and bad agent's morality (left) and competence (right) in Study 5, N=189. Error bars represent SEM. **P < 0.01, ***P < 0.001*



3.7 STUDY 6: REVISING IMPRESSIONS WHEN MORAL PREFERENCES CHANGE

Studies 1 through 5 show that beliefs about the morality of bad agents are more uncertain (and thus more volatile) than beliefs about the morality of good agents. Such results suggest that bad impressions are more rapidly updated than good impressions in the face of new, and potentially inconsistent, evidence. We hypothesized that this may reflect a mechanism by which people could revise their impressions of those who they infer threat by promoting cognitive flexibility in the service of cooperative but cautious behaviour. Here, we test this prediction directly using an adapted version of the moral inference task. Participants were randomized to learn about an agent who was initially either bad or good, but then began to make choices that were consistently either more or less moral than previously. Because beliefs about bad agents are more volatile, we predicted that participants would more strongly update their impressions of bad agents than good agents. We tested our hypothesis by comparing, for bad vs. good agents, the extent to which

participants updated their impressions, defined as the difference between character ratings before vs. after the agents' preferences shifted. Because this study investigated how people update character impressions in response to contradictory information, the design most closely resembled those implemented in past social psychology studies (Mende-Siedlecki et al., 2013; Reeder & Coovert, 1986).

3.7.1 METHODS

Participants

Four-hundred and eight U.S. residents were recruited from AMT and randomized to learn about an agent who was initially either bad or good, but then began to make choices that were consistently either more or less moral than previously. Participants provided informed consent and were compensated for their time. The study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098). Forty-four participants were excluded from the analysis as their behavioural performance was below chance (<50% accuracy). Final analysis was carried out on the remaining 364 participants. We confirm the pattern of results is similar when we include all participants in **Appendix A-E**.

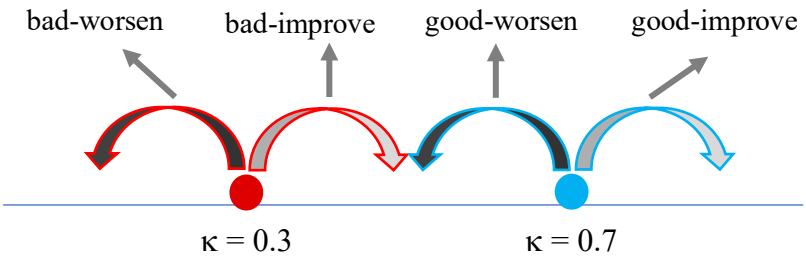
An *a priori* power analysis indicated that the study required 360 participants to have 80 percent power to detect a small to medium interaction effect ($f = 0.175$) in an analysis of variance (ANOVA). Thus, our study was sufficiently powered to observe an effect in our between-groups design. We pre-registered our sample size, experimental design, and planned analyses on the Open Science Framework (https://osf.io/5s23d/).

Experimental Procedure

Participants completed a modified version of the Moral Inference Task. In the task, participants predicted a sequence of 36 choices made by a single agent (using the experimental design implemented in Studies 1-3, **Figure 2.1**), and on each trial received immediate feedback about their accuracy. Every few trials, participants rated their impression of the agent's moral character and how certain they were about their impression. The study comprised a 2x2 factorial design with agent moral character (bad versus good) and shift direction (improve versus worsen) as between-subject independent variables (**Figure 3.11**).

Figure 3.11 Impression updating experimental design

Participants were randomized to learn about a bad agent (κ=0.3) or a good agent (κ=0.7) whose moral character either improved (κ+0.2) or worsened (κ–0.2).



Moral character: Between subjects we manipulated the moral character of the agent that participants observed (bad versus good). To manipulate moral character, we created agents with different preferences towards harming the victim, similar to our previous studies (bad agent: $\kappa = 0.3$; good agent: $\kappa = 0.7$). For the first 30 trials (phase 1) participants observed the two agents make choices for identical trial sequences. On every trial, the agents faced the same two options, but because the agents had different preferences

towards harming the victim, they often chose differently. We created the sequence of 30 trials using similar methods to those reported in **Chapter 2.1**, and simulated how the agent chose using **Equation 2.1**-**Equation 2.3**. In phase 1, we asked participants to provide subjective character and uncertainty ratings every 1-3 trials, for a total of 15 ratings.

Shift direction: Because we were interested in how participants update their impressions when an agent's behaviour becomes inconsistent with prior evidence, we manipulated the agents' preferences on the final 6 trials of the Moral Inference Task. For half of the participants, the agent became more moral than previously observed in the first 30 trials (improve condition) and for the other half the agent became less moral than previously observed in the first 30 trials (worsen condition). In the improve condition, agents became more harm-averse, and therefore required more money to inflict pain than previously (κ+0.2). In the worsen condition, agents became less harm-averse, and therefore required less money to inflict pain than previously (κ-0.2).

For the final 6 trials (phase 2), participants observed the agents make choices that were inconsistent with their previous preferences. Thus, in the improve condition, agents made prosocial choices where they would have previously chosen antisocially. In the worsen condition, agents made antisocial choices where they would have previously chosen prosocially. Together, this resulted in four conditions, manipulated between subjects: (1) bad agent's morality improves (bad-improve, $\kappa = 0.3 \rightarrow \kappa = 0.5$), (2) bad agent's morality worsens (bad-worsen, $\kappa = 0.3 \rightarrow \kappa = 0.1$), (3) good agent's morality improves (good-improve, $\kappa = 0.7 \rightarrow \kappa = 0.9$), and (4) good agent's morality worsens (good-worsen, $\kappa = 0.7 \rightarrow \kappa = 0.5$). In phase 2, we asked participants to provide subjective character and uncertainty ratings every second trial, for a total of 3 ratings.

In order to minimize the potential influence of prior expectations on participant predictions, we anchored prior expectations through explicit instruction. Specifically, we told participants that on average, people required $1 per shock to the victim. This prior expectation maps on to $\kappa = 0.5$ (i.e., equidistant from the initial preferences of the good and bad agents).

Statistical Analyses

The primary goal of Study 6 was to investigate whether participants more rapidly update their impressions of bad agents than good agents, particularly when agents show moral improvement. Consequently, we computed the magnitude that participants' impressions updated from phase 1 to phase 2. The update was defined as the difference between participants' phase 2 and phase 1 ratings (update = phase 2 – phase 1). For phase 1 ratings we took the average of the final 3 ratings in phase 1, and for phase 2 ratings we took the average of the 3 ratings in phase 2. We preregistered this definition of the update prior to collecting data (https://osf.io/5s23d//). We conducted a 2 (agent: bad versus good) x 2 (shift direction: improve versus worsen) analysis of variance (ANOVA) to obtain main effects and interaction effects. Because our dependent measure was not normally distributed, we split the data to complement the ANOVA with non-parametric statistics.

3.7.2 RESULTS

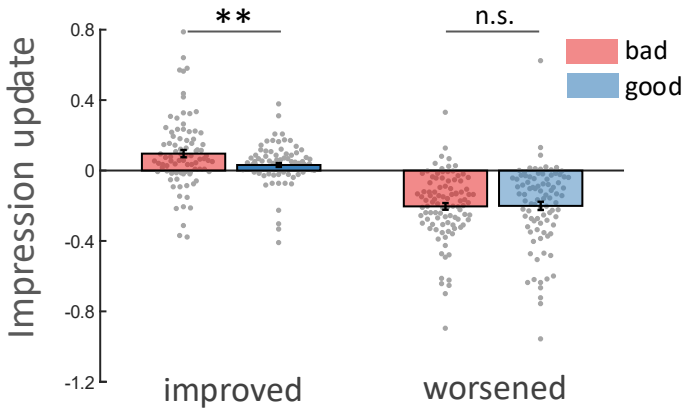
First, we investigated whether our results from phase 1 of the task replicated our previous findings from Studies 1-5. Again, participants indicated greater subjective uncertainty in their impression of the bad, relative to the good, agent (bad: 33.584±1.164; good: 27.945±1.347; $Z = 4.362$, $p < 0.001$; **Appendix E:**). This translated into faster

updating for the bad agent, as demonstrated by a larger ω (bad: -3.559±0.042; good: -3.928±0.034; $Z = 6.577$, $p < 0.001$; **Appendix B:**).

Next, we investigated impression updates following the agents' shift in behaviour. As predicted, we observed a main effect of agent on impression updating, where participants updated their character ratings more for bad agents than good agents (bad: 18.951±1.245; good: 14.928±1.316; F(1,360)=5.124, *P*=0.024; Z=3.541, *P*<0.001, **Figure 3.12** and **Figure 3.13a**). There was also a main effect of shift direction: updating was greater when morality worsened than when it improved (worsen: 22.083 ±1.389; improve: 11.468 ±1.010; F(1,360)=37.698, *P*<0.001; Z=6.372, *P*<0.001). This is consistent with past reports of the negativity bias in impression formation (Fiske, 1980; Pratto & John, 1991; Skowronski & Carlston, 1989), where people show stronger impression updating in response to inconsistent immoral behaviours relative to moral behaviours. Main effects were qualified by an interaction between agent and shift direction (F(1,360)=6.803, *P*=0.009; Chi-squared=57.227, *P*<0.001), where asymmetric updating was more pronounced when morality improved than when morality worsened (**Figure 3.12**). At first glance, this interaction may appear surprising, because our model only predicts a main effect of agent and does not differentiate between positive and negative updating. However, our theoretical framework proposes that people form more volatile beliefs about putatively bad agents due to an adaptive mechanism whereby potentially threatening cues increase attention and learning. Thus, when a "good" agent's behaviour suddenly worsens, participants may infer a potential threat, prompting their beliefs about the agent to become more uncertain and amenable to rapid updating.

Figure 3.12 Impression updating from inconsistent behaviour

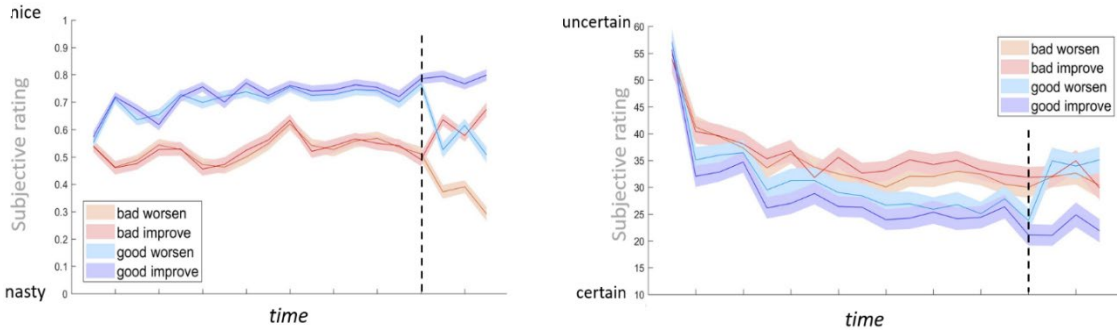
*In Study 6, participants more strongly updated their impressions of bad than good agents when moral character improved but not when it worsened. Error bars represent SEM. **P < 0.01*



Following from this prediction, in a secondary analysis we compared uncertainty before versus after the shift in our 2x2 factorial design. Post-change, for bad agents, uncertainty remained high, regardless of whether the agent's morality improved or worsened (**Figure 3.13b**; improved: Z=0.040, *P*=0.968; worsened: Z=1.233, *P*=0.218). However, for good agents, uncertainty increased when morality worsened (Z=-5.507, *P*<0.001), and decreased when morality improved (Z=2.252, P=0.024). The more uncertain participants became about the good agent whose morality worsened after the shift, relative to before, the more they updated their impression about that agent (Spearman's ρ =0.420, P<0.001).

Figure 3.13 Graphical depiction of temporal evolution of subjective ratings

*Temporal evolution of subjective character ratings (**a**) and uncertainty ratings (**b**) from Study 6. Dashed line represents the point at which the agent's behaviour worsened or improved. In phase 1 (to the right of the dotted line) ratings were made every 1-3 trials, for a total of 15 ratings. In phase 2 (to the left of the dotted line) ratings were made every second trial, for a total of 3 ratings.*



3.8 DISCUSSION

Using a novel Moral Inference Task, we have shown that moral learning is explained by an asymmetric Bayesian updating mechanism where beliefs about the morality of bad agents were more uncertain and volatile than beliefs about the morality of good agents. Notably, asymmetries in learning between good and bad agents were robust to a setting where agents' choices are noisy rather than deterministic. Asymmetries in learning about bad compared to good agents extended to learning about a trait unrelated to morality; participants' beliefs about bad agents were more uncertain and volatile than beliefs about good agents, but there was no difference in the volatility of beliefs about agents who differed in basketball ability. Furthermore, inferring bad character destabilized overall social impression formation, spilling over into learning about a non-moral trait. When moral behaviour improved, impressions were updated faster for putatively bad agents than good agents. Thus, the volatility of bad moral impressions may facilitate forgiveness by enabling initially bad impressions to be rapidly updated if behaviour improves.

By simultaneously measuring implicit beliefs about moral preferences that guided behavioural predictions, as well as explicit subjective impressions of moral character, our paradigm revealed that beliefs about preferences and subjective character impressions followed different dynamics. Consistent with previous work (Todorov et al., 2009), participants rapidly formed subjective impressions about moral character after just a few trials (**Figure 3.3**). Meanwhile, beliefs in the model integrated over more information and updated gradually over a longer timescale (**Figure 3.2**), reflecting the fact that the model estimates the precise exchange rate between money and pain, which cannot be inferred from a single trial. These different dynamics highlight how subjective moral impressions are often based on highly impoverished information: in our studies, participants were readily willing to judge the character of others well before they formed precise beliefs about their moral preferences. Why and how people jump to conclusions about others' character despite lacking sufficient information to accurately predict their behaviour remains an important question for further study.

Although theoretical models of person perception have claimed the independence of trait dimensions (namely warmth and competence) (Fiske et al., 2007), other evidence suggests that judgments across trait dimensions may share a positive relationship (Judd et al., 2005; Rosenberg et al., 1968). Our work lends further support to the possibility that the cognitive processing of different traits belonging to the same individual are related, and offers tools for addressing this question. By considering uncertainty of beliefs in addition to valence, future work may shed new light on how the mechanisms supporting different dimensions of person perception relate to one another.

Overall, our findings are consistent with research identifying a negativity bias in impression formation, where bad behaviours command more attention than good behaviours (Baumeister et al., 2001; Fiske, 1980; Pratto & John, 1991; Skowronski & Carlston, 1989), and research showing that uncertain attitudes are susceptible to change (Tormala & Rucker, 2007). Taken together, our results extend this literature to show that when considered within a Bayesian learning framework, a negativity bias naturally makes impressions more volatile, where impressions about bad agents are more rapidly updated than impressions about good agents. We suggest that by destabilizing overall impressions of others, the learning mechanism described here promotes cognitive flexibility in the service of building richer models of potentially threatening others. This mechanism provides an algorithmic solution to the problem of moral inference in a world where people sometimes make mistakes, and helps resolve the paradox of how people can forgive despite the potency of negative information in judging the moral character of others.

Chapter 4

4 OPTIMISTIC PRIOR EXPECTATIONS AND MORAL INFERENCE

This chapter incorporates sections from the Supplementary Materials of a paper published as:

Siegel, J.Z., Mathys, C., Rutledge, R.B., & Crockett, M. J. Beliefs about bad people are volatile. *Nature Human Behaviour*, 2.10 (2018): 750.

Doi: https://doi.org/10.1038/s41562-018-0425-1

Chapter 3 revealed that the computational mechanisms supporting moral inference are intrinsically related to the moral character of the agent: beliefs about putatively bad agents are especially uncertain and amenable to updating. A possible explanation for why people form more uncertain beliefs about the moral character of bad than good agents is a strong prior expectation that people will behave morally (Brañas-Garza et al., 2017; Rand, 2016), thus rendering the bad agent's behaviour more surprising. In this chapter, I

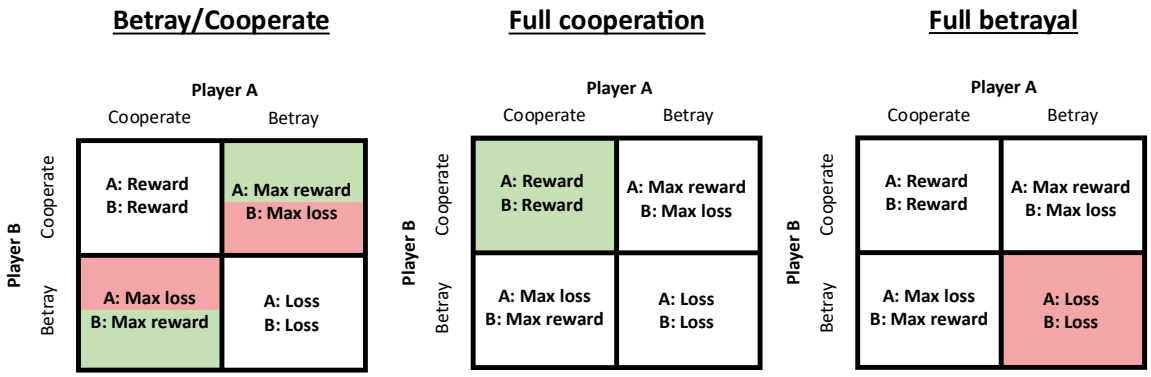
compliment new analysis using the data from Chapter 3, with two independent studies to investigate this plausible explanation (hereon known as the 'optimistic prior' hypothesis). The first study examines the level of harm aversion that people expect others to have within the experimental setting of the Moral Inference Task, quantitatively. The other study independently manipulates prior expectations about harm to investigate the effects of prior moral expectations on the underlying mechanisms of moral inference.

4.1 INTRODUCTION

The Prisoner's Dilemma is a behavioural economic game where two players with no means of communicating with each other face a dilemma. Each player is independently presented with an offer to either cooperate with or betray the other player. Betrayal can be beneficial because it earns the betraying individual a large reward, so long as the other player chooses to cooperate - in which case, the cooperator is maximally punished (betray/cooperate, **Figure 4.1**). Cooperation from both players earns each player some reward (full cooperation), though not as large as would be received in the former case. However, if both players decide to betray the other, both players are punished (full betrayal), though not to the same extent as the betray/cooperate case. Because neither player knows how the other will decide, in order to maximize rewards and minimize punishment they must predict how they *expect* the other will choose.

Figure 4.1 The Prisoner's Dilemma

Without any means of communication, two players decide whether to betray or cooperate with the other. If they chose differently, the betrayer receives the maximum reward and the cooperator receives the maximum loss (betray/cooperate). If both cooperate, they each receive some reward (full cooperation); if both betray, they each receive some loss (full betrayal)



When the game is played as a one-time encounter with a stranger, an expectation of betrayal promotes further betrayal in order to avoid maximal losses. Conversely, an expectation of cooperation may encourage a player to betray (for self-interested reasons) or cooperate (for prosocial reasons). Therefore, there is only one rational response when a player believes the other has ill intentions: betrayal. Yet a substantial amount of cooperative behaviour is observed in one-shot Prisoner's Dilemma experiments in the laboratory (Rand et al., 2014; Wong & Hong, 2005), suggesting that even in one-time encounters people often expect strangers to cooperate. Indeed, across a wide range of behavioural economic games, there is evidence that humans expect others to make prosocial decisions, and this expectation encourages people to make prosocial decisions themselves (Brañas-Garza et al., 2017; Yamagishi et al., 2013). Because prior expectations are fundamental to Bayesian belief updating, in this chapter I explore the plausible

hypothesis that optimistic expectations drive asymmetric belief updating and uncertainty between good and bad agents.

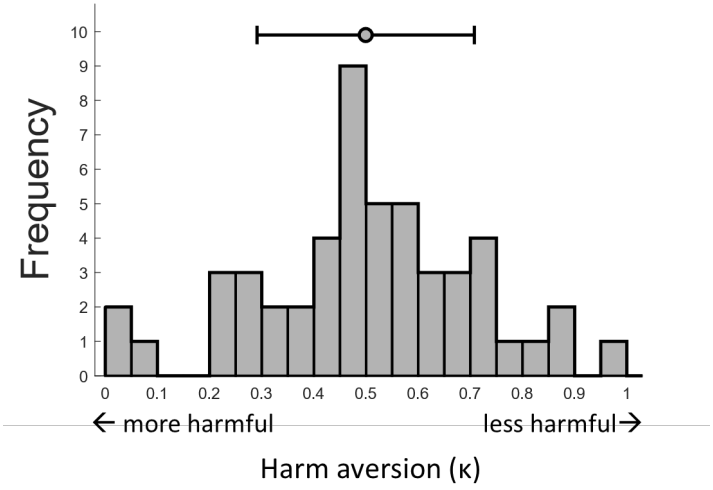
Despite clear evidence that people often have optimistic prior expectations about others' morality, these expectations are undeniably tied to the subjectivity of what is considered "good". For the Moral Inference Task, we did not probe inferences about prototypically "bad" behaviours such as lying, cheating and stealing because we expected people to have more objective prior expectations about behaviours with well-established norms. Instead, we chose to probe inferences on moral decisions that people are unlikely to have encountered before (trading money for electric shocks), and thus are unlikely to have strong prior expectations about. Because there is no established 'morally good' exchange rate between money and electric shocks, one observer may judge the choice to deliver 1 shock to an anonymous stranger in exchange for 1 pound as highly blameworthy, while another may judge that same choice undeserving of any blame. This raises the question whether participants have 'optimistic' prior expectations about decisions to trade money for electric shocks?

Studies report that, absent of additional information, people expect other's moral behaviour to be similar to their own (Brañas-Garza et al., 2017; Yamagishi et al., 2013). When human participants (N = 51) are asked to make real decisions between trading profit for themselves and electric shocks for an anonymous stranger, the distribution of harm aversion (κ) is roughly symmetrical around 0.5 (mean = 0.499, standard deviation (s.d.) = 0.210; **Figure 4.2**). Consequently, to simulate agents who were equally more harmful (bad agent) and less harmful (good agent) relative to typical human behaviour, we simulated agents to have preferences that were symmetrical around the average harm aversion when

real shocks and money are at stake (bad = 0.3, good = 0.7). Regardless, there remains a possibility that people expect prosocial behaviour within the context of the Moral Inference Task. A strong prior expectation for prosocial behaviour would render the bad agent's behaviour more surprising, as demonstrated by more uncertain, volatile beliefs. That is, if people expect the agents to have preferences that more closely resemble the preferences of the good agent, there would be less need to update those beliefs because the good agent's behaviour is more consistent with the prior belief (i.e., less surprising, smaller prediction errors). Thus, more uncertain, flexible beliefs about bad agents may be driven by an "optimistic" bias about other's harm aversion.

Figure 4.2 Distribution of harm aversion in a pilot sample of participants

The distribution of harm aversion when participants make real decisions between trading profit for themselves and electric shocks for a stranger. Error bars represent standard deviation.



In the following sections, I investigate the role of prior moral expectations in asymmetric uncertainty and belief updating between good and bad agents. First, I report additional analyses using the data from Chapter 3 that do not support the optimistic prior

hypothesis. Next, I introduce an independent study to directly investigate how people expect others to behave within the experimental setting of the Moral Inference Task. Finally, using an adapted version of the Moral Inference Task, I independently manipulate prior expectations of harm using facial stimuli and shed new light on the mechanisms that support flexible, uncertain beliefs about putatively 'bad' agents.

4.2 EVIDENCE AGAINST THE PRIORS HYPOTHESIS

Chapter 3 reported the findings from seven studies with data from over 1500 participants. Here, I report additional exploratory analyses using this data to elucidate the relationship between prior expectations and behaviour in the Moral Inference Task.

4.2.1 RELATIONSHIP BETWEEN SUBJECTIVE PRIOR EXPECTATIONS AND BEHAVIOUR

In Chapter 3, Studies 2-6, we asked participants to indicate how nasty or nice they *expect* agents will be, prior to observing either of the agents' choices. If prior expectations that agents will be 'good' increase uncertainty and volatility in beliefs about the bad agent, relative to good, then we would expect to see greater between-agent differences the nicer participants think agents will be. To investigate, we computed for each participant the difference in belief volatility, ω, between the good and bad agent ($\Delta\omega = \omega_{bad} - \omega_{good}$) and checked for correlations with subjective prior ratings. This analysis was conducted specifically for studies that include a within-subject manipulation of moral character (i.e., bad morality vs. good morality).

Across studies we found no consistent relationship between prior ratings and either of these dependent measures (see **Table 4.1** and **Figure 4.3**). No study showed a significant

positive relationship between prior ratings and Δω, as would be predicted by the optimistic prior' hypothesis. Nonetheless, to investigate the possibility that a sub-threshold relationship really does exist, we conducted a mini meta-analysis on the correlations across all studies that include a within-subject manipulation of moral character. Again, we found no evidence for a relationship between subjective prior expectations and the difference in belief volatility between agents (Z = -0.846, p = 0.398).

Table 4.1

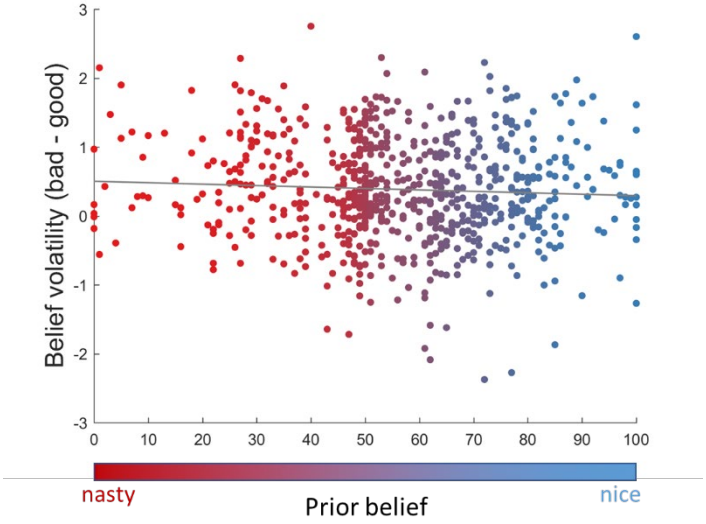
Correlation between prior trait rating and the difference in belief volatility (Δω) between good and bad agents.

		Prior trait rating* (mean ± SEM)	**Δω correlation ρ (p-value)**
Study 1		n.a	n.a
Study 2		48.724 ± 1.610	-0.179 (0.022)
Study 2s		60.621 ± 1.750	0.006 (0.952)
Study 3		52.007 ± 1.934	-0.029 (0.738)
Study 4	morality	56.954 ± 1.841	0.175 (0.068)
	competence	54.955 ± 1.404	0.106 (0.268)
Study 5	morality	61.355 ± 1.341	0.020 (0.788)
	competence	56.735 ± 1.192	0.033 (0.610)

*prior trait ratings are collected prior to observing any outcomes. For the morality task, participants are asked to indicate how nasty or nice they expect the agent to be on a scale ranging from 0 = *nasty* to 100 = *nice* (0 = *bad* to 100 = *good* for study 2s). For the competence task, participants are asked to indicate how skilled they expect the agent to be in basketball on a scale ranging from 0 = *beginner* to 100 = *expert*. Study 6 was omitted from this analysis because this was a between-subjects, rather than within-subjects, design. Thus, we cannot compute Δω for Study 6.

Figure 4.3 Prior expectations do not covary with asymmetric updating

Relationship between prior beliefs and the difference in volatility estimates between agents (Δω) across Studies 2, 2s, 3, 4 (morality condition), 5. Colors represent the valence magnitude of the prior belief; stronger red indicates worse expectations and stronger blue indicates more favourable expectations.



4.2.2 RELATIONSHIP BETWEEN MORAL PREFERENCES AND BEHAVIOUR

Previous work suggests that absent of additional information, people expect others preferences to be similar to their own preferences (Brañas-Garza et al., 2017; Hsee & Weber, 1997; Yamagishi et al., 2013). Consequently, it is possible that participants expect others to perform similarly to how they choose in the task. Prior to observing the agents' choices in Studies 2 and 3, participants indicated how *they* would decide if they were faced with similar decisions to profit from harming an anonymous person. Specifically, participants made a series of 20 hypothetical decisions that involved choosing between less money for themselves plus less shocks for anonymous person, or more money for themselves at the expense of more shocks for that person. We then adapted the decision

model in Equation 2.1 - Equation 2.3 to estimate participant's own harm aversion parameter ($\kappa_{subject}$).

Equation 4.1

$$V_{harm} = (1 - \kappa_{subject})\Delta m - \kappa_{subject}\Delta s$$

If people expect others to have similar preferences to their own, then we would expect participants who are more harm averse (i.e., larger values of $\kappa_{subject}$) to show greater between-agent differences in our task, under the optimistic prior hypothesis. Participants had an average $\kappa_{subject}$ = 0.445±0.022 in Study 2 and an average $\kappa_{subject}$ = 0.443±0.026 in Study 3. Given that people's own behaviour more closely resembles that of the bad agent than the good agent, it is unlikely that the bad agent's behaviour is more surprising than the good agent's. In a correlational analysis, we find the opposite relationship between participant's own preferences and $\Delta\omega$ than would be predicted by the optimistic prior hypothesis. Greater harm aversion was associated with smaller between-agent asymmetries in Bayesian belief updating (Study 2: ρ = -0.355, p<.001; Study 3: ρ = -0.154, p=.076).

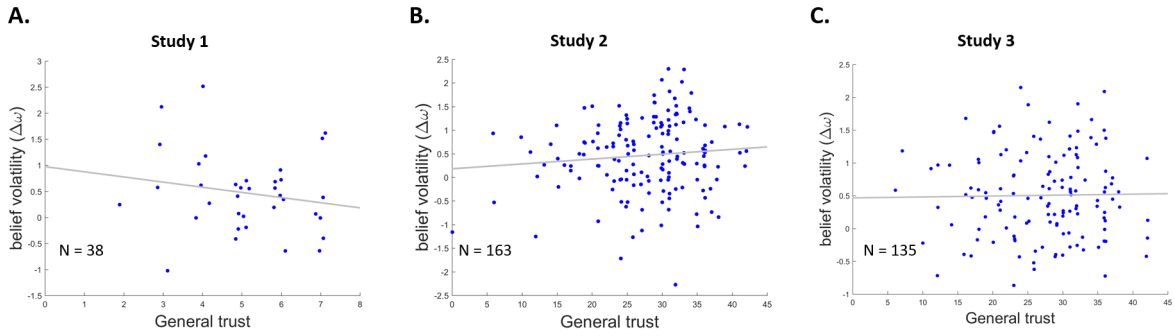
4.2.3 RELATIONSHIP BETWEEN GENERALIZED TRUST AND BEHAVIOUR

A prior expectation that people are generally morally good is likely related to beliefs about other's trustworthiness: the greater the expectation that people will be good, the more likely you are to believe others are trustworthy. In Study 1 we asked participants in a pre-testing questionnaire "To what extent do you feel you can trust other people that you interact with in your daily life?". Participants responded on a scale ranging from 1 (*very*

little) to 7 (*very much*). The optimistic prior hypothesis predicts that the more people generally believe that others are trustworthy, the more volatile beliefs will be for the bad agent relative to the good agent (larger Δω). In fact, we found we found no relationship between general trust and Δω (ρ = -0.178, p = 0.300; **Figure 4.4a**).

Figure 4.4 General trust and asymmetric Bayesian updating

Greater asymmetric belief updating (Δω) was not associated with general trust for Study 1 (A) ρ = -0.178, p = 0.300, study 2 (B) ρ = 0.065, p = 0.413 or study 3 (C) ρ = 0.020, p = 0.817.



In studies 2 and 3 participants completed a generalized trust scale, consisting of 6 items related to general beliefs about the trustworthiness and kindness of others (Yamagishi & Yamagishi, 1994). For example, "Most people are basically good and kind" and "Most people are trustworthy". Items were rated on a scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*) and summed for a single measure of 'generalized trust'. Again, we found no significant relationship between general trust and Δω in Study 2 or 3 (Study 2: ρ = 0.065, p = 0.413; Study 3: ρ = 0.020, p = 0.817; **Figure 4.4b-c**)

4.2.4 PRIOR EXPECTATIONS ABOUT BASKETBALL COMPETENCE VERSUS MORALITY

Another objection to the optimistic prior hypothesis is that prior expectations about morality and basketball competence were very similar in Chapter 3, Study 4 (morality = 56.339±1.845; competence = 54.580; Z = 0.678, p = 0.498), yet between-agent differences in belief volatility were restricted to the morality condition. However, prior expectations about skill may have been weaker (i.e., more uncertain) than those about morality. Consequently, we cannot rule out the possibility that prior expectations influence behaviour in the morality conditions but not the competence condition. To test this possibility, we checked whether participants expressed greater uncertainty in their explicitly stated prior expectations about an agent's basketball skill relative to moral character. We performed this analysis first for our between-subject design, Study 4, where participants either indicated their certainty about how skilled or how moral they expect an agent would be. We performed a similar analysis for our within-subject design, Study 5, where participants indicated their certainty in their expectations about an agent's morality and skill. In both studies, we found no significant differences in how certain participants were in their prior expectations about an agent's basketball skill and morality (Study 4: morality = 65.092 ± 2.355 , competence = 69.089 ± 2.595, Z = -1.536, p = 0.125; Study 5: morality = 56.079 ± 1.845 , competence = 57.217 ± 1.878, Z = -1.246, p = 0.213).

4.2.5 OTHER CONSIDERATIONS

Two other pieces of evidence from our previous experiments argue against an optimistic prior hypothesis. First, a prior expectation that people will behave morally cannot explain why inferring a bad moral character destabilizes beliefs about basketball

competence in Study 5. Second, in Study 6 we anchored participants to expect that most people require $1 per shock to the anonymous person, consistent with a prior belief of $\kappa = 0.5$. Yet we still find that beliefs about bad agents were more uncertain and volatile than beliefs about good agents. Together, this evidence does not support the hypothesis that asymmetries in learning result from prior expectations that people will be moral.

4.3 INVESTIGATING PRIOR EXPECTATIONS IN THE MORAL INFERENCE TASK

In general, it may be the case that people expect others to behave morally, or at the very least, not behave immorally. We don't expect that the majority of people that walk into a store are going to shoplift, and we might expect a majority of people who see someone suffering to try and help. The question we address here is whether *within the context of the Moral Inference Task*, people expect others to behave in a manner that more closely resembles the behaviour of the good agent than the bad agent. To investigate this possibility, we recruited participants in an independent study to determine how people expect others to behave when faced with the same moral decisions our agents faced in our previous experiments.

4.3.1 METHODS

Participants

Thirty U.S. residents were recruited from AMT to participate in a prediction task. All participants provided informed consent and were compensated for their time. This study was approved by the Medical Sciences Interdivisional Research Ethics Committee, University of Oxford (MSD-IDREC-C1-2015-098).

Experimental Procedure

Participants were fully briefed about our previous experiments where two participants arrive at the laboratory, and one of them makes *real* (i.e., not hypothetical) decisions about whether to profit by inflicting shocks on the other. After observing an example trial, participants were asked to indicate how they think "most people" decided in our previous experiments. Specifically, we asked them to predict which option was most commonly chosen by our participants, for a set of 34 trials. This allowed us to estimate participants' expected level of harm aversion (κ) within the context of our task. Feedback was not provided throughout the task. Crucially, we incentivized participants to be as accurate as possible in their predictions, because they would be rewarded financially for every choice for which they successfully predicted the majority response.

We modelled participants' predictions using the same decision model that was used to simulate agent choices (**Equation 2.1**-**Equation 2.3**), and extracted how harm averse they expected most people would be in this task, κ_e.

Equation 4.2

$$V_{harm} = (1 - \kappa_e)\Delta m - \kappa_e \Delta s$$

4.3.2 RESULTS

The optimistic prior hypothesis predicts that people will expect others' harm aversion, parametrized as κ_e, to be significantly greater than 0.5. This would demonstrate that people expect others to behave more similarly to the good agent than the bad agent, rendering the bad agent's choices in our task more surprising. In fact, our study reveals that participants expect others to behave slightly more similarly to the bad agent (κ_e =

0.445±0.043) though a one-sample Wilcoxon signed-rank test revealed that this was not significantly different from $\kappa = 0.5$ ($Z = -1.347$, $p = 0.178$). This study provides vital evidence that, at least within the context of our task, participants do not expect others to behave more similarly to the good agent than the bad agent.

4.4 MANIPULATING PRIOR EXPECTATIONS IN THE MORAL INFERENCE TASK

We found no evidence to suggest that people expect agents in the Moral Inference Task would decide in a manner more closely resembling the behaviour of the 'good' agent. Nor did we observe a relationship between the observed asymmetry in belief updating and measures that one would expect to be related to prior expectations of morality (e.g., generalized trust in others, personal moral preferences, subjective expectations of moral character). This raises the question: what role, if any, do prior expectations play in moral inference?

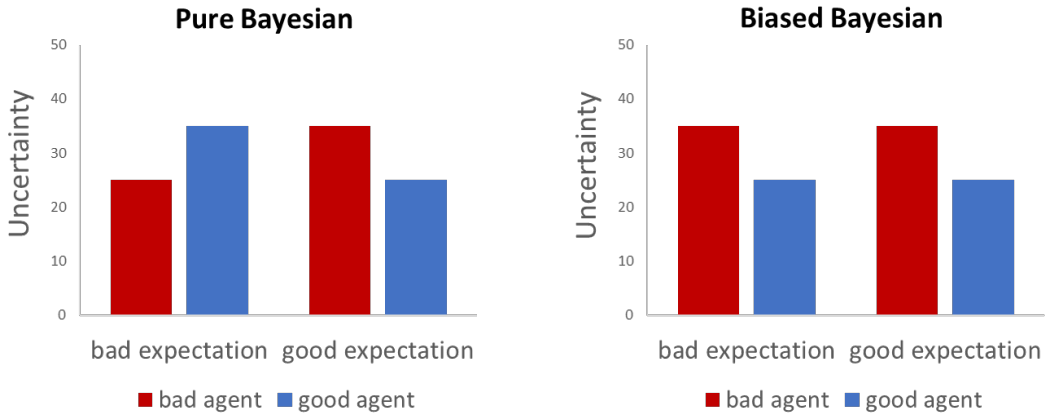
One possibility is that moral attributes bias learning in systematic ways. For instance, imagine you are stuck in an elevator with three people: a middle-aged business woman, a tall slender young man with a rucksack, and a large muscular man with ripped clothing and several tattoos on his face. Now, try to form an impression about each of their moral characters; no doubt you will have different expectations about each. Being stuck in the elevator, you start seeing more and more evidence corresponding to each of your expectations to the point where you feel you have a very accurate representation of their preferences. Despite equal information about each individual, chances are you might want to continue to keep an eye on the tattoo-bearing muscle man. Indeed, previous research suggests that people preferentially attend to individuals who they expect to be less

trustworthy and who may pose a threat to one's survival (Callan et al., 2013; Hackel, Looser, & Van Bavel, 2014; Pratto & John, 1991). This suggests that belief updating may be biased towards learning about bad agents because people are inherently motivated to learn about bad characters, not because their behaviour is more surprising.

We evaluate moral character from multiple cues, including physical appearances as well as behaviours. People form moral impressions after merely 40 milliseconds exposure to a face (Todorov et al., 2009) and these impressions influence subsequent decision-making and judgements of others actions; people judge harmful and helpful actions paired with an untrustworthy face more harshly than those paired with a trustworthy face (Baron, Gobbini, Engell, & Todorov, 2011). This suggests facial information sets moral expectations that biases subsequent learning from behaviours, but it is not yet known how the computations unfold. To investigate, the Moral Inference Task was adapted to probe how moral expectations derived from facial information shape learning about others whose behaviour is consistent versus inconsistent with prior expectations. From a purely Bayesian standpoint, beliefs about others should become more stable when behaviour is consistent with prior expectations and less stable when behaviour is inconsistent with prior expectations. Thus, an expectation of harm should decrease the volatility and uncertainty of beliefs about the bad agent, and increase the volatility and uncertainty of beliefs about the good agent ('pure Bayesian' hypothesis in **Figure 4.5**). However, if the valence of the agent's moral character (i.e., whether the agent is good or bad) influences learning and belief updating, beliefs about putatively 'bad' agents should be more volatile and uncertain, regardless of prior expectations ('biased Bayesian' hypothesis, **Figure 4.5**).

Figure 4.5 Prior moral expectations and uncertainty, hypotheses

Predicted uncertainty ratings as a function of prior expectations and agent morality. Manipulating moral expectations shape learning about harmful versus helpful moral agents (pure Bayesian hypothesis). Moral inference is shaped by the valence of the moral character (biased Bayesian hypothesis).



4.4.1 METHODS

Participants

Four-hundred and twenty-three US residents were recruited from AMT and randomized to learn about either a bad or a good agent in the Moral Inference Task, who was represented by either a low or high threat avatar. Participants provided informed consent and were compensated for their time. The protocol was approved by the Yale University Human Research Protection Program Institutional Review Board (IRB protocol ID: 2000022385) and the study complied with all relevant ethical regulations for work with human participants. Seventy-six participants were excluded from the analysis as their behavioural performance was below chance (<50% accuracy). Final analysis was carried out on the remaining 347 participants.

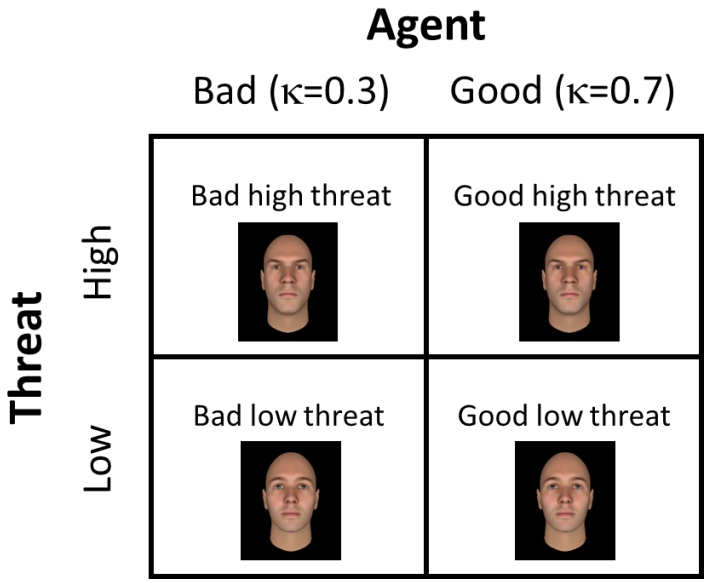
An *a priori* power analysis indicated that the study required 179 participants to have 80 percent power to detect a medium interaction effect (f = 0.25) in an ANOVA. Thus, our study was sufficiently powered to observe a medium sized effect in our between-groups design.

Experimental Procedure

Participants completed a modified version of the Moral Inference Task. In the task, participants predicted a sequence of 50 choices made by a single agent (using the experimental design implemented in Chapter 3, Studies 1-3, **Figure 2.1**), and on each trial received immediate feedback about their accuracy. Every few trials, participants rated their impression of the agent's moral character and how certain they were about their impression. Prior expectations were manipulated using face stimuli developed to vary on their level of perceived threat (Oosterhof & Todorov, 2008). The study comprised a 2x2 factorial design with agent moral character (bad versus good) and threat (high versus low) as between-subject independent variables.

Figure 4.6 Manipulating prior expectations experimental design

Participants were randomized to learn about a bad agent (κ=0.3) or a good agent (κ=0.7) who was represented by either a high threat or a low threat avatar.



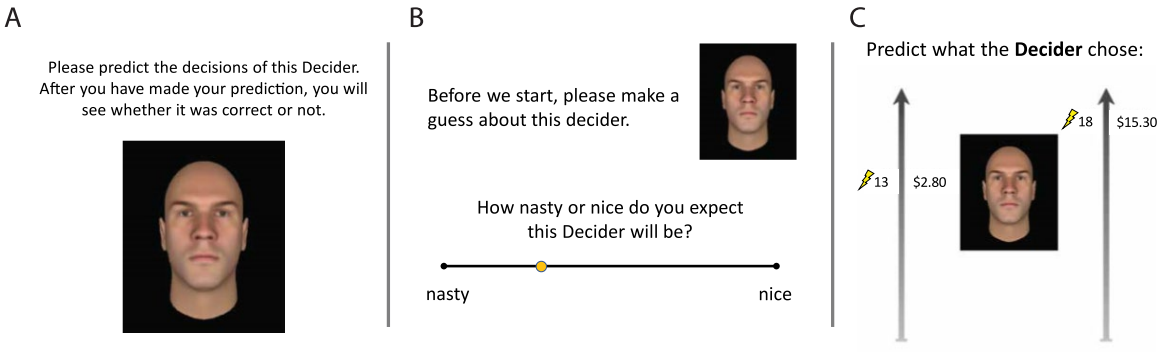
Moral character: Between subjects we manipulated the moral character of the agent that participants observed (bad versus good). To manipulate moral character, we created agents with different preferences towards harming an anonymous other person, similar to our previous studies (bad agent: $\kappa = 0.3$; good agent: $\kappa = 0.7$). We created the sequence of 50 trials using similar methods to those reported in **Chapter 2.1**, and simulated how the agent chose using **Equation 2.1**-**Equation 2.3**.

Threat: Because we were interested in how prior moral expectations shape learning about bad versus good agents, we manipulated how threatening the agents appeared to be before observing any of their choices in the Moral Inference Task. Facial cues indicative of threat are associated with inferences of harmful intent, and crucially, the ability to carry out those intentions (Oosterhof & Todorov, 2008). Consequently, to manipulate

expectations of harm we used validated stimuli developed from a computational model that constructs face variations along the dimension of threat (Oosterhof & Todorov, 2008). Participants were instructed that they would be predicting the decisions of a past participant (known as the "Decider") who will be represented by an avatar. The avatar was presented to the participants prior to observing any of the agent's choices (**Figure 4.7a**), and participants were asked to indicate their expectations about the agent's moral behaviour in the task (**Figure 4.7b**). For half of the participants, the avatar was extremely threatening (high threat condition) and for the other half the avatar was minimally threatening (low threat condition). To ensure the manipulation remained salient throughout the task, the avatar was presented in the middle of the screen for each binary prediction (**Figure 4.7c**) and in the corner of the screen for each rating. At the end of the task we asked participants how threatening they thought the avatar looked on a continuous scale ranging from 0 (*not at all threatening*) to 100 (*very threatening*).

Figure 4.7 Threat manipulation experimental design

*Prior to observing the choices of the agent (known as the 'Decider'), an avatar representing the agent was presented to the participant (**A**) and they were asked to indicate their expectation of the agent's behaviour on a scale from nasty to nice (**B**). The avatar was presented in the middle of the screen on every trial (**C**).*



Prior to observing the avatar, we told participants that on average, people required $1 per shock to the victim. This prior expectation maps on to $\kappa = 0.5$ (i.e., equidistant from the initial preferences of the good and bad agents). This provided participants with a baseline for behavioural expectations and thus any deviations in prior expectations can be attributed to our manipulation of threat.

Statistical Analyses

The primary goal of this study was to investigate whether manipulating prior expectations would influence the pattern of uncertainty and belief updating observed in Chapter 3. Specifically, if optimistic moral expectations account for increased uncertainty and volatility of negative moral beliefs, then decreasing moral expectations via threatening facial cues should increase the certainty and stability of negative moral beliefs. That is, the effect observed in Chapter 3 should decrease when moral expectations worsen as indicated by a significant interaction between our two manipulations (agent and threat). Consequently, we conducted a 2 (agent: bad versus good) x 2 (threat: low versus high) univariate ANOVA to obtain main effects and interaction effects. Because our dependent measure was not normally distributed, we split the data to complement the ANOVA with non-parametric statistics.

4.4.2 RESULTS

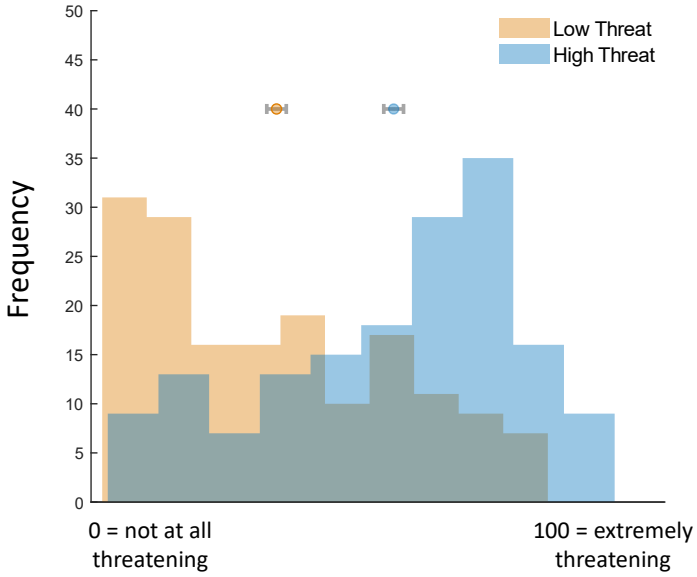
Manipulation checks

First, we verified that participants judged the high threat avatar as more threatening than the low threat avatar (high: 51.323±1.772; low: 30.297±1.749; $F(1,327) = 67.639$, $p<.001$, $\eta^2 = 0.171$; $Z = -7.508$, $p <.001$). However there was substantial variation in how

threatening participants believed the stimuli looked (**Figure 4.8**): 21% of the participants in the low threat condition rated the avatar above the midpoint (>50) of the scale ranging from 0 (*not at all threatening*) to 100 (*very threatening*), while 43% of participants in the high threat condition rated the avatar below the midpoint (<50).

Figure 4.8 Perceived threat from face stimuli

The high threat avatar was rated significantly more threatening than the low threat avatar. Colored points represent means and error bars represent standard error of the mean.



Next, we investigated whether participants' prior moral expectations varied as a function of the threat manipulation. Participants expected the agent in the high threat condition to be significantly nastier in their moral behaviour than the agent in the low threat condition (high: 42.358±1.265; low: 50.930± 1.312; $F(1,345) = 22.133$, $p < .001$, $\eta^2 = 0.060$; $Z = 4.620$, $p < .001$). Notably, the more threatening participants found the avatar, the worse their moral expectations ($\rho = -0.342$, $p < .001$). We also verified that participants in the high and low threat conditions were equally confident about their prior moral expectations (high: 60.938±1.751; low: 63.357± 1.789; $F(1,345) = 0.934$, $p = .334$, $\eta^2 =$

0.003; Z = -1.028, p = .304). This validates the use of our stimuli for manipulating perceptions of threat and moral expectations, and supports previous reports suggesting that threat evaluations are associated with worse moral expectations (Oosterhof & Todorov, 2008).

Subjective impressions of agents' moral character

Because we implemented a fully between-subject design, we could investigate whether evaluations of two agents who behave exactly the same differ as a function of their representative avatar. An ANOVA with agent (bad versus good) and threat (high versus low) as factors revealed significant main effects of agent ($F(1,343) = 132.280$, $p<.001$, $\eta^2 = 0.298$) and threat ($F(1,343) = 10.820$, $p = 0.001$, $\eta^2 = 0.027$) on final character ratings. The interaction between agent and threat was not significant ($F(1,343) = 1.310$, $p = .253$, $\eta^2 = 0.004$). The main effect of threat demonstrated that participants evaluated the perceptually more threatening agent more harshly, despite behaving identically to the less threatening agent. Thus, perceptions of threat influenced moral character impressions independently from the moral actions.

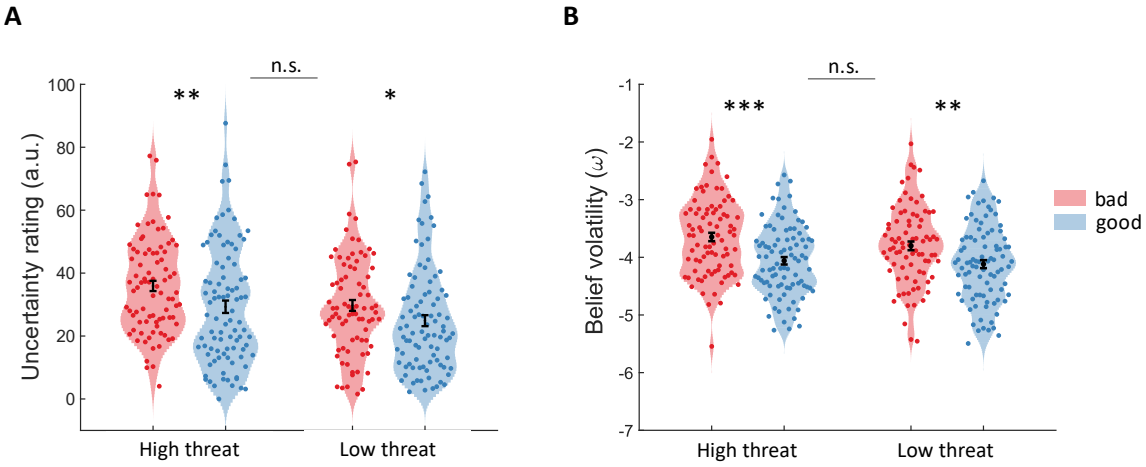
Certainty of subjective impressions and volatility of beliefs

Replicating the findings from Chapter 3, participants expressed greater uncertainty about their impressions of the bad agent than the good agent ($F(1,343) = 10.291$, $p=.001$, $\eta^2 = 0.029$; $Z = -3.746$, $p<.001$). We also observed a main effect of threat ($F(1,343) = 8.852$, $p=.003$, $\eta^2 = 0.025$; $Z = -2.933$, $p=.003$), indicating that participants were more uncertain about their impressions of the more threatening avatar. A main goal of the study was to investigate whether the learning asymmetry observed in Chapter 3 is shaped by moral

expectations. However, the interaction between agent and threat was not significant ($F(1,343) = 0.251$, p=.617, $\eta^2 = 0.001$), suggesting that manipulating prior expectations did not significantly impact the learning asymmetry (**Figure 4.9a**). Indeed, participants were significantly more uncertain about the bad agent relative to the good agent in both the high threat ($Z = -2.875$, p=.004) and low threat ($Z = -2.373$, p=.018) conditions.

Figure 4.9 Learning asymmetries are robust to manipulating prior expectations

*Uncertainty ratings (**A**) and belief volatility ω (**B**) as a function of agent (bad versus good) and threat (high versus low). Error bars represent SEM. ***P < 0.001, **P < 0.01, *P < 0.05, n.s. = not significant, a.u. = arbitrary units.*



Model estimated belief volatility, ω, was higher for the bad agent than the good agent ($F(1,343) = 26.962$, $p < .001$, $\eta^2 = 0.073$; $Z = -4.910$, $p < .001$). The main effect of threat did not reach significance ($F(1,343) = 2.265$, $p = .113$, $\eta^2 = 0.007$; $Z = -1.274$, $p = .203$), nor did the interaction between agent and threat ($F(1,343) = 0.462$, $p = .497$, $\eta^2 = 0.001$). Simple effects analysis revealed that ω was higher for the bad agent than the good agent in both the high threat ($Z = -3.988$, $p < .001$) and low threat ($Z = -2.981$, $p = .003$) conditions. Together, these findings suggest that learning asymmetries between good and bad agents are robust to manipulating prior expectations about others harmfulness.

4.4.3 DISCUSSION

In this study, our goal was to investigate the role of prior moral expectations on moral inference. We successfully manipulated moral expectations using previously validated facial stimuli that subtly varied along the dimension of threat; agents represented by more threatening faces were expected to be less moral than agents represented by less threatening faces. According to Bayesian inference, information that is inconsistent with prior expectations motivates belief updating in order to minimize uncertainty. Thus, manipulating prior moral expectations are hypothesized to have opposite effects on uncertainty and belief updating for good and bad agents: negative moral expectations should enhance belief certainty and rigidity for bad agents and reduce belief certainty and rigidity for good agents, while the reverse pattern is hypothesized for positive moral expectations. The present study found that manipulating prior expectations about moral behaviour did not follow the hypothesized pattern: Beliefs about bad agents were more volatile and uncertain than beliefs about good agents, and these effects did not covary with prior moral expectations.

An alternative hypothesis is that *negative* moral expectations motivate learning because it is especially important to learn about potential harms. The ability to systematically adapt learning according to perceived harms may have evolved to optimize learning in complex social environments to aid survival. Consistent with this theory, participants were significantly more uncertain about their character impressions when learning about a perceptually more threatening agent, regardless of whether the agent behaved in line with their expectations. While the exact relationship between *subjective* impression uncertainty and learning is not well-established, it's possible that threat-

induced uncertainty functions to motivate objective learning about preferences and belief updating. This is in line with a recent report from the non-social literature, showing that people are more uncertain about their decision-making strategies when learning to avoid losses than seek gains, which in turn enabled participants to flexibly adapt decision-making strategies when environmental contingencies changed (Lebreton, Bacily, Palminteri, & Engelmann, 2019).

While we observed the same pattern of results in model estimate ω, the main effect of threat did not reach significance. That is, although participants expressed greater uncertainty about their impressions of the more threatening avatar, this did not translate to significantly more flexible belief updating for the more threatening avatar. However, it's possible that our manipulation using computerized 2-d face stimuli was too weak to impact the mode's global estimate of belief volatility, ω. We instructed participants to predict the decisions of a past participant "who would be represented by an avatar". For participants, it was clear that the avatar was merely a 2-d computer model that was not created by the "agent" whose behaviour they were going to predict (**Figure 4.7**). This left little reason to believe that inferences drawn from the avatar should help predict the agent's behaviour. Despite this, cumulatively the high threat avatar was rated as more threatening than the low threat avatar and expected to be more harmful. However the extent to which the stimuli were perceived as threatening varied substantially across participants (**Figure 4.8**): many participants in the high threat condition rated the avatar as less threatening than some participants in the low threat condition rated the low threat avatar. Therefore, it's possible that a more robust manipulation would reveal a main effect of threat on overall belief

volatility, ω. Future work would benefit from a more robust, objective manipulation to test whether perceptions of social threat motivate belief updating.

4.5 CONCLUSION

This Chapter embarked on an investigation of the role prior moral expectations play in asymmetric learning about more harmful versus less harmful agents. Bayesian inference predicts that new information consistent with prior expectations serves to narrow the distribution over all possible moral preferences, thereby decreasing uncertainty about that agent's preference. Following from this prediction, a plausible explanation for more volatile, uncertain beliefs about the bad agent's preferences in the Moral Inference Task is that people hold optimistic prior expectations about other's morality. Revisiting the data from Chapter 3, I performed new analyses to test whether prior moral expectations explain the observed pattern of results and found no evidence to this effect. Incentivizing a separate group of participants to predict, in the context of decisions to profit from others' pain, how "most people" would choose found no evidence for optimistic prior expectations either. Independently manipulating prior expectations using threatening and non-threatening facial stimuli revealed that negative moral expectations were associated with more uncertain character ratings, regardless of whether the agent's behaviour was consistent with prior expectations or not. Thus, although moral inference generally follows the same principles of Bayesian updating (with precision-weighted prediction errors driving belief updating), there appears to be additional mechanisms at play. Even with a reasonably good estimate of another's preferences, moral information may systematically bias uncertainty and belief updating in order to preserve cognitive resources for reducing uncertainty about salient others.

Chapter 5

5 EXPOSURE TO VIOLENCE DISRUPTS THE DEVELOPMENT OF SUBJECTIVE IMPRESSIONS AND ADAPTIVE TRUST BEHAVIOUR

This chapter is an extended version of a paper published as:

Siegel, J.Z., Estrada, S., Crockett, M.J., & Baskin-Sommers, A. Exposure to violence affects the development of moral impression and trust behaviour in incarcerated males. *Nature Communications*, 10.1 (2019): 1942.

Doi: https://doi.org/10.1038/s41467-019-09962-9

Exposure to community violence is a reliable predictor of negative life outcomes (e.g., problems with health, mental health, chronic aggression). Notably, individuals exposed to violence are more likely to engage in antisocial behaviour and, as a result, exposure to violence dramatically increases the likelihood of involvement in the justice and social service systems. Theoretical accounts suggest that disruptions in learning underlie the link between exposure to violence and maladaptive social behaviours (e.g.,

aggression, antisocial behaviour). However, empirical evidence specifying these processes is sparse. Here, we investigated how exposure to violence affects the ability to learn about the harmfulness of others and use this information to adaptively modulate trust behaviour in a sample of currently incarcerated males. Participants predicted the choices of two agents who repeatedly decided whether to inflict painful electric shocks on another individual in exchange for money. The agents differed substantially in their harmfulness, in that the "good" agent required more compensation to harm than the "bad" agent. Participants periodically rated their subjective impressions of the agent's moral character, as well as their certainty of their impressions. After completing the learning task, we assessed how participants interacted with each agent in a one-shot trust game. Results indicated that exposure to violence did not impact the ability to accurately develop beliefs about the agents' harm preferences and predict their choices. However, exposure to violence disrupted the ability to form moral impressions that dissociated between agents with distinguishable harm preferences. Consequently, participants with higher exposure to violence had more difficulty adjusting their trust behaviour towards the two different agents. Our findings reveal a novel cognitive process that may explain the emergence of maladaptive behaviour related to exposure to violence.

5.1 INTRODUCTION

Exposure to community violence, whether it is witnessing someone get chased or hurt, hearing gunshots in the neighborhood, or being directly chased, assaulted or shot at, is a significant public health concern. In the United States, over three-quarters of youth have been exposed to some form of community violence in their lifetime (Finkelhor, Turner, Shattuck, & Hamby, 2013, 2015). In general, both cross-sectional and longitudinal

research finds that exposure to violence places young people at risk for persistent academic underachievement (Delaney-Black et al., 2002), physical health problems (e.g. difficulty sleeping, headaches, heart disease, immune disease) (Bailey et al., 2005; Moffitt & Tank, 2013)), mental health problems (e.g. depression, anxiety, post-traumatic stress, antisocial personality (D. Baskin & Sommers, 2015; Fowler, Tompsett, Braciszewski, Jacques-Tiura, & Baltes, 2009; Moffitt & Tank, 2013)), and interpersonal problems (e.g. problems with trust, lower levels of empathy (Guo et al., 2013)). Additionally, individuals exposed to violence are more likely to engage in antisocial behaviour (D. Baskin & Sommers, 2015; Fowler et al., 2009; Javdani, Abdul-Adil, Suarez, Nichols, & Farmer, 2014), show earlier and more chronic aggressive behaviour (DuRant, Pendergrast, & Cadenhead, 1994), and hold beliefs that can normalize or romanticize aggression (Guerra, Huesmann, & Spindler, 2003). As a result, exposure to violence dramatically increases the likelihood of involvement in the justice and social service systems (Hawkins et al., 2000).

Exposure to violence predisposes some individuals to diverse forms of negative life experiences and mental health problems, as well as comprises a prominent risk factor for a lifetime mired in aggression. Chronic exposure to violence, whether in a larger community or justice system context, shapes cognition in a way that is likely to distort perceptions of what is considered harmful behaviour and how to react to harmful behaviour. For example, Dodge and colleagues (Dodge et al., 1990) found that physically abused children attended less to social cues and were biased to interpret the behaviour of neutral actors as hostile, relative to children who did not experience abuse. These information processing deficits predicted aggressive behaviour six months later and mediated the relationship between abuse and aggression. The research provides initial evidence that exposure to violence may

lead to maladaptive behaviour through the development of distorted social information processing patterns. However, the precise social cognitive processes that may underlie these distortions in individuals exposed to violence is unclear. At the core of several theories about the relationship between exposure to violence and aggressive/antisocial behaviour is the role of learning (Albert Bandura, 1978; Guerra et al., 2003; Huesmann & Kirwil, 2007; Ng-Mak, Salzinger, Feldman, & Stueve, 2004; Ng-Mak et al., 2002). However, empirical evidence identifying and specifying the way in which learning is disrupted and can affect behaviour in individuals exposed to violence remains elusive.

One aspect of learning that is especially relevant to adaptive social behaviour is learning about whether other individuals might harm us. Harmfulness is a core dimension of moral character (Haidt & Joseph, 2004; Schein & Gray, 2018) and is therefore important for inferring the morality of others (i.e., moral infenece). We suggest that there exists two, distinct, components of moral inference. On the one hand, people use social cues to objectively update their beliefs about others' moral preferences by gradually accumulating information over time to predict future outcomes (i.e., in a Bayesian manner). On the other hand, people form subjective impressions about moral character that emerge rapidly and effortlessly (Engell et al., 2007; Todorov et al., 2009). These beliefs and moral impressions are used to adaptively learn and decide whom to trust in social interactions (Haidt & Joseph, 2004; Stanley et al., 2011). For example, in Chapter 3 participants entrusted more money in a one-shot trust game to agents who were less willing to harm others for profit and ascribed better moral character (subjective impression) to those agents compared to agents who were more willing to harm for profit. Notably, subjective character impressions predicted trust game behaviour. Together, these components of learning about other's serve

as powerful informational tools; for the purpose of survival, humans are evolutionarily inclined to identify potential foes and avoid them through adaptive social decision-making (Alexander, 1987; Gintis, 2000). However, life experiences, such as exposure to violence, are likely to shape moral inferences and resulting social behaviours. Prior research linking exposure to violence to normalized views of aggression and aberrations in interpersonal functioning raises the possibility that exposure to violence may impact learning about the morality of others, and by extension, behaviours that rely on trust. To date, however, there has been no research on exposure to violence and social learning about harm.

To examine the relationships among exposure to violence, moral inference, and trust behaviour, we administered the Moral Inference Task to a sample of incarcerated males. While a sample of currently incarcerated individuals is not the same as a sample from the general population, this type of sample does serve as an informative sample in which to explore how differences in exposure to violence impact moral inference. It is well-documented that exposure to violence among the incarcerated covers the full continuum of potential experiences compared to the general population where scores are often restricted in range and narrowly centered around a few points within that range (D. Baskin & Sommers, 2014; Finkelhor, Turner, Ormrod, & Hamby, 2010; Fitzpatrick & Boldizar, 1993). Moreover, by studying a sample of currently incarcerated individuals, we are better poised to investigate the variation in exposure to violence within a sample that is already demonstrating the theorized behavioural effects of such exposure.

In this task, participants predicted and observed the choices of two agents who repeatedly decided whether to inflict painful electric shocks on another individual in exchange for money. The two agents substantially differed in their preferences towards

harm (i.e., their exchange rate between money and pain). On each trial, participants predicted the choice made by the agent and received immediate feedback about their accuracy. Every three trials, participants rated their overall impression of the agent's moral character (on a scale from "nasty" to "nice") and their certainty of that impression. This task enabled us to measure two distinct components of moral inference: the ability to develop accurate *beliefs* about the agents' objective exchange rates between money and pain (a quantity that is used to predict their choices), and the use of those estimates to form subjective, global *impressions* about other's moral character. After the Moral Inference Task, participants engaged in a one-shot trust game (Berg et al., 1995) with each of the agents. All participants completed a battery that assessed exposure to violence, as well as a clinical assessment measuring different aspects of antisociality to address potential confounds.

5.2 METHODS

Participants

The present sample included 119 males from a high-security correctional institution in Connecticut. We used a prescreen of institutional files and assessment materials to exclude justice-involved individuals who: were not between the ages of 18 and 75; scored below 70 on a brief measure of IQ (Shipley Institute of Living Scale (Zachary, 1991)) performed below the fourth-grade level on a standardized measure of reading (Wide Rage Achievement Test-III (Wilkinson, 1993)) had diagnoses of schizophrenia, bipolar disorder, or psychosis, not otherwise specified; were currently taking psychotropic medication; or had a history of medical problems (e.g., uncorrectable auditory or visual deficits, head

injury with loss of consciousness greater than 30 min, seizures, neurological disorders) that may impact their comprehension of the materials. The Yale University Human Investigation Committee approved the procedures used in the present study. The study complied with all relevant ethical regulations for work with human participants and all participants provided written informed consent. See **Table 5.1** for participant demographic information.

Table 5.1

Participant demographic information

	Descriptive Statistics				
	Minimum	Maximum	Mean		Std. Deviation
	statistic	statistic	statistic	Std. Error	statistic
Age on Date of Interview	20.00	58.00	34.98	0.911	9.940
Individual Income Past Year	1.00	10.00	1.61	0.165	1.796
Family Income Past Year	1.00	11.00	7.82	0.351	3.667
Highest Level of Education	5.00	16.00	11.59	0.168	1.829
PCL-R Total Score	5.30	37.0	23.37	0.649	7.080
Symptom Count Adult APD	0.00	7.00	3.82	0.163	1.775
ETV Total Score	1.00	13.00	8.04	0.295	3.219
CTQ Total Score	25.00	97.00	43.37	1.490	16.251
Total # violations in prison	0.00	93.00	7.45	1.106	12.070
Years of incarceration	0.33	30.80	6.83	0.661	7.179

Experimental Procedure

Primary measures:

Moral Inference Task. Participants completed the Moral Inference Task described in Chapter 2. To incentivize participants to learn about the moral preferences of the agents, before beginning the task they were instructed that "it is important that you pay close attention to the deciders and learn about their behaviour, because you will interact with

each of them in a computerized trust game at the end of the study which could earn you reward points."

Trust game. After completing the moral learning task, participants played a trust game with each of the agents. In the game, participants were endowed with 100 "points" that they could entrust with each agent. Any amount that they entrusted with the agent would be tripled, and the agent could then choose how much of the tripled amount to return to the participant. We instructed participants that the percent returned by each agent had been predetermined, and thus the agents were not playing actively. We set the returned amount to correspond to the agents' actual harm tendencies, such that the good agent behaved less selfishly than the bad agent and therefore returned a larger proportion of the entrusted points. The final number of points was tallied and the top five earners were added to a "leader board" that was on display to all study participants in the testing room (*Note*: The Connecticut Department of Correction did not allow researchers to pay justice-involved individuals). Of particular interest for this study was the difference in amount entrusted with the good agent vs. the bad agent (Δentrust = amount entrusted with the good agent – amount entrusted with the bad agent).

Exposure to Violence Scale (ETV). The ETV scale (Selner-O'Hagan, Kindlon, Buka, Raudenbush, & Earls, 1998) was used to measure lifetime exposure to violent events. The questionnaire consisted of 13 items, documenting the types of both experienced and observed violence (e.g., "Have you been hit, slapped, punched, or beaten up?" and "Have you seen someone else get attacked with a weapon, like a knife or bat?"). Participants were asked to respond to each item based on a dichotomous choice (*yes*/*no*). If *yes* was selected, participants indicated the number of times they experienced this situation in their lifetime.

The two scales, experienced and observed, showed moderate overlap ($\rho = 0.607$, $p < 0.001$). Thus, we examined a total exposure to violence score using a sum of all 13 items. Internal consistency for ETV total score was .86. ETV scores were normally distributed (skewness: -0.605, kurtosis: -0.810). Ninety-nine percent of the sample reported experiencing at least one exposure to violence in their lifetime and approximately 30% of the sample reported experiencing over 9 (the median) different exposures to violence in their lifetime. Lifetime frequency of exposure to violence ranged from 2 times to 11,465 times (median = 88).

General Trust Scale (Yamagishi & Yamagishi, 1994). The General Trust scale was used to measure general beliefs about the honesty and trustworthiness of others. Participants were asked to indicate to what extent they agree (*1*) or disagree (*5*) with six statements (e.g., 'Most people are trustworthy'). The scores from each statement were averaged together to produce a continuous measure of generalized trust.

Covariates:

Hare Psychopathy Checklist Revised (PCL-R). The PCL-R (R. Hare, 2003) used information gleaned from a life-history interview and a review of institutional files to score participants on the presence of 20 different items (e.g., superficial charm, shallow affect, impulsivity, criminal versatility). A score of 0, 1, or 2 was given for each item according to the degree to which a characteristic was present. PCL-R total scores ranged from 0 to 40. The reliability and validity of the PCL-R has been well established (R. Hare, 2003; R. D. Hare et al., 1990). Inter-rater reliability for 24% of the sample was .991 (alpha).

Antisocial Personality Disorder (APD). Participants were assessed for APD during a semi-structured diagnostic interview. The interview evaluated the age and frequency of

engagement in behaviours outlined in the Diagnostic Statistical Manual-5(American Psychiatric Association, 2013) (DSM). A diagnosis of APD was given if there was evidence of conduct disorder (CD) prior to age 15 and sufficient adult antisocial symptoms (e.g., aggression, irresponsibility). Inter-rater reliability for 32% of the sample was .989 (Cohen's kappa).

Childhood Trauma Questionnaire (CTQ). We used the CTQ (Bernstein et al., 2003), a 28-item questionnaire, to assess maltreatment experiences prior to age 18. It consisted of five clinical scales: emotional abuse, physical abuse, emotional neglect, and sexual abuse. Items were rated on a 5-point Likert-type scale with response options ranging from "Never True" to "Very Often True." For the present study, the total score was examined. For this sample, the total score demonstrated good internal consistency (Cronbach's $\alpha = 0.824$).

Computational modelling

Three computational models were compared to describe how participants learned the agents' preferences and predicted their choices. We fit the HGF (Mathys et al., 2011, 2014), which identified participant-specific parameters to describe each individual participant's learning process. Beliefs about an agent's harm preference were updated using a Bayesian reinforcement learning algorithm, with precision-weighted prediction errors driving belief updating at the different levels of the hierarchical model. Second, we fit a Rescorla Wagner model, in which beliefs were updated by prediction errors with a fixed learning rate. Third, we fit a modified Rescorla Wagner model, in which beliefs were updated by prediction errors with separate fixed learning rates for helpful and harmful outcomes. For details about the alternative models, see Table 2.2. The log-model evidence (LME) indicated that the

HGF model (sum LME = -5920) outperforms both a simple single learning rate RW model (sum LME = -6376) and a RW model with separate learning rates for positive and negative outcomes (sum LME = -6055). We validated these findings using formal Bayesian Model Selection, which is a random-effects procedure that takes into account inter-subject heterogeneity (Rigoux et al., 2014; Stephan et al., 2009). To this end, we used LME data to compare between the HGF and our two RW models. This analysis yielded a protected exceedance probability indistinguishable from 1 for the HGF model for both agents, indicating effectively a 100% probability that the HGF model better explains the data than the other models included in the comparison.

Statistical analysis

All data analysis was completed in Matlab (Mathworks) and PASW Statistics 24 (SPSS/IBM). All statistical tests were two-sided. We used robust linear regression models with a bisquare weighting function to analyze the z-scored trial-by-trial rating data (impression and certainty ratings). We used nonparametric statistical tests that do not make any assumptions about the underlying distributions of variables (e.g., Spearman's ρ and signed rank tests). To investigate whether the relationship between ETV score and differences in social behaviour were mediated by differences in participants final harm judgments of the agents we used the PROCESS macro for SPSS (Hayes, 2012).

5.3 RESULTS

Beliefs and predictions about moral preferences:

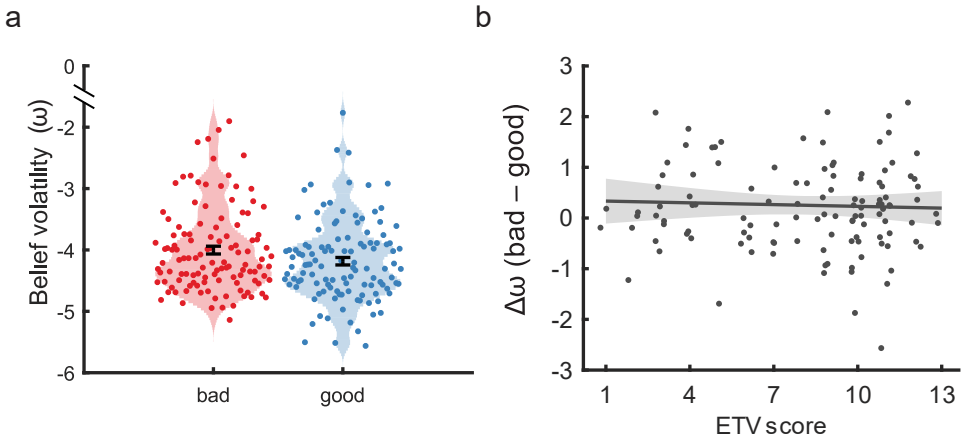
We first investigated participants' ability to develop accurate beliefs about the agents' objective harm preferences and predict their decisions. On average, participants

predicted accurately 72% of the good agent's choices and 77% of the bad agent's choices. There was no relationship between ETV score and prediction accuracy for either agent (Spearman's ρ, good: ρ = -0.065, p = .483; bad: ρ = 0.043, p = .639). This suggests that participants with higher exposure to violence were equally motivated to learn the harm preferences of the agents, relative to those with lower exposure to violence.

Next, we examined how rapidly participants updated their beliefs about the agents' preferences in response to feedback. Replicating previous findings (Chapter 3), beliefs about the bad agent's preferences were more rapidly updated in response to feedback, as indicated by a higher ω, than beliefs about the good agent's preferences (signed rank test, Z = -2.328, p = .020; **Figure 5.1a**). ETV score was not significantly related to ω for either agent (Spearman's ρ, good: ρ = 0.025, p = .785; bad: ρ = -0.014, p = .879), nor was it related to the difference in ω between good and bad agents (Δω: ρ = 0.008, p = .929; **Figure 5.1b**). Together, these results suggest that objective harm learning was largely intact in this sample and did not covary with exposure to violence.

Figure 5.1 Learning rate does not covary with exposure to violence

*(**a**) Beliefs about the bad agent's harm preferences were more volatile than beliefs about the good agent's harm preferences. (**b**) Between-agent asymmetries in belief updating were not related to participant's ETV score, suggesting that exposure to violence does not significantly impact the underlying processes of objective social learning. Error bars represent standard error of the mean. Error bands represent 95% confidence intervals.*



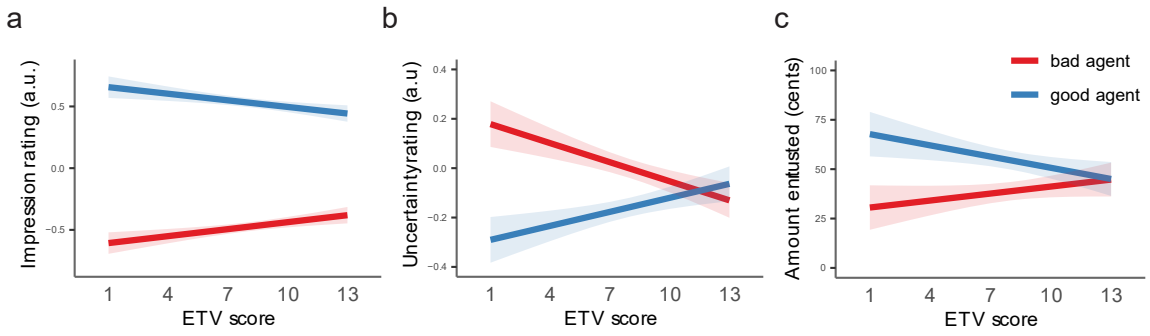
Subjective impressions of the agent's moral character:

Despite the fact that ETV score did not impact learning the *objective* features of agents' preferences, we observed a strong effect of ETV score on participants' *subjective* global impressions of the agent's moral character. A robust linear regression was used to predict subjective moral impression ratings as a function of agent (good vs. bad), ETV score, and their interaction. We included an additional regressor to control for trial number, and found no effects on impression ratings (main effect of time: $\beta = 0.003 \pm 0.003$, $t = 0.955$, $p = .340$). Replicating previous chapters, in general, participants formed more favourable impressions of the good agent's moral character than the bad agent's moral character (main effect of agent: $\beta = -1.300 \pm 0.075$, $t = -17.284$, $p < .001$). Higher ETV scores predicted more negative impressions of the good and bad agents' moral character (main effect of ETV: $\beta = -0.018 \pm 0.006$, $t = -2.902$, $p = .004$). There was a significant interaction between ETV

score and type of agent (interaction: $\beta = 0.037 \pm 0.009$, $t = 4.222$, $p < .001$), indicating that for participants with higher ETV scores, there was less differentiation in their impressions of the good and bad agents' moral character (**Figure 5.2a**).

Figure 5.2 Diminishing effects of agent with increasing exposure to violence.

Participants with higher ETV scores showed less differentiation in their subjective impressions of good versus bad agent's moral character (A) and reported smaller discrepancies in the uncertainty of their impressions of good and bad agents (B). Higher ETV scores also resulted in smaller discrepancies in the amounts that participants entrusted with good versus bad agents in a one-shot trust game (C). Y-axis in figures A and B denote standardized values (z-scored). Error bands represent 95% confidence intervals.



To further investigate the interaction between type of agent and exposure to violence on subjective impression ratings, we ran separate regressions on ratings for the good and bad agent. Specifically, we asked whether diminishing effects of agent with higher ETV scores were driven by the good agent, the bad agent, or both. These analyses revealed that the effects were not specifically driven by either agent alone. Higher ETV scores predicted more favourable impressions of the bad agent (main effect of ETV: $\beta = 0.019 \pm$, $p = .004$) and less favourable impressions of the good agent ($\beta = -0.019 \pm$, $p = .001$; see **Figure 5.2a**).

Participants with higher ETV scores were no more likely to predict worse harm intentions of the agents in the task before they observed any of their choices ($\rho = -0.082$, $p = 0.361$) and were no less trusting of others in general (as indicated by scores on the General

Trust scale; ρ = -0.040, p = 0.665). These findings suggest that exposure to violence affects how participants in this sample form subjective impressions about other's moral character through observing their choices, rather than affecting prior moral expectations about others.

Certainty of subjective impressions:

Work from previous chapters in non-incarcerated samples indicates that adults hold more certain positive impressions of others and more uncertain negative impressions, which is hypothesized to serve the adaptive social function of enabling people to more flexibly update negative impressions that turn out to be inaccurate (Rand et al., 2009; Siegel, Mathys, Rutledge, & Crockett, 2018). Consistent with previous research, uncertainty decreased over time as participants were exposed to more information about the agents' harm preferences (main effect of time: β = -0.018±0.003, t =-5.969, p < .001). Furthermore, participants expressed greater uncertainty in their impressions of the bad agent, relative to the good agent (main effect of agent: β = 0.513±0.078, t = 6.605, p < .001).

To investigate whether exposure to violence affected participants' uncertainty in their impressions of the agents' moral character, we performed a robust linear regression to investigate the effects of agent, ETV score, and their interaction on participants' ratings of uncertainty about their impressions of the agents. Participants with higher ETV scores were more uncertain in their impressions overall (main effect of ETV: β = 0.019±0.006, t = 2.982, p = .003), and this interacted with the effect of agent (interaction: β = -0.045±0.009, t = -4.973, p < .001; **Figure 5.2b**). Consistent with our findings on subjective impressions, the interaction indicated that participants with higher ETV scores expressed smaller differences in their uncertainty ratings between good and bad agents, such that they

became more uncertain that the good agent was good, and less uncertain that the bad agent was bad. Notably, smaller differences in impression ratings between good and bad agents predicted less discrepant uncertainty ratings ($\rho = 0.336$, $p < .001$). Results from our robust linear regressions did not change after controlling for age and education (see Appendix F:).

To investigate the nature of the interaction between type of agent and exposure to violence on uncertainty ratings, we ran separate regressions on ratings for the good and bad agent conditions. Mirroring the subjective impression results, these analyses revealed that the effects were not specifically driven by either agent alone. Higher ETV scores predicted less uncertain impressions of the bad agent (main effect of ETV: $\beta = -0.025 \pm 0.007$, $t = -3.795$, $p < .001$) and more uncertain impressions of the good agent ($\beta = 0.019 \pm 0.006$, $t = 3.126$, $p = .002$; see **Figure 5.2b**).

Trust behaviour

Although exposure to violence impaired participants' ability to form distinct subjective impressions of agents with different harm preferences, it is unclear whether this has consequences for social behaviour. To address this question, we asked participants to engage in a one-shot trust game with each of the agents, after predicting all the agents' choices in the Moral Inference Task. Previous work, including work from Chapter 3 and Chapter 4, has shown that non-incarcerated adults adjust their behaviour in the trust game according to the harm preferences of the agent with whom they are interacting (i.e., people entrust significantly less money with those who treat others poorly than those who treat others well (Delgado et al., 2005)).

To investigate whether adaptive trust behaviour was diminished in participants with higher ETV scores, we entered the amount participants entrusted in a repeated measures general linear model with agent (good vs. bad) as the within-subject factor and ETV score as a covariate. Consistent with previous research, participants entrusted more points with the good agent than the bad agent (main effect of agent: $F_{(1,119)} = 6.202$, $p = .014$, $\eta^2 = 0.056$). The effect of agent was significantly moderated by ETV score (interaction between agent and ETV: $F_{(13,119)} = 2.142$, $p = .017$, $\eta^2 = 0.210$; **Figure 5.2c**). The interaction indicated that higher ETV scores predicted smaller discrepancies in the amount participants entrusted with the good versus the bad agent. Specifically, those with higher ETV scores entrusted significantly less with the good agent ($\rho = -0.220$, $p = .016$), and consequently ended up earning fewer points overall ($\rho = -0.325$, $p < .001$). ETV scores did not significantly affect the amount that participants entrusted with the bad agent ($\rho = 0.119$, $p = .198$). Thus, exposure to violence was associated with maladaptive trusting behaviour, specifically when interacting with those who are less willing to harm others, and this had a negative impact on their overall earnings.

Given the relationship between exposure to violence and differential trust behaviour (Δentrust, calculated as amount entrusted with good agent – amount entrusted with bad agent), it is possible that participants' final subjective impressions of the agents' moral character (Δjudgment, calculated as final impression of the good agent – final impression

of the bad agent) account for (i.e., mediate) that relationship[1]. Regression analysis was used to investigate the hypothesis that Δjudgment mediates the effect of exposure to violence on Δentrust. Results indicated that ETV score was a significant predictor of Δjudgment, effect = -0.025 ± 0.012, p = .041, and that Δjudgment was a significant predictor of Δentrust, effect = 32.582 ± 6.500 p < .001. These results support the mediational hypothesis. Exposure to violence remained a significant predictor of Δentrust after controlling for the mediator, Δjudgment, effect = -2.341 ± 0.863, p = 0.008. Approximately 25% of the variance in Δentrust was accounted for by the predictors (R^2 = .251). The indirect effect was tested using a bootstrap estimation approach with 5000 samples. We found a significant indirect effect of Δjudgment on Δentrust, effect = -0.812, 95% CI [-1.705, -0.054], suggesting that impressions about other's moral character account for differences in social behaviour among participants with higher levels of exposure to violence. Thus, higher exposure to violence was associated with increasingly maladaptive trust behaviour as mediated by decreased subjective impression sensitivity.

Despite these results, some might question the validity of the trust game in the current sample given their incarceration status. Therefore, we examined whether the extent to which participants adjusted their trust behaviour according to the agents' harm preferences predicted social behaviour in prison. Less discrepant behaviour towards good and bad

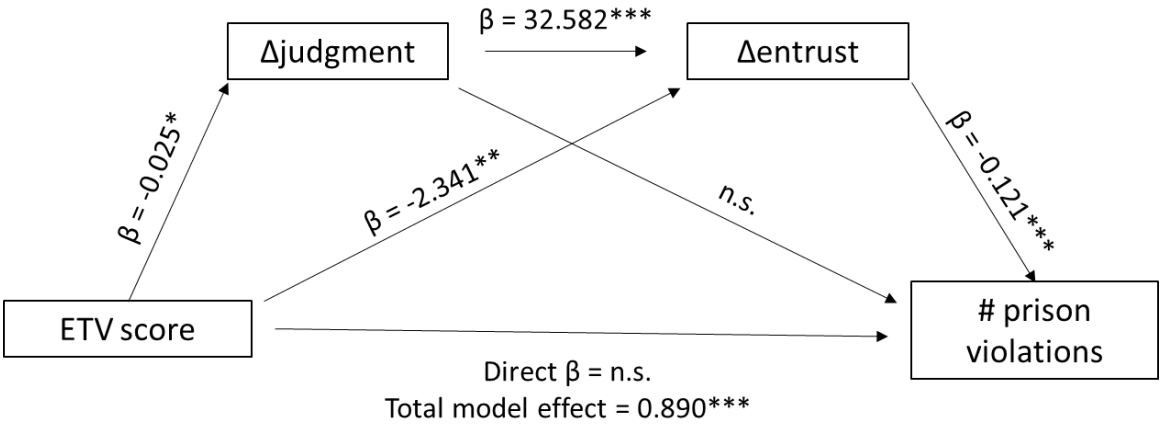
[1] It is possible that objective learning and trust are associated, such that a participant who is less able to predict the agents' choices would have greater difficulty distinguishing trust behavior for agents who behave differently. In fact, we find a significant association between accuracy and trust behavior (P < 0.001), where increased accuracy was associated with a greater tendency to adapt trust behavior to agents with different moral preferences. However, there was no impact of ETV score on that relationship. Together, these findings suggest that exposure to violence *does not* impact the association between the ability to learn preferences of others, and moreover, use that information to engage in trust behavior. However, as demonstrated in the main analysis, exposure to violence *does* impact the ability to form subjective impressions based on distinguishable behaviors, and subsequently adapt trust behavior accordingly.

agents was associated with more behavioural violations in prison (ρ = -0.208, p = .023), and specifically with aggressive violations against persons (ρ = -0.217, p = .020). This suggests the ability to adjust trust behaviour based on impressions of other's moral character, as measured by our task, captures variance in real-world social behaviour.

However, it's possible that this relationship between less discrepant behaviour towards good and bad agents and more behavioural violations in prison is largely explained by the relationship with ETV scores. Indeed, higher ETV scores predict more behavioural violations in prison (ρ = 0.450, p <.001). However, we predicted that this relationship would be mediated by the extent to which participants differentiated in their subjective impressions and trust behaviour between the good and bad agent. Consequently, we applied a serial multiple mediation analysis using the PROCESS macros for SPSS (Hayes, 2012) (model 6) that allowed us to determine the causal link between mediators with a specified direction of causal flow. We investigated whether the relationship between exposure to violence and prison violations was mediated by trust behaviour (Δentrust) as a function of impression sensitivity (Δjudgment). ETV score was only a marginally significant predictor of prison violations after impression sensitivity and trust behaviour were accounted for (effect = 0.622 ± 0.340, p = 0.070). The indirect effects were tested using a bootstrap estimation approach with 5000 samples. These results indicated the indirect serial coefficient was significant, effect = 0.099 ± 0.071, 95% CI = [0.002, 0.274] (for complete serial multiple mediation results, see **Figure 5.3**), suggesting that disruptions in the ability to form distinguishable impressions resulting from higher ETV scores, translates into maladaptive trust behaviour, which in turn leads to a greater number of direct violations in prison.

Figure 5.3 Serial multiple mediation analysis

*n.s = not significant; *P < 0.05; **P < 0.01; ***P < 0.001*



Previous work has shown that exposure to violence is associated with antisocial behaviour and psychopathic traits (Baskin-Sommers & Baskin, 2016; Baskin-Sommers et al., 2016; Guerra et al., 2003; Kimonis, Frick, Munoz, & Aucoin, 2008). Here we also found that ETV score was associated with increased antisociality, as indicated by higher PCL-R score and increased symptoms of APD (PCL-R total score: $\rho = 0.394$, $p < .001$; APD symptoms: $\rho = 0.261$, $p = .004$). This leaves open the question of whether the effects of exposure to violence on subjective impressions observed here are a primary consequence of exposure to violence, or an indirect consequence of possessing characteristics that predispose exposure to violence, such as psychopathy or antisocial personality disorder. To assess whether ETV score had a direct, as opposed to indirect, effect on subjective impressions and uncertainty ratings, and social behaviour, we entered each covariate (PCL-R total score and total number of antisocial personality disorder symptoms) separately into our regressions with ETV score. Across all measures, we found that interactions between agent and ETV score remain significant even when we include the interaction between

agent and each covariate (see **Appendix F:** for all analyses including covariates). An alternative possibility is that the observed effects of exposure to violence on impressions reflect a general impact of traumatic experiences, rather than being specific to community violence. To investigate, we entered scores from the CTQ into our regression with ETV score. Again, we found that the interactions between agent and ETV score remain significant even after controlling for CTQ (see **Appendix F:** for full analysis). Together, this suggests that being exposed to violence had a direct effect on subjective impressions of moral character and social behaviour and that findings could not be entirely explained by antisocial psychopathology or childhood trauma.

5.4 DISCUSSION

The ability to infer other's intentions and predict their behaviour is crucial for successful social interactions. In particular, learning whether others are likely to harm us is important for consequential social decisions like deciding whom to trust. However, there are environmental experiences that may impact how we learn about harm and use this information to make adaptive social decisions. Exposure to violence is one environmental experience that is associated with aberrations in beliefs about harm (Guerra et al., 2003; Ng-Mak et al., 2002). As a result, exposure to violence is related to behaviours that reflect a lack of trust and prosociality (e.g., aggression, crime), increasing contact with systems of social control.

The current data suggest that, in a sample of currently incarcerated males, exposure to violence adversely impacts some components of social learning, but not all. Participants with higher ETV scores showed an ability to develop accurate beliefs about others by

objectively encoding their harm preferences. However, exposure to violence appeared to disrupt the formation of subjective, global impressions of other's moral character from observed harm behaviour (i.e., character impression formation). Participants with higher ETV scores formed more positive and less uncertain impressions of harmful agents and more negative and less certain impressions of helpful agents. Moreover, these differences in subjective impressions associated with higher ETV scores led to maladaptive trust behaviour, such that participants with higher ETV scores extended less trust than optimal when interacting with a "good" agent. Finally, the link between exposure to violence and maladaptive trusting behaviour was mediated by the disturbances in impression formation. In turn, this led to significantly more violations in prison, suggesting that the effects of ETV on real social behaviour in prison is predicted by subjective impressions and trust behaviour as measured by our task. On the whole, these findings raise the intriguing possibility that exposure to violence does not fundamentally disrupt all components of social learning, but instead may produce a problem with generating global subjective social impressions and translating those impressions into adaptive social decision-making.

Our findings are consistent with evidence that the ability to *learn* the value of information is cognitively and neurally distinct from the ability to *use* learned information to guide decision-making and behaviour (Noonan et al., 2010; Zapparoli et al., 2018). Cognitively, participants with higher ETV scores were able to learn harm preferences of different agents but formed subjective impressions that appeared to "normalize" beliefs about harm in the bad agent (Guerra et al., 2003; Ng-Mak et al., 2002), seeing that agent as more similar to the good agent. This dissociation between learning and using learned information is consistent with neural lesion studies show that the lateral orbitofrontal cortex

(OFC) is associated with learning the value of stimuli, whereas the medial OFC is associated with translating stimulus value representations into decisions (Noonan, Kolling, Walton, & Rushworth, 2012). Although there are no imaging studies that directly examine exposure to violence, studies looking at the combination of physical abuse and adversity (e.g., seeing intimate-partner violence, bullying, physical and sexual abuse) note structural and functional abnormalities in the OFC (De Brito et al., 2013; McCrory, Brito, A, & Viding, 2011; McCrory, Brito, & Viding, 2010), and individuals (both incarcerated and non-incarcerated) prone to aggressive and antisocial behaviour also display abnormal medial OFC structure and function (Blair, 2001, 2004; Buckholtz, 2015; Nelson & Trainor, 2007). Taken together, findings from these parallel literatures suggest that diminished use of learned information may be a consequence of OFC dysfunction in individuals exposed to violence.

An especially unfortunate consequence of disrupted use of harm learning may be a pervasive inability to develop healthy social relationships with trustworthy individuals and a greater likelihood of placing trust in the "wrong" people. Consistent with findings in the present study showing the impact of subjective impressions on trust behaviour, research from the fields of sociology, psychology, and economics highlight that individuals who reside in communities with high rates of crime and disorder experience mistrust in their interactions with strangers, prosocial members of their community, and institutions (Elijah Anderson, 1994; Besbris, Faber, Rich, & Sharkey, 2015; Raudenbush, 2016). Justice-involved individuals tend to reside in these types of communities characterized by crime and disorder, where social interactions and systems that may provide pathways out of serious and persistent offending are absent, severely debilitated, or sparse (Elijah

Anderson, 1994; Wilson, 2009). Moreover, once caught within the justice system, it is likely that these types of interactions are reinforced through new exposures to violence and negative social interactions (Boxer, Middlemass, & Delorenzo, 2009). Combined, community context and justice-involvement translates to lower access to informal and formal resources, homebased learning, and chronic re-exposure to violence (Buka, Stichick, Birdthistle, & Earls, 2001; Hetey & Eberhardt, 2018). The resulting pattern is that some individuals, like justice-involved individuals, are more likely to live in communities of unrelenting social and economic deprivation. These environmental characteristics do not solely impact the incarcerated but spills over to other community members: those who are incarcerated are released and their behaviour and experiences impact family members, social acquaintances, and strangers in their own communities. For that matter, the disproportionate presence of incarcerated individuals in disadvantaged communities is not seen as aberrant but is often just part of living in these communities (Goffman, 2009). Thus, the combination of environmental characteristics and disruptions in the cognitive processes at the individual level are critical for the development and maintenance of trust and may ensnare individuals in a trajectory that continually reinforces maladaptive social connections, ultimately limiting chances for economic stability (Fehr, 2009; Knack & Keefer, 1997; LaPorta, Lopez-de-Silane, Shleifer, & Vishny, 1996) and psychosocial wellbeing (Moffitt & Tank, 2013).

Before concluding, methodological and conceptual limitations should be noted. The present sample is limited to incarcerated offenders, and thus we do not know whether or how incarceration status may impact the relationship between exposure to violence and harm learning. However, it is important to note that all task main effects replicated previous

findings in non-incarcerated samples. For instance, Chapters 3 and 4 use the same task and show that people form less positive, more uncertain, and more volatile beliefs about the bad agent, relative to the good agent, and adjust their trust behaviour according to the harm preferences of the agents. We observe the same pattern of results in our sample of incarcerated individuals. Moreover, length of incarceration (see **Appendix F:**) and other correlates known to increase risk for incarceration did not impact the reported exposure to violence effects. Ultimately, being currently incarcerated is just one type of adverse outcome related to exposure to violence that should not be seen as excluding the importance of the lived experience of exposure to violence for these individuals (Casciano & Massey, 2008; Garbarino & Sherman, 1980; Goffman, 2015; Monahan, King, Shulman, Cauffman, & Chassin, 2015; Tangney, Stuewig, & Mashek, 2007).

While it may be useful to replicate the findings in a sample of non-incarcerated individuals, this raises important experimental considerations. From a scientific perspective, using a sample with sufficient variability in ETV scores, and whose experience with exposure to violence has led to great personal cost, is essential. Notably, the distribution of ETV scores in our sample of incarcerated individuals covers the full range of the scale. Endeavors in samples typical of psychology research, such as university or crowdsourced samples, often suffer from restricted range in ETV scores. Nonetheless, to test for generalizability, future research should replicate the present findings in a sample of non-incarcerated individuals whose ETV scores are reflective of a range of experiences.

A final consideration is that implementing shocks in the harmfulness learning task is not as extreme a behaviour as what might be seen in the real world (e.g., sexual assault, murder) for individuals exposed to violence or involved in the justice system. Therefore, it

is possible that the objective learning of other's harm preferences could be different with more extreme behaviours. Future research should continue to investigate components of learning in those exposed to violence and vary the stimuli used to assess learning that consider cultural and situational contexts.

The relationship between exposure to violence and negative life experiences is undeniable. However, an understanding of how this environment shapes cognition and behaviour is less clear. The present study identifies a specific deficit in the ability of incarcerated individuals exposed to violence to adapt social behaviour towards agents with distinguishable harm preferences. Continuing to identify and specify the processes that are altered by exposure to violence will be crucial for understanding how individuals experience, incorporate, and react to their particular social environment.

Chapter 6

6 Moral inference in Borderline Personality Disorder

Borderline Personality Disorder (BPD) is a serious mental illness characterized by marked disturbances in interpersonal relationships, including difficulties with trust and forgiveness often resulting in the premature termination of relationships. Here, we tested a hypothesis that forming impressions about the moral character of others is disturbed in BPD. We predicted that BPD would be associated with slower updating of initially bad moral impressions, and that this deficit would be restored following treatment in a Democratic Therapeutic Community (DTC), which has shown some promise in ameliorating interpersonal disturbances in BPD. Participants predicted and observed the choices of two agents who repeatedly decided whether to inflict painful electric shocks on a victim in exchange for various amounts of money. The two agents differed substantially in their morality: the "good" agent required more compensation to inflict pain on others than the "bad" agent. Periodically, participants rated their subjective impressions of the agent's morality, and the certainty of those impressions. We used a hierarchical Bayesian learning model to describe participants' evolving beliefs about the character of the good

and bad agents. We show that the effect of BPD on Bayesian belief updating is intrinsically related to the morality of the agent. Relative to a sample of matched control participants who did not classify for a clinical diagnosis of BPD (N=106), patients with BPD (N=20) formed more certain beliefs about bad agents and more uncertain beliefs about good agents. Thus, beliefs about bad agents were slower to update in BPD patients, whereas beliefs about good agents were faster to update. We found that DTC treated BPD patients (N=23) were faster to update beliefs about bad agents relative to untreated patients. The results provide a mechanistic explanation for social deficits in BPD and demonstrate the potential for combining objective behavioural paradigms with computational modelling as a tool for assessing treatment outcomes.

6.1 INTRODUCTION

Borderline Personality Disorder (BPD) is a serious mental illness hypothesized to affect up to 2% of the general population (Torgersen, Kringlen, & Cramer, 2001). Marked disturbances in interpersonal relationships constitute one of the core symptom domains of BPD, causing tremendous suffering for patients and their social network. Difficulties related to interpersonal relationships have been identified as significant predictors of the most detrimental outcomes of BPD (Berk, Jeglic, Brown, Henriques, & Beck, 2007), contributing to substantial economic and societal costs including high rates of suicide and intensive use of high-cost medical care (American Psychiatric Association, 2001; Berk et al., 2007; Kjær, Biskin, Vestergaard, & Munk-Jørgensen, 2015; van Asselt, Dirksen, Arntz, & Severens, 2007). Longitudinal studies indicate that symptoms related to interpersonal relationships are among the hardest to treat; serious social deficits often persist even after years of rigorous and resource exhaustive treatment (Bateman & Fonagy, 2009; Giesen-

Bloo et al., 2006; Gunderson et al., 2011; Zanarini, Frankenburg, Reich, & Fitzmaurice, 2010). Research identifying the mechanisms of impairment in social functioning in BPD is therefore paramount for relieving interpersonal and societal burdens.

Building and maintaining successful social relationships depends on the ability to form accurate representations of others' mental states (e.g., intentions, beliefs, desires) (Frith & Frith, 2012). However, research indicates that individuals with BPD are negatively biased and hypervigilant in their socio-cognitive perceptions, often interpreting others' actions as threatening or hostile (Barnow et al., 2009; Fertuck et al., 2018; Fertuck, Grinband, & Stanley, 2013; Nicol, Pope, Sprengelmeyer, Young, & Hall, 2013; Preißler, Dziobek, Ritter, Heekeren, & Roepke, 2010; Unoka, Fogd, Füzy, & Csukly, 2011). Negatively biased perceptions have adverse consequences on the patients' social network, leading to relationships that appear volatile in nature. For example, social interactions in BPD are associated with a pattern of rapid shifting from periods of admiration to dislike of social partners (Bender & Skodol, 2007) and the termination of close relationships in response to even minor slights (Clifton et al., 2007).

A growing body of theoretical and empirical work suggests that social learning plays an important role in socio-cognitive and interpersonal disturbances in BPD (Fonagy, Luyten, Allison, & Campbell, 2017). However, empirical evidence identifying and specifying the underlying mechanisms of social learning in BPD is sparse. In a recent study (Fineberg et al., 2018), participants played the Social Valuation task (Behrens et al., 2008), an interactive game that requires participants to learn and update beliefs about the trustworthiness of a partner from social cues. The partner's trustworthiness varies across trials; thus, participants must track social cues and update beliefs about trustworthiness

accordingly. Surprisingly, participants with BPD exhibited blunted responses to social cues, relative to controls, as indicated by a slower learning rate. On the one hand, this finding appears inconsistent with clinical observations that patients shift rapidly from periods of admiration to dislike in response to even minor slights (Bender & Skodol, 2007), and well-documented empirical work showing *hypersensitivity* to social cues in BPD (Bertsch et al., 2013; Dyck et al., 2009; Frick et al., 2012; Koenigsberg et al., 2009). On the other hand, the finding is consistent with research showing that BPD patients are insensitive to social cues in a 10-round trust game, failing to coax a partner back into cooperation following a rupture of trust (King-Casas et al., 2008).

A possible explanation for the discrepancy in findings is that belief updating in the Social Valuation task was assessed using a computational model insensitive to dynamics of the learning environment. For example, in the Social Valuation task (Fineberg et al., 2018) a single learning rate captured participant behaviour across all trials. However, given that the trustworthiness of the partner in the Social Valuation task varied across trials, it is unclear whether and how learning may vary as a function of the *valence* of the belief at any given moment. Moreover, it is well documented that learning is intrinsically related to the *uncertainty* of beliefs (Behrens, Woolrich, Walton, & Rushworth, 2007; Mathys et al., 2011), where more uncertain beliefs are more rapidly updated. Thus, whether and how the uncertainty of beliefs shapes learning in BPD is unknown.

Subjective uncertainty of social judgments in BPD has been investigated in a small number of studies, however the findings are inconsistent. Schilling and colleagues (Schilling et al., 2012) found that BPD patients report greater confidence in their affective assessments in a theory of mind task wherein patients had to identify the emotional valence

of eyes. Conversely, Kaletch, Kruger, and colleagues (Kaletsch et al., 2014) observed that patients report less confidence in their evaluations of socially relevant body movements. Consistent with this finding, two studies found that BPD patients reported lower confidence in their evaluation of the affective valence of social cues in an emotion recognition task; Thome and colleagues (Thome et al., 2016) found that this was particularly pronounced for positive social cues while Niedtfeld found no effect of valence on confidence (Niedtfeld, 2017). A final study found no significant differences between BPD patients and healthy control participants in overall emotion recognition confidence, although BPD patients expressed greater confidence in their inaccurate recognition of sadness and surprise (Lowyck et al., 2016). Together, this work speaks to the complexity of social evaluations in BPD and suggests that the context of the evaluation may play a role in confidence judgments.

Here we apply our novel computational assay of moral inference to probe both the valence and uncertainty of momentary social beliefs. The previous chapters indicate that healthy adults hold more uncertain and less rigid beliefs about those for whom they infer a bad moral character relative to a good moral character. This is hypothesized to reflect an adaptive mechanism for sustaining relationships when others sometimes behave badly by allowing negative beliefs to be rapidly updated when presented with new information. Thus, the uncertainty of beliefs about morally bad agents may be an important aspect of healthy social functioning. Given that patients with BPD often hold grudges and present difficulty forgiving others (Sansone et al., 2013; Thielmann et al., 2014), we hypothesize that BPD patients may lack this adaptive mechanism for sustaining relationships and present more certain beliefs about more harmful agents.

Understanding the mechanisms underlying interpersonal problems in BPD is essential to the development and assessment of effective treatment. Democratic Therapeutic Community (DTC) treatment is one of the most widespread psychosocial treatments in the UK with a strong focus on developing cooperative strategies to help patients effectively navigate their social environment (Whiteley, 2004). DTC offers a therapeutic environment for patients with severe personality disorders to develop skills that support adaptive social functioning. Treatment usually lasts for a period of 18 months, where patients spend a minimum of 2-3 days each week in the facility. The treatment adopts a democratic, authoritarian-free therapeutic environment where patients are active participants in each other's treatment. There is no immediate, clear distinction between patients and staff. Patients and staff alike are held accountable for their actions by the community. All participants share responsibility for the daily running of the community; patients must interact and cooperate with each-other to accomplish daily tasks including grocery shopping, cleaning, cooking, and group activities.

DTC was associated with improved social functioning in a 24 month follow-up study (Pearce et al., 2017), as measured by the Social Functioning Questionnaire, an 8-item validated measure of social functioning (Tyrer et al., 2005). Moreover, one of the strongest outcomes reported by participants following DTC is more pleasant social relations (Debaere et al., 2016). Despite the strong focus on social learning, the mechanisms through which DTC treatment effects change, and whether treatment indeed shapes how patients learn from their social environment, are unknown. As a secondary goal, the present research aimed to address these questions by assessing moral inference in a group of BPD

patients following DTC treatment, and compare behaviour to a group of untreated BPD patients.

6.2 METHODS

Participants

Non-BPD group. For the control group, we aimed to collect a sample of adults from the United Kingdom using the online crowdsourcing application, Prolific (www.prolific.ac). Online crowdsourcing enabled us to collect a sample of control participants that were precisely matched to our patient population. This method has the potential to improve the validity and generalizability of research by providing a source for comparison groups for unique samples who may come from specific environments at a relatively low cost of resources and time (Azzam & Jacobson, 2013). To ensure that our sample of healthy adults was closely matched to our sample of BPD patients, we aimed to recruit five healthy adults who matched each patient in sex, age and education (+/- 4 years from the age of the patient). We also ensured that the matched participants received the same variant of the moral learning task (i.e., same sequence of trials).

Participants completed the study on the web application framework, Heroku, and were subsequently directed to a Qualtrics survey to complete additional questionnaires to assess clinically relevant personality traits. All control participants completed the McLean Screening Inventory (MSI, see *Additional Measures* below) for BPD and were excluded from the analysis if they tested positive for the presence of clinically relevant BPD (MSI score > 6). All participants provided written informed consent prior to participation and were compensated for their time. The Yale University Human Investigation Committee

approved the procedures used, ethics number 2000022385. The study complied with all relevant ethical regulations for work with human participants. The final sample of matched controls included 106 adults who did not classify for BPD.

BPD group. Participants were treatment-seeking patients with a primary diagnosis of BPD recruited from referrals to the Oxford Complex Needs Service. The Structured Clinical Interview for axis II disorders (SCID-II) was administered by trained clinicians to establish BPD diagnosis. Inclusion criteria were: diagnosis of BPD, aged between 18 and 65, not currently being treated in group therapy, no current drug or alcohol dependence, and no psychiatric hospital admission in the preceding month. Individuals were excluded if they had a previous or current neurological condition, were unable to provide informed consent, were pregnant or breastfeeding, or met criteria for an Axis I illness (e.g., anxiety, mood, eating disorders). Nine patients were taking antidepressant or antipsychotic medication or both at the time of participation. The study was approved by the local National Health Service ethics committee in Oxford, ethics number 14/SC/1430. The final sample include 20 patients with BPD.

DTC group. Patients with a primary diagnosis of BPD who completed DTC treatment within three years of participation were recruited from the Oxford Complex Needs Service database. Eligible patients were contacted by post and sent a copy of the information sheet along with an invitation to participate in the study. The Structured Clinical Interview for axis II disorders (SCID-II) was administered to interested individuals by trained clinicians to establish BPD diagnosis. Inclusion criteria were: diagnosis of BPD, aged between 18 and 65, completed DTC at the Oxford Complex Needs Service within the past three years, and no current drug or alcohol dependence. Individuals were excluded on

the same basis as patients in the BPD group. Eleven patients were taking antidepressant or antipsychotic medication or both at the time of participation. The study was approved by the local National Health Service ethics committee in Oxford, ethics number 14/SC/1430. The final sample included 23 patients with BPD who had completed DTC treatment.

Behavioural testing of patients (BPD and DTC groups) took place at the Oxford Complex Needs Service.

Experimental Procedure

We employed the Moral Inference Task as described in Chapter 2. Because non-BPD and patient groups completed the task under very different experimental settings (online versus the clinic), we wanted to confirm that the groups were equally motivated to learn about the agents and predict their decisions. Consequently, after predicting all the choices for a given agent, we explicitly asked participants to indicate on a continuous scale from 0 (*very unmotivated*) to 100 (*very motivated*) "How motivated to be accurate did you feel during the task?".

Additional Measures

Borderline evaluation of severity over time (BEST). We used the BEST (Pfohl et al., 2009) to assess the severity of BPD symptomology in patients with BPD at the time of participation. The BEST is a 15-item questionnaire which measures thoughts, emotions, and behaviours (positive and negative) typical of BPD. Positive behaviours were not measured in this study, and thus participants responded to only 12 of the 15 items. Each item asks participants to rate their experience with each of the items since their last clinical session; the lowest score of 1 means that it caused little or no problems, and the highest

score of 5 means that it caused extreme distress, severe difficulties with relationships, and/or kept them from completing tasks. The scores from the 12 items were added together to yield a score between 12 and 60, where higher scores indicated greater BPD severity.

Personality Inventory for DSM-5, brief form (PID-5-BF). We used the PID-5-BF (Krueger, Derringer, Markon, Watson, & Skodol, 2012), a 25-item self-report questionnaire, to assess clinically relevant personality traits that do not necessarily constitute a personality disorder. The PID-5-BF constitutes five personality trait domains: negative affect, detachment, antagonism, disinhibition, and psychoticism. Each item on the questionnaire asks participants to rate how well the item describes him or her generally on a scale from 0 (*very false or often false*) to 3 (*very true or often true*). The scores from all items were added together to produce a score between 0 and 75, with higher scores indicating greater general overall personality dysfunction.

McLean Screening Instrument for BPD (MSI). The MSI (Zanarini et al., 2003) was used as a screening measure for the presence of clinically relevant BPD in the control group. The validated instrument consists of ten true-false self-report questions to assess the occurrence of symptoms typically found in BPD, such as "*Have you deliberately hurt yourself physically (e.g. punched yourself, cut yourself, burned yourself)*". The screen is regarded as positive when seven or more of the symptoms are true.

General Trust Scale (Yamagishi & Yamagishi, 1994). The General Trust scale was used to measure general beliefs about the honesty and trustworthiness of others. Participants were asked to indicate to what extent they agree (*1*) or disagree (*5*) with six statements (e.g., 'Most people are trustworthy'). The scores from each statement were averaged together to produce a continuous measure of generalized trust.

Self Report Psychopathy - Revised, short form (SRP-R-SF). We used the SRP-R-SF (C. Neumann & Pardini, 2012), a 29-item self-report questionnaire, to assess psychopathic personality traits across patients and control participants. The instrument constitutes four factors of psychopathy: affective callousness, interpersonal manipulation, antisociality, and erratic lifestyle. Each item on the questionnaire asks participants to rate the extent to which they thought the item reflected their own beliefs using a 5-point likert scale (1 = *strongly disagree* to 5 = *strongly agree*).

Computational modelling

Three computational models were compared to describe how participants learned the agents' preferences and predicted their choices. We fit the HGF (Mathys et al., 2011, 2014), which identified participant-specific parameters to describe each individual participant's learning process. Beliefs about an agent's harm preference were updated using a Bayesian reinforcement learning algorithm, with precision-weighted prediction errors driving belief updating at the different levels of the hierarchical model. Second, we fit a Rescorla Wagner model, in which beliefs were updated by prediction errors with a fixed learning rate. Third, we fit a modified Rescorla Wagner model, in which beliefs were updated by prediction errors with separate fixed learning rates for helpful and harmful outcomes. For details about the alternative models, see **Table 2.2**. The log-model evidence (LME) indicated that the HGF model (sum LME = -7149) outperforms both a simple single learning rate RW model (sum LME = -7444) and a RW model with separate learning rates for positive and negative outcomes (sum LME = -7192). We validated these findings using formal Bayesian Model Selection. To this end, we used LME data to compare between the HGF and our two RW models. This analysis yielded a protected exceedance probability indistinguishable from 1

for the HGF model for both agents, indicating effectively a 100% probability that the HGF model better explains the data than the other models included in the comparison.

Analysis

For the present study we were interested in the computational processes during moral inference in BPD. Previous work suggests that BPD affects the stability of beliefs *over time* (Bender & Skodol, 2007). Due to the temporal emphasis of belief stability in BPD, we optimized our analysis to investigate the temporal dynamics of learning. The variance in the posterior belief (σ_2) reflects a dynamic learning rate which dictates trial-by-trial belief updating as a function of the precision (i.e., inverse uncertainty) of beliefs about the agent's moral preference (κ). σ_2 is monotonically related to the global estimate of belief volatility, ω. However, ω is a global estimate that captures individual differences in the influence of historical and newly arriving information on existing beliefs, while σ_2 captures the weight a prediction error has towards updating beliefs on the current trial. Thus, σ_2 has the sensitivity to capture fluctuations in updating over time, given individual differences in ω. We used robust linear regression models with a bisquare weighting function to analyze the model estimated dynamic learning rates, character ratings, and uncertainty ratings. Further analyses used nonparametric statistical tests that do not make any assumptions about the underlying distributions of variables (e.g., Spearman's ρ and signed rank tests).

6.3 RESULTS

Moral inference in BPD

In an initial series of analyses, we investigated whether moral inference in treatment-seeking BPD patients differed from a sample of non-clinically diagnosed adults matched

for sex, age, and education (i.e., the non-BPD group; **Table 6.1**). As expected, patients with BPD scored significantly higher on the Personality Inventory for DSM-V (Z = -4.997, p < .001; Error! Not a valid bookmark self-reference.) indicating significantly higher levels of clinically relevant personality traits than non-BPD participants. The BPD and non-BPD group demonstrated similar levels of trait psychopathy (Z = 1.437, p = .151).

On average, participants predicted accurately 73% of the good agent's choices and 79% of the bad agent's choices. The BPD and non-BPD groups were similarly accurate in their predictions of the good and bad agent (% accuracy: bad agent: Z = -1.103, p = 0.270; good agent: Z = 0.295, p = 0.768) and reported similar levels of motivation to be accurate in their predictions (motivation rating: bad agent: Z = - 0.879, p = 0. 379; good agent: Z = -1.704, p = 0.088). This suggests that participants with BPD were equally motivated to learn the harm preferences of the agents, relative to those without BPD, despite completing the task under different experimental settings (online vs. the clinic).

Table 6.1

Participant demographic information, BPD vs. non-BPD. SEM = Standard error of the mean.

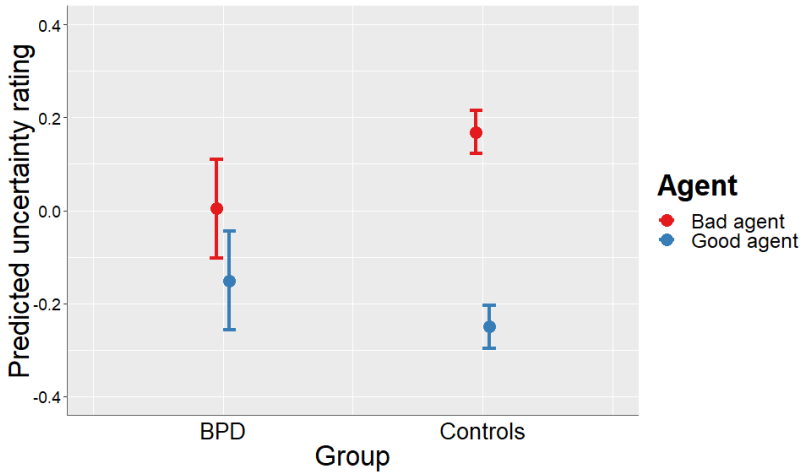
	BPD group (N=20)		**Non-BPD group (N=106)**			
	Mean	**SEM**	**Mean**	**SEM**	***Z*-stat**	***p*-value**
Age on date of participation	39.500	2.561	40.957	1.140	-0.612	0.540
Highest level of education	2.412	0.195	2.587	0.094	-0.861	0.389
General trust score	18.600	2.021	30.245	0.678	-4.997	<0.001
Personality inventory for DSM-V	39.950	3.042	18.740	1.202	5.269	<0.001
Psychopathy	42.053	2.024	38.387	0.795	1.437	0.151
Prior moral expectations	39.300	5.901	54.118	2.550	-2.269	0.023
Prior uncertainty rating	62.800	5.614	62.434	2.134	-0.327	0.743

To examine subjective moral impressions and certainty ratings, we standardized each dependent variable and entered them into separate robust linear regression models with agent (bad vs. good) as a within subject variable and group (BPD vs. non-BPD) as a between subject variable. We included an additional regressor to control for rating number. Certainty ratings were reverse scored such that higher values indicated greater uncertainty. Examining subjective impression ratings revealed that participants formed more negative impressions about the 'bad' agent than the 'good' agent (mean±SEM, $\beta = -1.178 \pm 0.027$, $t = -44.299$, $p < .001$). The main effect of group and the interaction between agent and group were not significant. Thus, the valence of moral impressions did not vary as a function of BPD diagnosis.

Subjective uncertainty ratings revealed a different pattern of results. Consistent with previous findings, impressions of the bad agent were more uncertain than impressions of the good agent ($\beta = 0.418 \pm 0.032$, $t = 13.099$, $p < .001$). The BPD group was marginally more uncertain in their impressions overall ($\beta = 0.098 \pm 0.057$, $t = 1.738$, $p = .082$) and this was qualified by a significant interaction between agent and group ($\beta = -0.263 \pm 0.080$, $t = -3.284$, $p = .001$; **Figure 6.1**). To examine the interaction, we fit the model to uncertainty ratings of the bad and good agents separately. This revealed that BPD was associated with *less* uncertain impressions of the bad agent ($\beta = -0.162 \pm 0.058$, $t = -2.805$, $p = .005$), and marginally *more* uncertain impressions of the good agent ($\beta = 0.098 \pm 0.055$, $t = 1.761$, $p = .078$), relative to the non-BPD group.

Figure 6.1 Predicted effect of BPD and agent on uncertainty ratings

BPD predicted less uncertain impressions of the bad agent and more uncertain impressions of the good agent. Error bars represent 95% confidence intervals.

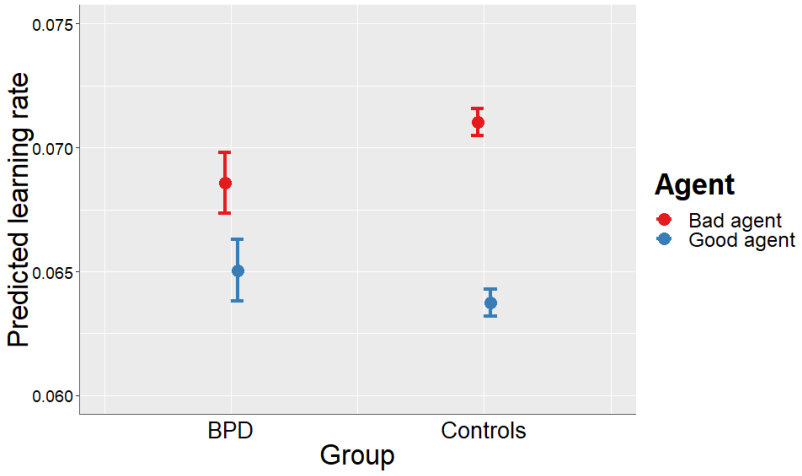


In Bayesian updating uncertainty informs the rate at which new evidence is incorporated into beliefs; greater uncertainty predicts that new evidence will be more readily incorporated into existing beliefs (Behrens et al., 2007; Mathys et al., 2011). Consequently, we entered trial-wise dynamic learning rates from our Bayesian computational model into a robust linear regression model with agent and group as independent regressors to examine whether the pattern observed for subjective uncertainty ratings was reflected in belief updating. We included an additional regressor to control for trial number. Replicating findings from previous chapters, beliefs were updated at a faster rate for the bad agent relative to the good agent, as indicated by a significant main effect of agent on dynamic learning rates ($\beta = 0.322 \pm 0.017$, $t = 18.601$, $p < .001$). We also found that the BPD group updated beliefs marginally faster than the control group ($\beta = 0.058 \pm 0.031$, $t = 1.880$, $p = .060$), consistent with subjective uncertainty ratings. Main effects were qualified by a significant interaction between agent and group (β = -

0.167±0.044, t = -3.827, p < .001; **Figure 6.2**). Fitting the model to dynamic learning rate data for good and bad agents separately revealed that the BPD group was slower to update beliefs about the bad agent (β = -0.109±0.034, t = -3.222, p = .001) and faster to update beliefs about the good agent (β = 0.062±0.027, t = 2.287, p = .022) relative to the control group. Together, these findings suggest that BPD patients have more confident and more rigid beliefs about putatively 'bad' agents, but less confident and more flexible beliefs about putatively 'good' agents.

Figure 6.2 Predicted interaction effect on learning rate

Effect of agent (bad versus good) and group (BPD versus controls) on dynamic learning rates. Error bars represent 95% confidence intervals.



A plausible explanation for the observed pattern of results is that the BPD group *expected* agents to have more harmful preferences, thus rendering the bad agent's behaviour less surprising than the good agent. In other words, the bad agent's behaviour would be more consistent with patient's prior expectation (and therefore increase confidence and rigidity of posterior beliefs) while the good agent's behaviour would be

less consistent with patient's prior expectations (thus, decrease confidence and rigidity of posterior beliefs). Negative expectations and low initial trust are common vulnerabilities in BPD (Barnow et al., 2009; Critchfield, Levy, Clarkin, & Kernberg, 2008; Ebert et al., 2013; Fertuck et al., 2018, 2013; King-Casas et al., 2008). In the present study, we found a similar vulnerability towards negative expectations. Participants with BPD indicated more harmful expectations before observing any of the agents' choices than non-BPD participants (Z = -2.491, p = .013; **Table 6.1**) but groups were similarly certain about their expectations (Z = -0.327, p = .743). Participants with BPD were also less trusting of others in general (as indicated by lower scores on the General Trust scale; Z = -4.810, p < .001).

While prior expectations are likely to shape the dynamics of learning, controlling for these variables in the present study yielded a similar pattern of results (group X agent interaction controlling for prior expectations, uncertainty: β = -0.261±0.080, t = -3.265, p = .001; learning rate: β = -0.004±0.001, t = -3.692, p < .001). To further explore the possibility that negative prior expectations explain differences between the BPD and non-BPD group, we fixed prior beliefs in the Bayesian reinforcement learning model using participants' prior character expectations measured at the start of the learning task. Again, the relationship between BPD and learning rates was robust (β = -0.008±0.002, t = -4.232, p < .001).

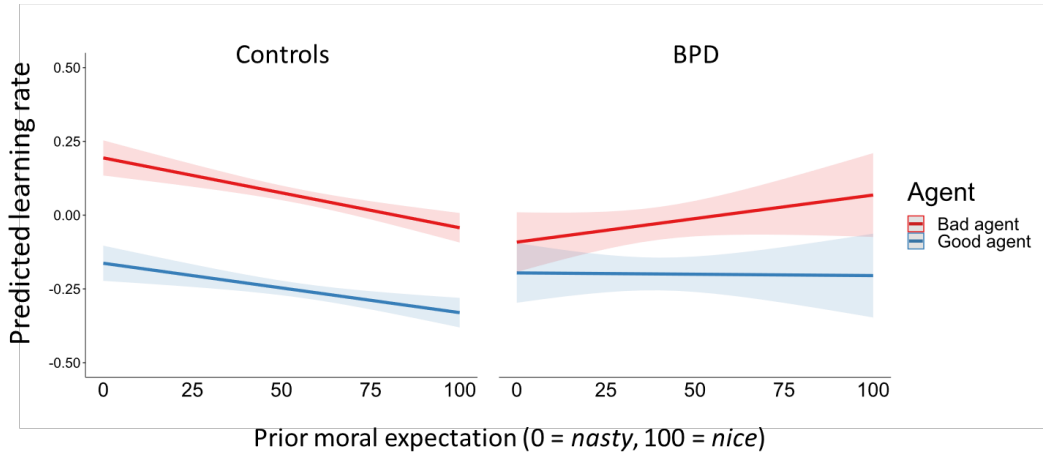
Although we observed similar effects controlling for prior expectations, it's possible that prior expectations interact with between group differences in Bayesian updating in systematic ways. Chapter 4 revealed that prior moral expectations are unlikely to account for asymmetric learning about good and bad agents in healthy adults. We hypothesize that humans have evolved to adapt learning according to contextual information, such as

perceived threat, to aid survival. In turn, this adaptive mechanism may enable healthy adults to discount expectations to build richer models of agents when harmful outcomes are expected (i.e., in response to negative moral expectations). One possibility is that patients with BPD lack the mechanism for adapting learning according to moral information. That is, while healthy adults may be able to override prior expectations and rapidly adjust their learning for putatively bad agents, this adaptive mechanism may be absent in BPD. As a result, learning may be more sensitive to prior expectations in BPD. If this is the case, we would expect learning in BPD to be more strongly influenced by prior moral expectations than learning in non-BPD participants.

In line with this prediction, we found a significant three-way interaction between prior expectations, BPD diagnosis, and agent ($\beta = 0.004 \pm 0.002$, $t = 2.214$, $p = .027$). To unpack the interaction, we performed a similar regression splitting the data as a function of BPD diagnosis. Consistent with findings from Chapter 3, prior expectations were not associated with differences in learning rates between the good and bad agent in the control group ($\beta = -0.001 \pm 0.001$, $t = -1.454$, $p = .146$, **Figure 6.3**). Conversely, prior expectations predicted asymmetric learning rates for good and bad agents in the BPD group: more pessimistic expectations were associated with a smaller learning asymmetry ($\beta = 0.003 \pm 0.001$, $t = 2.250$, $p = .025$, **Figure 6.3**). This suggests that the mechanisms underlying the ability to rapidly adapt learning towards moral information in healthy adults may be absent in BPD.

Figure 6.3 Prior moral expectations moderate belief updating in BPD

Effect of prior moral expectations on estimated learning rates for the control (i.e., non-BPD) group (left) and BPD group (right). Prior moral expectations were measured on a continuous scale before observing any of the agent's choices. The scale asked participants to indicate how nasty or nice they expected the agents would be in the task. Error bands represent 95% confidence intervals.



The impact of Therapeutic Community (TC) treatment on BPD

As a secondary goal, we aimed to assess the effect of DTC treatment on moral inference in BPD. To this end, we recruited a sample of BPD patients who had completed DTC treatment and were matched in sex, age, and education to the BPD group that had not undergone DTC treatment (**Table 6.2**). We confirmed that the severity of BPD symptomology in DTC treated patients was significantly lower than untreated BPD patients, as indicated by significantly lower scores on the BEST (BPD = 41.444±1.975; DTC: 26.867±1.956; $Z = 3.690, p < .001$). Meanwhile, BPD and DTC patient groups scored similarly on measures of psychopathy (BPD = 42.053 ± 2.024; DTC: 40.217 ± 2.628; $Z = 0.999$, $p = .318$) and personality disorder (BPD = 39.950 ± 3.042; DTC: 33.478 ± 3.029; $Z = 1.572$, $p = .116$).

Table 6.2

Participant demographic information, untreated vs. treated BPD. SEM = Standard error of the mean.

	BPD (N=20)		DTC (N=23)			
	Mean	**SEM**	**Mean**	**SEM**	***Z*-stat**	**p-value**
Age on date of participation	39.500	2.561	41.609	2.205	-0.573	0.567
Highest level of education	2.412	0.195	2.632	0.211	-0.748	0.455
General trust score	18.600	2.021	24.652	1.321	-2.244	0.025
Personality inventory for DSM-V	39.950	3.042	33.478	3.029	1.572	0.116
Psychopathy	42.053	2.024	40.217	2.628	0.999	0.318
Prior moral expectations	39.300	5.901	43.522	4.963	-0.585	0.559
Borderline evaluation of severity	41.444	1.975	26.867	1.956	3.690	<0.001

The treated and untreated BPD groups were similarly accurate in their predictions of the agents' behaviour (% accuracy: bad agent: BPD = 0.778±0.019 , DTC = 0.779±0.014, Z = 0.930, p = .352; good agent: BPD = 0.729±0.028, DTC = 0.774±0.029, Z = 0.184, p = .854), and indicated similar motivations to be accurate on a post-task rating (motivation rating: bad agent: Z = - 1.681, p = . 093; good agent: Z = -0.926, p = .354). This suggests that treated and untreated BPD groups were similarly motivated to learn the agents' harm preferences.

A robust linear regression with agent (bad vs. good) and group (BPD vs. DTC) indicated that subjective impressions of the bad agent were significantly worse than the good agent (β = -1.618±0.044, t = -36.832, p < .001). The DTC treated group expressed more favourable subjective impressions than patients in the untreated BPD group (β = 0.146±0.046, t = 3.197, p = .001). Main effects were qualified by a significant interaction (β = -0.236±0.064, t = -3.668, p < .001). The interaction indicated that the DTC treated group expressed more favourable subjective impressions of the good agent (β =

0.151±0.043, t = 3.507, p < .001), and marginally worse impressions of the bad agent (β = -0.090±0.048, t = -1.869, p = .062).

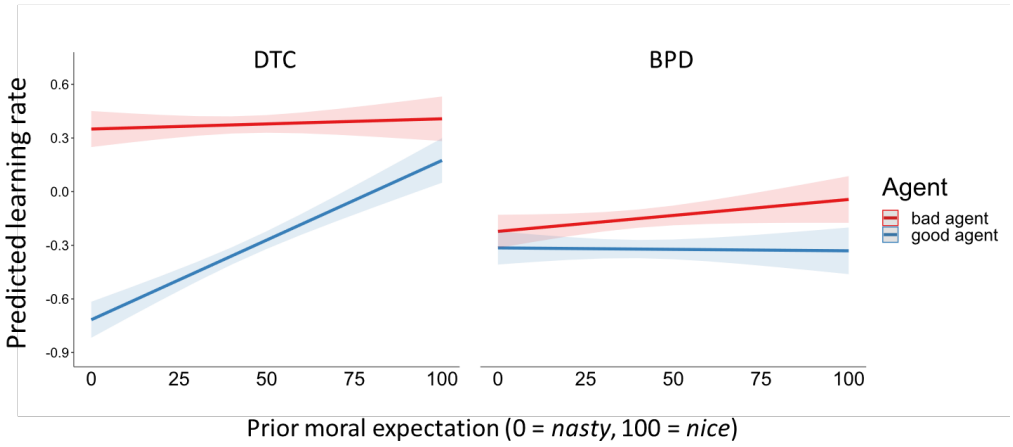
In our previous analysis we found that, relative to the non-BPD control group, BPD patients expressed *less* uncertain impressions about the bad agent, but *more* uncertain impressions about the good agent. Here, we observed a significant interaction between agent and group on uncertainty ratings (β = 0.277±0.095, t = 2.904, p = .003). DTC treatment was associated with more uncertain impressions about the bad agent (β = 0.188±0.067, t = 2.802, p = .005) while treatment did not affect uncertainty of impressions about the good agent (β = -0.081±0.068, t = -1.196, p = .232). Model estimated dynamic learning rates revealed a similar interaction between agent and group (β = 0.589±0.052, t = 11.588, p < .001). Beliefs about the bad agent updated at a faster rate in the DTC treated group than the untreated BPD group (β = 0.543±0.040, t = 13.698, p < .001). No effect of DTC treatment on dynamic learning rates was observed for the good agent (β = -0.030±0.030, t = -0.989, p = .323). Thus, treatment was associated with increased uncertainty and more flexible beliefs about the bad agent, specifically.

Patients in the DTC treated and untreated BPD group had similar expectations about the agents' morality (BPD = 39.300±5.901, DTC = 43.521±4.963, Z = 0.585, p = .559; **Table 6.2**). Nevertheless, we observed a significant interaction between prior expectations, agent, and group on learning rates (β = -0.010±0.002, t = -4.752, p < .001; Figure 6.4). To unpack the interaction, we fit the regression model separately for untreated BPD and DTC treated groups. Again, we found that worse expectations were associated with a smaller updating asymmetry between agents in the BPD group (β = 0.003±0.001, t = 2.250, p = .025). However, the opposite pattern was observed for the DTC treated group: worse

expectations were associated with a larger updating asymmetry between agents (β = -0.007±0.002, t = -4.615, p < .001). These findings suggest that even though DTC and BPD groups had similar moral expectations, the groups differed in how expectations subsequently shaped learning.

Figure 6.4 Prior moral expectations moderate belief updating

Effect of prior moral expectations on estimated learning rates for the DTC group (left) and BPD group (right). Prior moral expectations were measured on a continuous scale before observing any of the agent's choices. The scale asked participants to indicate how nasty or nice they expected the agents would be in the task. Error bands represent 95% confidence intervals.



6.4 DISCUSSION

Chapter 3 and Chapter 4 indicated that healthy adults hold less rigid beliefs about those who they infer a bad moral character relative to a good moral character. This is hypothesized to reflect an adaptive mechanism for sustaining relationships when others sometime behave badly. In BPD, relationships are characterized as intense and unstable (American Psychiatric Association, 2013). Relative to healthy adults, BPD patients' social networks have a greater number of relationships that are terminated (Clifton et al., 2007), they often hold grudges and present difficulty forgiving others (Sansone et al., 2013;

Thielmann et al., 2014). This suggests that patients with BPD may lack an adaptive mechanism for maintaining relationships.

Using a precise measure of moral inference, we provide evidence that the effect of BPD on social learning is intrinsically related to the *valence* of the belief, which may explain paradoxical findings in the literature. The finding of more rigid negative beliefs in BPD is consistent with patients exhibiting less coaxing following a rupture of trust (King-Casas et al., 2008), slower learning rates in the Social Valuation task (Fineberg et al., 2018), and difficulty forgiving others (Thielmann et al., 2014). Conversely, more flexible positive beliefs in BPD may explain the ease patients have in terminating relationships or clinical observations that the patient can shift rapidly from a period of admiration to dislike in response to even minor slights (Bender & Skodol, 2007).

The research on the development of BPD may help explain these findings. An extensive history of negative social experiences is all too common for patients with BPD. Patients often describe a childhood mired in abuse, neglect and violence, with sexual abuse by a close relative (before the age of 19) consistently reported more often in BPD compared to depressed or personality-disordered controls (Zanarini, 2000). Such adversities are likely to coin an individual's expectations about other's moral behaviours and trustworthiness. This is evidence by significantly lower monetary units entrusted by patients with BPD in trust games relative to controls (King-Casas et al., 2008; Unoka et al., 2009). Here, we also found evidence supporting a negative bias in patients' social expectations: BPD patients reported significantly worse expectations about the agents' harm preferences and scored lower on a measure of general trust than healthy control participants.

In optimal Bayesian inference, learning is intricately tied to prior expectations about the environment. Observations that are consistent with prior expectations help reinforce them, while those that are inconsistent may be used to update expectations. However the previous chapters indicate that moral inference departs from Bayes optimality in an important way: typical human observers maintain more uncertain beliefs about the character of bad agents even when observations are consistent with prior expectations. This provides the flexibility to update beliefs about bad agents when they turn out to be wrong, preserving social relationships in the wake of accidental harms. While this mechanism may be advantageous in social environments generally characterized by stable moral preferences, it can be costly when social cues are unreliable predictors of future behaviour. The development of blunted belief updating when observing bad agents in BPD may therefore function as an *adaptive* response to adversity and volatility in the patient's social environment. That is, if those closest to an individual abruptly flip between loving kindness and violence, flexible updating of beliefs may be maladaptive because social cues are unreliable predictors of future behaviour. By shutting down the gateway for learning when behaviour aligns with negative expectations, rigidity then provides a protective mechanism that prevents responding to unreliable social cues.

At the same time, two explanations, that are not necessarily mutually exclusive, may account for the finding that positive beliefs are more uncertain and flexible in BPD than non-BPD control participants. First, patients have more robust negative moral expectations, and therefore maintain greater uncertainty when behaviour misaligns with expectations in a Bayes optimal sense. While this remains a theoretical possibility, BPD patients and non-BPD control participants reported similar levels of confidence in their

prior expectations suggesting that, at the very least, such differences did not reach metacognitive awareness. Second, patients may infer harmful intentions (i.e., social threat) even when the agent's actions do not necessarily warrant this interpretation. Increased sensitivity to social cues (Frick et al., 2012) and a higher propensity to misattribute social information with malevolent motives (Barnow et al., 2009) are common features of BPD. Such hypervigilance may impact the ability for patients to trust positive behaviours at face value. Thus, the pattern of learning and belief updating in BPD may have developed as a computationally optimal way of responding to *perceived* threats, in a similar manner to how healthy adults respond to agents who are threatening at face value.

To summarize, the findings suggest that BPD may impact the development of cognitive processes important for the ability to adapt learning as a function of moral information. In turn, patients rely heavily on prior expectations manifested from mistrust born from adversity in the individual's social environment. Exercising caution in response to moral actions that misalign with antisocial expectations may be an adaptive response to common interactions in the patient's environment but maladaptive outside them. Therefore, patients must learn when and how to adjust their learning in different social contexts to maintain healthy relationships.

DTC offers a safe environment for patients with BPD to learn the skills necessary for successful social functioning and has shown some promise in ameliorating social difficulties. Consequently, we applied the Moral Inference Task to investigate the mechanisms through which DTC effects change by comparing a group of patients with BPD who had completed DTC treatment to a comparable group of untreated patients. DTC was associated with more uncertain negative beliefs, and consequently faster updating from

social cues. This suggests that DTC may positively impact social interactions by increasing patients' openness to learning from their social environment, allowing information to be integrated over longer timescales before establishing a negative evaluation. No differences were found between BPD and DTC groups when learning about the putatively good agent.

The findings suggest that DTC treatment may be especially focused on the development of social skills for adverse social interactions. Indeed, a key component of the DTC environment is that patients cannot "escape" or "write off" other members of the community. Instead, patients are forced to recognize that at times, they must cooperate with people they are not fond of and learn to forgive those who mistakenly, or even intentionally, betray them. However, the social environment in DTC is biased towards individuals with personality disorders, whose behaviour is characterized by volatile, aggressive tendencies. Thus, it is unclear whether patients are significantly exposed to *positive* interactions partners in DTC, which is essential for the development of stable positive beliefs.

The present study did not assess traumatic experiences in childhood, such as sexual and physical abuse. Hence, it's unclear whether differences in prior expectations between patients with BPD and non-BPD control participants can be attributed to childhood experiences. More work is needed to assess whether there exists a relationship between childhood trauma, negative moral expectations and social learning. It's possible that the development of rigid beliefs about harmful agents is causally related to victims of childhood trauma, and therefore may generalized to victims who do not necessarily express the set of symptoms for a clinical diagnosis of BPD.

Before concluding, several other methodological and conceptual limitations should be addressed. In the present study many patients in the DTC treated group were taking

psychotropic medication during the time of participation. It's possible that these pharmacological treatments drove increased flexibility and belief updating for the harmful agent, rather than DTC treatment. However, a similar number of participants in the treated and untreated BPD group were taking psychotropic medication (DTC treated N = 11, untreated BPD N = 9), which limits the likelihood that medication use fully accounted for the observed differences in DCT treatment. We also observe a similar pattern of results when controlling for medication use (see **Appendix I:** and **Appendix L:**). Nonetheless, future work should assess the effects of DCT on social learning in a sample of BPD patients free from psychotropic medication.

6.5 CONCLUSIONS

We provide a unifying framework for understanding both volatility and rigidity in BPD. By combining computational modelling and a novel moral inference paradigm, we show that BPD is associated with more rigid negative beliefs and more volatility positive beliefs, which may be attributed to greater reliance on prior expectations in BPD. Our findings suggest that DTC may shape social interactions in BPD by increasing patients' openness to learning about adverse social interaction partners.

The set of methods presented here captures the richness in the pathology of BPD and may have significant clinical implications. Currently, there exists no assessments for BPD that does not rely on clinician or patient subjective reports. Subjective diagnoses have a tendency to vary from one clinician to the next, consequently the development of more objective measures is needed to provide increased accuracy of clinical diagnosis and better assess the effectiveness of treatment. Computational modeling of social learning dynamics

may prove a useful tool for investigating longitudinally how aspects of learning and impression updating might predict the course of treatment. Undoubtably, there also exists individual differences in the underlying mechanisms of social dysfunction in BPD. Using the tools presented here, we may be better equipped to identify these individual differences and develop more personalized treatment that target causal mechanisms on an individual basis.

Chapter 7

7 GENERAL DISCUSSION

7.1 SUMMARY OF EXPERIMENTAL FINDINGS

7.1.1 PART I: THE COMPUTATIONAL MECHANISMS OF MORAL INFERENCE

Part I of this dissertation examined the cognitive mechanisms of moral inference in humans. A large body of work in social psychology suggests that negative moral impressions are especially difficult to change because inaccurately attributing good character to bad people increases one's chances of being harmed (Baumeister et al., 2001; Skowronski & Carlston, 1989). However, the tendency to rigidly form negative moral impressions can be costly as well: attributing bad character from rare and minor transgressions leaves little room for people to make mistakes and can impede the development of stable relationships necessary for healthy social functioning. Responding to immoral acts with leniency and forgiveness may enable the development of successful relationships despite occasional harms. Although evolutionary and economic models provide descriptive accounts of these behaviours (Fudenberg et al., 2012; Grim, 1995; Nowak & Sigmund, 1992; Rand et al., 2009), the cognitive mechanisms that enable them are not well understood. Advances in non-social perceptual learning may hold a clue. This

work suggests that a pupil-linked arousal system may encode uncertainty signals that facilitate learning from future outcomes (Allen et al., 2016; Nassar et al., 2012). If this mechanism extend to learning about socially threatening others, who are also believed to evoke arousal, this may enable negative moral impressions to be flexibly updated from new information (Öhman, 1986). The experiments described in Chapters 3-4 set out to examine the cognitive mechanisms of moral inference to elucidate the processes that support adaptive functioning in social relationships.

Chapter 3 employed a novel behavioural task designed to assess two separate measures of moral inference: the ability to learn and predict others' objective moral preferences and the formation of explicit moral character impressions. In the Moral Inference Task, participants predicted whether agents chose to profit from harming an anonymous stranger. Intermittently participants rated their impression of the agent's moral character and how confident they were about their impression. One agent (the 'good' agent) required significantly more money to harm the stranger than the other (the 'bad' agent). No information about the agents was provided to participants prior to observing their choices. Therefore, to optimally predict the agents' choices participants had to gather information across trials to infer the agent's exchange rate between money and harm. A Bayesian reinforcement learning model was fit to participants' trial-wise predictions to capture the influence of historical information and newly arriving information on existing beliefs about the agents' moral preferences.

Moral inference was described by an asymmetric Bayesian updating mechanisms where beliefs about the bad agent were more volatile than beliefs about the good agent, relying more heavily on new over historical information (Study 1-3). This is consistent

with the finding that prior information signaling good moral character blunts caudate prediction error signals in repeated interactions to a greater degree than prior information signaling bad moral character (Delgado et al., 2005). Participants also expressed greater subjective uncertainty about their impressions of the bad relative to the good agent, suggesting that meta-cognitive processes play a role in moral inference. Predicting sports performance did not elicit a similar asymmetry when learning about agents who significantly differed in their skill level in the absence of moral information (Study 4). However, beliefs about skill-level were more uncertain and volatile when the agent also presented signs of immorality (Study 5), suggesting that properties of an agent's moral character influence social inference more generally. The vulnerability of negative moral beliefs to new information, over historical information, may contribute to the ability to forgive others for past harms (Study 6).

The findings from Chapter 3 raise two plausible hypotheses for why beliefs about bad agents are especially uncertain and volatile. One hypothesis is that the bad agents violated participants' expectations to a greater degree than the good agents. This makes beliefs about the bad agents more amenable to Bayesian updating, by which belief updates are optimized to minimize surprise (Mathys et al., 2011). An alternative hypothesis is that participants were especially motivated to learn about socially threatening agents. Because both threat and unexpected outcomes evoke arousal (Allen et al., 2016; Nassar et al., 2012; Storbeck & Clore, 2008) either one may drive uncertainty signals to facilitate learning about putatively bad agents.

An 'optimistic' prior hypothesis predicts that asymmetric updating will be related to the extent to which participants believe that others are good and trustworthy. Additionally,

if in the absence of additional information people expect others to have preferences similar to their own (Brañas-Garza et al., 2017; Hsee & Weber, 1997; Yamagishi et al., 2013), then participants who are more averse to harming a stranger should show a larger asymmetry in updating beliefs about good and bad agents. Each of these predictions were tested in Chapter 4 with no supporting evidence found (Section 4.2, page 116). Incentivizing a separate group of participants to predict, in the context of decisions to profit from others' pain, how "most people" would choose also found no evidence for optimistic expectations (Section 4.3, page 122).

The final study described in Chapter 4 further explored the role that moral expectations play in moral inference (Section 4.4, page 124). The study was designed to arbitrate between alternative explanations for asymmetric updating by dissociating moral expectations using facial cues (high versus low threat) from the agent's morality (bad versus good). The data revealed that beliefs about bad agents were more volatile and uncertain than beliefs about good agents, and these effects did not covary with prior moral expectations. Consistent with this, Delgado et al. (Delgado et al., 2005) found that the magnitude of prediction error signals were similar when agents' behaviour was consistent versus inconsistent with prior expectations of morality. Together, the results support a theoretical model for moral inference whereby moral expectations are rapidly discounted to enable beliefs to be more or less flexibly updated on a person-by-person basis.

7.1.2 PART 2: MALADAPTIVE MORAL INFERENCE AND SOCIAL INTERACTIONS

Any discussion about the mechanisms for adaptive social inference would be incomplete without considering the potential consequences that ensue when these

mechanisms break down. Inappropriate learning and decision-making lie at the heart of many populations characterized by maladaptive social functioning. Computational modelling may be particularly useful for understanding the mechanisms underlying disfunction. Consequently, Part II of this dissertation applied the Moral Inference Task to examine learning in two populations associated with abnormal interpersonal functioning: individuals exposed to violence and patients with BPD.

Exposure to community violence is a prominent risk factor for aggression, antisocial behaviour and incarceration. How and why exposure to violence leads to maladaptive social behaviours is not well understood. Perhaps one of the most fundamental ingredients for healthy social functioning is learning who is good and trustworthy versus who is not. Consequently, Chapter 5 applied the Moral Inference Task to investigate how exposure to violence affects the ability to learn others' morality and use this information to adaptively modulate trust behaviour in a sample of incarcerated males. The data suggest that exposure to violence adversely impacted some components of moral inference, but not all. Although all participants were able to learn the agents' moral preferences and predict their decisions, only those with little exposure to violence were able to use this information to make adaptive trust decisions, placing more trust in the good agent than the bad. Meanwhile, those with the highest exposure to violence trusted the good and bad agents equally. In particular, participants with high exposure to violence extended less trust than optimal when interacting with the good agent, which resulted in missed opportunities.

The relationship between exposure to violence and maladaptive trusting behaviour was mediated by disturbances in impression formation. Consistent with evidence that violence normalizes beliefs about harm (Ng-Mak et al., 2002), exposure to violence

predicted more lenient impressions of the bad agent. Meanwhile, those with the highest exposure made harsher evaluations of the good agent, consistent with evidence that individuals who have experienced violence themselves interpret the behaviour of neutral actors as hostile (Dodge et al., 1990). Disturbances in impression formation and trust behaviour in turn predicted real antisocial behaviour in prison. The findings suggest that chronic exposure to violence leaves resounding effects on the ability to use other's actions to distinguish those we should avoid from those we should befriend. In turn, this may lead individuals to respond inappropriately in social interactions.

BPD is a serious mental illness similarly characterized by marked disturbances in interpersonal relationships, with symptoms often persisting after years of resource intensive treatment. Part I of this dissertation indicates that healthy adults hold more flexible beliefs about those who they infer a bad moral character relative to a good moral character. This is hypothesized to reflect an adaptive mechanism for sustaining relationships when others sometime behave badly. In BPD, relationships are characterized as intense and unstable (American Psychiatric Association, 2013). Relative to healthy adults, BPD patients' social networks have a greater number of relationships that are terminated (Clifton et al., 2007), they often hold grudges and present difficulty forgiving others (Sansone et al., 2013; Thielmann et al., 2014). This suggests that patients with BPD may lack an adaptive mechanism for maintaining relationships. Consequently, Chapter 6 set out to test the hypothesis that BPD would be associated with slower updating of initially bad moral impressions, and that this deficit would be restored following treatment in a Democratic Therapeutic Community (DTC), which has shown some promise in ameliorating interpersonal disturbances in BPD.

In line with this prediction, the results indicate that the effects of BPD on moral inference is intrinsically related to the morality of the agent. BPD was associated with less uncertain and flexible beliefs about putatively bad agents, relative to a sample of matched healthy control participants. This may explain findings that patients exhibit less coaxing following a rupture of trust (King-Casas et al., 2008), slower learning rates when inferring other's trustworthiness (Fineberg et al., 2018), and difficulty forgiving others (Thielmann et al., 2014). Conversely, BPD was associated with more uncertain and flexible beliefs about putatively good agents. This may explain the ease patients have in terminating relationships or clinical observations that social evaluations shift rapidly from a period of admiration to dislike in the wake of minor slights (Bender & Skodol, 2007). The effects of DTC treatment were also related to the morality of the agent. While there were no effects of DTC on moral inference for good agents, DTC-treated BPD patients were faster to update beliefs about bad agents relative to untreated patients. The results provide a mechanistic explanation for social deficits in BPD and suggest that DTC treatment may shape social functioning by increasing patients' openness to learning about adverse social interaction partners.

7.2 SYNTHESIS OF EXPERIMENTAL FINDINGS

7.2.1 NEURAL MECHANISMS UNDERLYING ASYMMETRIC BELIEF UPDATING

Through what biological mechanisms might inferences about moral character bias human learning? Psychologically, reactions to threatening cues and subsequent regulatory responses are often recognized as bottom-up and top-down processes, respectively. For example, initial reactions to a villian in a film (i.e. bottom-up saliency) are implicitly

controlled by the realization that this person presents no immediate danger because they are merely acting (top-down regulation). Studies have highlighted a role for the amygdala and medial prefrontal cortex (mPFC) in the competition between bottom-up and top-down processes, where mPFC is believed to regulate amygdala output and subsequent behavioural responses via inhibitory control (Ochsner & Gross, 2005; Ochsner et al., 2009; Quirk & Beer, 2006; Quirk, Likhtik, Pelletier, & Paré, 2003). Outputs of the amygdala include direct projections to the hypothalamus and brainstem, as well as projections to all major neuromodulatory systems to globally affect cortical excitability (Kapp, Supple, & Whalen, 1994; Krettek & Price, 1978). These latter projections are believed to induce a state of heightened vigilance rendering prior beliefs more susceptible to environmental inputs (i.e., increase the learning rate).

Traditionally, the amygdala is thought to evaluate stimuli along a general valence dimension ranging from aversive to appetitive (Irwin et al., 1996; Morris et al., 1996; P J Whalen, 1998; Stillman, Van Bavel, & Cunningham, 2015). In the context of social impression formation, this has been realized in the perceived trustworthiness of others, where the amygdala is progressively sensitive to increasingly untrustworthy faces (Baron et al., 2011; Engell et al., 2007; Winston, Strange, O'Doherty, & Dolan, 2002). While this suggests that the amygdala signals social information that is potentially threatening, work mapping the neural responses of facial expressions provides additional insights into amygdala function. For example, the amygdala appears to be especially responsive to fearful faces compared to angry faces (Whalen et al., 2001). While both angry and fearful expressions provide social information about threats, only angry expressions inform the observer of its source. Accordingly, the inherent ambiguity of fearful expressions appears

to particularly activate amygdala responses to threat. This suggests that the amygdala may function to increase the susceptibility of beliefs to environmental inputs in order to aid in the resolution of predictive uncertainty, bearing a striking resemblance to the role of uncertainty in Bayesian inference.

A critical question is whether the amygdala specifically functions to resolve uncertainty in response to potential threats, or functions more generally across threatening and non-threatening contexts. Using facial expressions of surprise, Kim and colleagues have helped resolve this ambiguity. Expressions of fear and surprise are similarly indicative of a significant, uncertain event. However, in comparison to fearful expressions, which are universally interpreted as negative in nature, expressions of surprise can be interpreted either positively or negatively. Those who interpreted expressions of surprise negatively showed greater amygdala reactivity in the ventral portion of the amygdala and weaker regulatory activity of the mPFC, while the reverse was true for those who interpreted expressions positively (Hackjin Kim, 2005; Kim, Somerville, Johnstone, Alexander, & Whalen, 2003). Meanwhile, the dorsal portion of the amygdala showed comparable increases in activity regardless of whether the surprised face was interpreted positively or negatively. This highlights a dual function of the amygdala in signaling information about both social threat and uncertainty, and suggests that mPFC may regulate how the amygdala responds to uncertain events as a function of one's impression. Indeed, ample evidence suggests a role for the mPFC in forming impressions about others' mental states and demonstrate that this structure is particularly sensitive when one's impression is positive, rather than to negative (Cooper, Kreps, Wiebe, Pirkl, & Knutson, 2010; Harris, McClure, Bos, Cohen, & Fiske, 2007). This suggests that the mPFC may exert greater inhibitory

control over amygdala activity when an agent is believed to possess a good moral character, relative to a bad moral character. Consequently, in the studies presented in this dissertation observing an agent acting immorally may augment amygdala representations of uncertainty and recruit sensory processes to boost learning from environmental inputs.

7.2.2 LIMITATIONS

Limitations associated with particular studies have been described in their relevant chapters. However, despite the robustness of our findings, there are some general methodological limitations that should be discussed. First, accepting money in exchange for shocks that are painful but not dangerous is a relatively mild moral transgression. Mild transgressions represent the vast majority of transgressions that will be personally experienced by most individuals, and thus the mechanisms we identify may explain everyday changes in beliefs about the moral character of others. However, it is unclear how these results will generalize to learning about more extreme transgressions, such as assault, rape, or murder.

There may be good reason to believe that the results will not generalize to extreme behaviours that are highly diagnostic about an individual's character. Research in non-social perceptual learning suggests a curvilinear relationship between uncertainty and arousal, where uncertainty is highest for moderate levels of arousal and weakest for either overly high or low levels of arousal (Nassar et al., 2012). This is consistent with the Yerkes-Dawson 'inverted-U" relationship between arousal and learning, where learning is optimal at moderate levels of arousal (Salehi, Cordero, & Sandi, 2010; Yerkes & Dodson, 1908). Evaluations of immoral actions are sensitive to the magnitude of harm (Shenhav & Greene, 2010) and physiological arousal is greater for moral actions that are evaluatively worse

(McDonald, Defever, & Navarrete, 2017). If more extreme acts of immorality elicit significantly greater levels of autonomic arousal, this predicts that negative beliefs that are based on extreme behaviours may be more certain and resistant to updating.

This hypothesis may explain evidence for the negativity bias in impression formation, which suggests that the difficulty in revising an initial bad impression is directly related to the extremity of behaviours; impressions made from less extreme behaviours are easier to update in light of disconfirming evidence than more extreme behaviours. (Skowronski & Carlston, 1992; Wojciszke et al., 1993). Accordingly, the relative flexibility of bad impressions over good impressions may decrease when behaviours are highly diagnostic of an agent's morality due to higher levels of autonomic arousal. From an evolutionary perspective, leaving room for further learning when behaviours are highly diagnostic of bad moral character may pose a risk to one's wellbeing. If arousal is related to uncertainty in a curvilinear fashion, then more extreme behaviours should be associated with greater certainty and therefore prevent people from giving the "wrong people" the benefit of the doubt. Therefore, when we take into account an inverted-U relationships between arousal and uncertainty, the framework outlined in this dissertation can account for adaptive behaviour in response to both moderately bad people, who should be given a second chance, and extremely bad people, whose relationship should be terminated for good reason.

This raises a new potential explanation for the finding that BPD is associated with more certain beliefs about bad agents and less certain beliefs about good agents, relative to non-BPD control participants. According to Linehan's biosocial model (Linehan, 1993), an underlying biological vulnerability to emotional dysfunction interacts with social

experiences (e.g., childhood abuse) to contribute to the development of interpersonal dysfunction in BPD. Biological vulnerabilities include higher baseline levels of emotional arousal, heightened emotional reactivity and delayed return to baseline following emotional arousal (Gratz, Rosenthal, Tull, Lejuez, & Gunderson, 2010; Henry et al., 2001; Stiglmayr et al., 2005). Therefore, if arousal plays a role in moral inference than the observed differences between patients with BPD and non-BPD control participants may arise due to underlying biological differences in physiological arousal. Previous work has proposed how baseline levels of arousal interact with the emotional intensity of social cues to produce deficits in BPD (Daros, Zakzanis, & Ruocco, 2013). Specifically, higher baseline levels of arousal are hypothesized to enhance processing for low intensity social cues but impede processing for higher intensity social cues. If learning is optimal at moderate levels of arousal, then higher baseline levels of arousal in BPD may enhance learning for stimuli with low emotional intensity (good agent), whereas when stimuli are at a higher emotional intensity (bad agent) the same baseline arousal may interfere with learning.

7.3 FUTURE DIRECTIONS

7.3.1 EXAMINING AUTONOMIC AROUSAL IN MORAL INFERENCE

Chapter I postulated a theoretical framework where arousal associated with social threat (Fouragnan, 2013; Öhman, 1986) drives asymmetric learning about good and bad agents because arousal is linked to belief updating in non-social perceptual learning (Jepma & Nieuwenhuis, 2010; Nassar et al., 2012; Urai et al., 2017). Evidence suggests that signs of immorality induce an enhanced state of arousal in perceivers (Fouragnan, 2013; Öhman,

1986), however, whether arousal in the context of the Moral Inference Task differs when evaluating good and bad agents is unknown. The closest insight for this question comes from Chapter 4, where prior moral expectations were manipulated using face stimuli developed to vary on their level of perceived threat (Oosterhof & Todorov, 2008). Regardless of the agents' behaviour, beliefs about agents represented by a highly threatening face were more uncertain than beliefs about agents represented by a face that was less threatening. Because threatening faces are evaluated as more arousing than friendly or neutral faces (Noordewier et al., 2019; Roelofs et al., 2010; Schupp et al., 2004), it's possible that differences in arousal drove greater uncertainty about beliefs about agents represented by more threatening avatars. Although the data appears consistent with the hypothesized mechanism, the studies presented in this dissertation cannot fully speak to the precise role of arousal because autonomic responses were not measured. To determine concretely whether asymmetric Bayesian updating stems from differences in arousal, future work must measure physiological responses, such as pupillometry, galvanic skin response, or heart rate.

Emotions are at the very core of autonomic responses and are therefore an important consideration for future studies assessing the role of arousal in moral inference. Both anger, anxiety, and enthusiasm are associated with increased attention, heart rate, breathing, and general activity of the autonomic nervous system (Kreibig, 2010; Valentino, Hutchings, Banks, & Davis, 2008). Yet cognitive theories of emotion suggest that each derives from different cognitive appraisals about the current state of the environment, and can trigger distinct motivations and behaviours (Frijda, 1988; Frijda, Manstead, & Bem, 2000). For instance, anger increases with certainty about a negative outcome while anxiety increases

with uncertainty about a negative outcome. Consequently, inducing anger triggers the use of cognitive heuristics to motivate immediate action, while inducing anxiety triggers more systematic information processing to encourage the integration of new information and minimize uncertainty (Tiedens & Linton, 2001).

This raises the prediction that arousal induced by anxiety, rather than other emotions induced by immoral actors (e.g., anger), mediates the relationship between moral character and learning because only anxiety enhances the motivation to seek out and encode new information. In support of this prediction, Valentino et al. (Valentino et al., 2008) manipulated the perceived threat of a political candidate using newspaper articles and gave participants the opportunity to learn more about the candidates on a website after rating how angry, anxious, and enthusiastic the article made them feel. Compared to when the candidate was depicted as a low threat, those who read the article about the highly threatening candidate sought out significantly more information about the candidate and remembered more information in a subsequent memory test. Crucially, the impact of candidate threat on subsequent information seeking and memory was mediated by anxiety ratings, specifically, as opposed to anger or enthusiasm.

Valentino et al. (Valentino et al., 2008) show that emotions induced by a target influences how people learn about that individual. However, autonomic responses linked to emotions fluctuate over time and often without awareness. Therefore, a key question to explore is whether the relevance of the emotion to the task at hand is important for mediating the relationship between moral character and learning. That is, does it matter whether anxiety arises from the individual you are learning about versus an incidental, unrelated threat? Subjecting individuals to a threat of receiving painful electric shocks has

successfully induced a continuous state of anxiety in healthy adults (Robinson et al., 2013). Threat-induced anxiety states enhance the automatic processing of neutral sensory stimuli (Lojowska et al., 2019), suggesting that the mechanisms supporting the relationship between threat and information processing may generalize to stimuli unrelated to the threatening cue. If anxiety induced by threat of shock increases belief uncertainty and subsequent belief updating, then inducing anxiety using threat of shock may increase uncertainty of beliefs about agents who are not threatening. The findings from these studies may be important for understanding both adaptive and pathological social behaviours arising from individual differences in anxiety and mood.

7.3.2 INFERRING THE STABILITY OF MORAL TRAITS

Finally, the present research only addressed one type of uncertainty, namely, "informational uncertainty" (also known as estimation uncertainty), which is highest for new contexts and generally decreases over time as knowledge is gained. Uncertainty, however, can be parsed into several distinct forms. For instance, people's motivations and intentions may fluctuate over time, creating uncertainty about the stability of their traits. This form of uncertainty, often known as "environmental uncertainty", has a normative relationship with learning: greater uncertainty about the stability of the environment increases the difficulty of predicting outcomes. For example, assuming the chef at a restaurant influences the quality of the food, uncertainty about how often the chef comes in to cook makes it harder to predict whether the food experienced on the next visit will be like the last. This raises the question: how stable do people think others' moral preferences are?

Previous work has examined this question by asking how people infer the probability of an agent providing helpful advice when the agent's motivation to be helpful or harmful fluctuates over time (Behrens et al., 2008; Diaconescu et al., 2014, 2017). This work revealed that people are less likely to rely on their beliefs about the agent's current moral intentions (to be harmful or helpful) when intentions are believed to be unstable, relative to stable. In other words, inferring that someone rapidly changes from providing helpful advice to harmful advice makes beliefs about the accuracy of the advice less certain, and thus motivates more exploratory behaviour. However, the extent to which people believe that inferences about moral character made in one context will generalize to others is unclear. For instance, someone may be very kind to you, however, you may be very uncertain about how kind they are towards their family, their co-workers, stranger, etc. If you believe that people have stable moral preferences across contexts, then you will infer only small fluctuations in an individual's morality when they interact with other people. Conversely, if you believe that people adapt moral preferences depending on who they interact with, you will infer large fluctuations in morality. This raises the question for future work as to whether inferences about someone's morality – as either good or bad - based on how they behave in one context predict inferences about how much their morality will change across contexts.

The diagnosticity theory for the negativity bias in impression formation postulates that bad people often behave morally whereas good people rarely behave immorally (Skowronski & Carlston, 1987, 1989). This predicts that people infer greater instability about bad agents' morality across contexts than good agents. Crucially, in Bayesian inference the relationship between beliefs about the stability of the environment and

estimation uncertainty is normative (Mathys et al., 2011): an inference that the environment is highly unstable decreases certainty that one's current belief about the environment is accurate and consequently more weight should be given to new information, using a *faster* learning rate. Conversely, in stable conditions the relative weight given to recent information over historical information should decrease over time, using a *slower* learning rate. This raises the intriguing question whether faster learning rates for bad agents over good agents was truly capturing differences in the extent to which beliefs about moral character were uncertain (referring to the variance in posterior distributions) versus differences in the extent to which participants believed moral preferences would change from trial to trial.

In the Moral Inference Task, participants were instructed to predict the decisions of agents who chose whether to increase their own profit at the expense of electric shocks to a stranger. Because the context of the decisions remained constant throughout the task, there was little reason to believe that the agent's motivations or intentions would change from trial to trial. Nonetheless, it is important to address the possibility that more uncertain and volatile beliefs about bad agents can be explained by inferences that immoral preferences in the Moral Inference Task are less generalizable to new contexts, and therefore less diagnostic about character, than moral preferences. New contexts might include those where agents make decisions about whether to increase their own profit at the expense of electric shocks to a friend or family, or decisions about whether to increase the amount of money donated to a charity at the expense of electric shocks to a stranger, friend or family. To investigate inferences about the stability of beliefs as a function of moral character, I hope to include a range of such contexts, while manipulating both the

moral character (good versus bad) and stability of moral character across contexts (high versus low). Comparing the HGF model outlined in Chapter 2 to a model that includes a third hierarchical level, which estimates beliefs about the stability of moral traits across contexts, I can verify whether people infer on the stability of moral preferences to optimally predict moral decisions and how inferences about morality influence these processes.

7.4 CONCLUDING REMARKS

Those around us may occasionally commit immoral acts and those closest to us may hurt us. To cement bonds we must have mechanisms for maintaining relationships despite actions that break our trust and mar positive impressions. Understanding these mechanisms requires a deeper understanding of how moral impressions are formed and evolve over time. Furthermore, it is critical to understand how maladaptive behaviours arise when these processes go awry. Viewing the brain as an organ of approximate Bayesian inference can help understand how the brain represents others' morality and uses this information for adaptive decision-making. The studies described in Part I of this dissertation indicate that moral inference is described by an asymmetric Bayesian updating mechanism, where beliefs about the morality of bad agents are more uncertain (and therefore more amenable to updating) than beliefs about the morality of good agents. Our model and data reveal a cognitive mechanism that rapidly discounts prior expectations to permit flexible updating of beliefs about potentially threatening others, a mechanism that could facilitate forgiveness when initial bad impressions turn out to be inaccurate. Part II of the dissertation reveal novel cognitive processes that may explain the emergence of maladaptive behaviour related to exposure to violence and BPD. Collectively, the findings provide the basis for a

new program of research investigating the role of autonomic responses in moral inference, its neurological basis, and clinical implications.

BIBLIOGRAPHY

Aksoy, O., & Weesie, J. (2014). Hierarchical Bayesian analysis of outcome- and process-based social preferences and beliefs in Dictator Games and sequential Prisoner's Dilemmas. *Social Science Research*, *45*, 98–116. https://doi.org/10.1016/j.ssresearch.2013.12.014

Albert Bandura. (1978). Social Learning Theory of Aggression. *Journal of Communication*, *28*(3), 12–29. https://doi.org/10.1111/j.1460-2466.1978.tb01621.x

Alexander, R. (1987). *The Biology of Moral Systems (Foundations of Human Behavior)*. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0202011747

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *ELife*, *5*. https://doi.org/10.7554/eLife.18103

American Psychiatric Association. (2001). *Practice Guideline for the Treatment of Patients with Borderline Personality Disorder*. American Psychiatric Pub.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*.

Anderson, Elijah. (1994). The Code of the Streets. *Monthy Atlantic*, (273), 81–94.

Anderson, Eric, Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The Visual Impact of Gossip. *Science*, *332*(6036), 1446–1448. https://doi.org/10.1126/science.1201574

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394–400. https://doi.org/10.1037/h0022280

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Axelrod, R. M. (2006). *The Evolution of Cooperation*. Basic Books.

Azzam, T., & Jacobson, M. R. (2013). Finding a Comparison Group: Is Online Crowdsourcing a Viable Option? *American Journal of Evaluation*, *34*(3), 372–384. https://doi.org/10.1177/1098214013490223

Bailey, B. N., Delaney-Black, V., Hannigan, J. H., Ager, J., Sokol, R. J., & Covington, C. Y. (2005). Somatic Complaints in Children and Community Violence Exposure. *Journal of Developmental & Behavioral Pediatrics*, *26*(5), 341. https://doi.org/10.1097/00004703-200510000-00001

Barnow, S., Stopsack, M., Grabe, H. J., Meinke, C., Spitzer, C., Kronmüller, K., & Sieswerda, S. (2009). Interpersonal evaluation bias in borderline personality disorder. *Behaviour Research and Therapy*, *47*(5), 359–365. https://doi.org/10.1016/j.brat.2009.02.003

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience*, *6*(5), 572–581. https://doi.org/10.1093/scan/nsq086

Baskin, D., & Sommers, I. (2014). Exposure to Community Violence and Trajectories of Violent Offending. *Youth Violence and Juvenile Justice*, *12*(4), 367–385. https://doi.org/10.1177/1541204013506920

Baskin, D., & Sommers, I. (2015). Trajectories of Exposure to Community Violence and Mental Health Symptoms Among Serious Adolescent Offenders. *Criminal Justice and Behavior*, *42*(6), 587–609. https://doi.org/10.1177/0093854814556882

Baskin, T. W., & Enright, R. D. (2004). Intervention Studies on Forgiveness: A Meta-Analysis. *Journal of Counseling & Development*, *82*(1), 79–90. https://doi.org/10.1002/j.1556-6678.2004.tb00288.x

Baskin-Sommers, A. R., & Baskin, D. (2016). Psychopathic Traits Mediate the Relationship Between Exposure to Violence and Violent Juvenile Offending. *Journal of Psychopathology and Behavioral Assessment*, *38*(3), 341–349. https://doi.org/10.1007/s10862-016-9535-0

Baskin-Sommers, A. R., Baskin, D. R., Sommers, I., Casados, A. T., Crossman, M. K., & Javdani, S. (2016). The impact of psychopathology, race, and environmental

context on violent offending in a male adolescent sample. *Personality Disorders*, *7*(4), 354–362. https://doi.org/10.1037/per0000168

Bateman, A., & Fonagy, P. (2009). Randomized Controlled Trial of Outpatient Mentalization-Based Treatment Versus Structured Clinical Management for Borderline Personality Disorder. *American Journal of Psychiatry*, *166*(12), 1355–1364. https://doi.org/10.1176/appi.ajp.2009.09040539

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

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

Bender, D. S., & Skodol, A. E. (2007). Borderline Personality as a Self-Other Representational Disturbance. *Journal of Personality Disorders*, *21*(5), 500–517. https://doi.org/10.1521/pedi.2007.21.5.500

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Berk, M. S., Jeglic, E., Brown, G. K., Henriques, G. R., & Beck, A. T. (2007). Characteristics of Recent Suicide Attempters with and without Borderline Personality Disorder. *Archives of Suicide Research*, *11*(1), 91–104. https://doi.org/10.1080/13811110600992951

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., … Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, *27*(2), 169–190. https://doi.org/10.1016/S0145-2134(02)00541-0

Bertsch, K., Gamer, M., Schmidt, B., Schmidinger, I., Walther, S., Kästel, T., … Herpertz, S. C. (2013). Oxytocin and Reduction of Social Threat Hypersensitivity in Women

With Borderline Personality Disorder. *American Journal of Psychiatry*, *170*(10), 1169–1177. https://doi.org/10.1176/appi.ajp.2013.13020263

Besbris, M., Faber, J. W., Rich, P., & Sharkey, P. (2015). Effect of neighborhood stigma on economic transactions. *Proceedings of the National Academy of Sciences*, 201414139. https://doi.org/10.1073/pnas.1414139112

Blair, R. J. R. (2001). Neurocognitive models of aggression, the antisocial personality disorders, and psychopathy. *Journal of Neurology, Neurosurgery & Psychiatry*, *71*(6), 727–731. https://doi.org/10.1136/jnnp.71.6.727

Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain and Cognition*, *55*(1), 198–208. https://doi.org/10.1016/S0278-2626(03)00276-8

Boxer, P., Middlemass, K., & Delorenzo, T. (2009). Exposure to Violent Crime During Incarceration: Effects on Psychological Adjustment Following Release. *Criminal Justice and Behavior*, *36*(8), 793–807. https://doi.org/10.1177/0093854809336453

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

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology*, *49*(5), 811–821. https://doi.org/10.1016/j.jesp.2013.04.005

Brañas-Garza, P., Rodríguez-Lara, I., & Sánchez, A. (2017). Humans expect generosity. *Scientific Reports*, *7*, 42446. https://doi.org/10.1038/srep42446

Briscoe, M. E., Woodyard, H. D., & Shaw, M. E. (1967). Personality impression change as a function of the favorableness of first impressions. *Journal of Personality*, *35*(2), 343–357. https://doi.org/10.1111/j.1467-6494.1967.tb01433.x

Brown, R. P. (2003). Measuring Individual Differences in the Tendency to Forgive: Construct Validity and Links with Depression. *Personality and Social Psychology Bulletin*, *29*(6), 759–771. https://doi.org/10.1177/0146167203029006008

Buckholtz, J. W. (2015). Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences*, *3*, 122–129. https://doi.org/10.1016/j.cobeha.2015.03.004

Buka, S. L., Stichick, T. L., Birdthistle, I., & Earls, F. J. (2001). Youth Exposure to Violence: Prevalence, Risks, and Consequences. *American Journal of Orthopsychiatry*, *71*(3), 298–310. https://doi.org/10.1037/0002-9432.71.3.298

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond Bipolar Conceptualizations and Measures: The Case of Attitudes and Evaluative Space. *Personality and Social Psychology Review*, *1*(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2

Callan, M. J., Ferguson, H. J., & Bindemann, M. (2013). Eye movements to audiovisual scenes reveal expectations of a just world. *Journal of Experimental Psychology: General*, *142*(1), 34–40. https://doi.org/10.1037/a0028261

Carr, C. T., & Walther, J. B. (2014). Increasing Attributional Certainty via Social Media: Learning About Others One Bit at a Time. *Journal of Computer-Mediated Communication*, *19*(4), 922–937. https://doi.org/10.1111/jcc4.12072

Casciano, R., & Massey, D. S. (2008). Neighborhoods, employment, and welfare use: Assessing the influence of neighborhood socioeconomic composition. *Social Science Research*, *37*(2), 544–558. https://doi.org/10.1016/j.ssresearch.2007.08.008

Clifton, A., Pilkonis, P. A., & McCarty, C. (2007). Social Networks in Borderline Personality Disorder. *Journal of Personality Disorders*, *21*(4), 434–441. https://doi.org/10.1521/pedi.2007.21.4.434

Cone, J., & Ferguson, M. J. (2015). He Did What?: The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T., & Knutson, B. (2010). When Giving Is Good: Ventromedial Prefrontal Cortex Activation for Others' Intentions. *Neuron*, *67*(3), 511–521. https://doi.org/10.1016/j.neuron.2010.06.030

Cornwell, B. R., Garrido, M. I., Overstreet, C., Pine, D. S., & Grillon, C. (2017). The Unpredictive Brain Under Threat: A Neurocomputational Account of Anxious

Hypervigilance. *Biological Psychiatry*, *82*(6), 447–454. https://doi.org/10.1016/j.biopsych.2017.06.031

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York, NY, US: Oxford University Press.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology*, *92*(2), 208–231. https://doi.org/10.1037/0022-3514.92.2.208

Critchfield, K. L., Levy, K. N., Clarkin, J. F., & Kernberg, O. F. (2008). The relational context of aggression in borderline personality disorder: Using adult attachment style to predict forms of hostility. *Journal of Clinical Psychology*, *64*(1), 67–82. https://doi.org/10.1002/jclp.20434

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. https://doi.org/10.1073/pnas.1408988111

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. https://doi.org/10.1038/nn.4557

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., … Dolan, R. J. (2015). Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making. *Current Biology: CB*, *25*(14), 1852–1859. https://doi.org/10.1016/j.cub.2015.05.021

Daros, A. R., Zakzanis, K. K., & Ruocco, A. C. (2013). Facial emotion recognition in borderline personality disorder. *Psychological Medicine*, *43*(9), 1953–1963. https://doi.org/10.1017/S0033291712002607

Davis, M., & Whalen, P. J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry*, *6*(1), 13. https://doi.org/10.1038/sj.mp.4000812

De Brito, S. A., Viding, E., Sebastian, C. L., Kelly, P. A., Mechelli, A., Maris, H., & McCrory, E. J. (2013). Reduced orbitofrontal and temporal grey matter in a community sample of maltreated children. *Journal of Child Psychology and Psychiatry*, *54*(1), 105–112. https://doi.org/10.1111/j.1469-7610.2012.02597.x

De Bruin, E. N. M., & van Lange, P. A. M. (2000). What People Look for in Others: Influences of the Perceiver and the Perceived on Information Selection. *Personality and Social Psychology Bulletin*, *26*(2), 206–219. https://doi.org/10.1177/0146167200264007

Debaere, V., Vanheule, S., Van Roy, K., Meganck, R., Inslegers, R., & Mol, M. (2016). Changing encounters with the other: A focus group study on the process of change in a therapeutic community. *Psychoanalytic Psychology*, *33*(3), 406–419. https://doi.org/10.1037/a0036862

Delaney-Black, V., Covington, C., Ondersma, S. J., Nordstrom-Klee, B., Templin, T., Ager, J., … Sokol, R. J. (2002). Violence Exposure, Trauma, and IQ and/or Reading Deficits Among Urban Children. *Archives of Pediatrics & Adolescent Medicine*, *156*(3), 280–285. https://doi.org/10.1001/archpedi.156.3.280

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the Intentions of Others by Hierarchical Bayesian Learning. *PLoS Comput Biol*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634. https://doi.org/10.1093/scan/nsw171

Dibbets, Pauline, Adolphs, Laura, Close, Ingeborg, Herings, Anke, Kiggen, Maiken, Kinneging, Maaike, … RS: FPN CPS III. (2012). Reversal of Attitude: The Influence of Counter-Attitudinal Information. *Journal of Social Sciences*, *8*(3), 390–396.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290. https://doi.org/10.1037/h0033731

Dodge, K. A., Bates, J. E., & Pettit, G. S. (1990). Mechanisms in the cycle of violence. *Science*, *250*(4988), 1678–1683. https://doi.org/10.1126/science.2270481

DuRant, R. H., Pendergrast, R. A., & Cadenhead, C. (1994). Exposure to violence and victimization and fighting behavior by urban black adolescents. *Journal of Adolescent Health*, *15*(4), 311–318. https://doi.org/10.1016/1054-139X(94)90604-1

Dyck, M., Habel, U., Slodczyk, J., Schlummer, J., Backes, V., Schneider, F., & Reske, M. (2009). Negative bias in fast emotion discrimination in borderline personality disorder. *Psychological Medicine*, *39*(5), 855–864. https://doi.org/10.1017/S0033291708004273

Ebert, A., Kolb, M., Heller, J., Edel, M.-A., Roser, P., & Brüne, M. (2013). Modulation of interpersonal trust in borderline personality disorder by intranasal oxytocin and childhood trauma. *Social Neuroscience*, *8*(4), 305–313. https://doi.org/10.1080/17470919.2013.807301

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Elliott, R., & Dolan, R. J. (1998). Activation of Different Anterior Cingulate Foci in Association with Hypothesis Testing and Response Selection. *NeuroImage*, *8*(1), 17–29. https://doi.org/10.1006/nimg.1998.0344

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–1519. https://doi.org/10.1162/jocn.2007.19.9.1508

Eysenck, M. (2012). *Attention and Arousal: Cognition and Performance*. Springer Science & Business Media.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of Direct Social Experience on Trust Decisions and Neural Reward Circuitry. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00148

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association*, *7*(2–3), 235–266. https://doi.org/10.1162/JEEA.2009.7.2-3.235

Fertuck, E. A., Grinband, J., Mann, J. J., Hirsch, J., Ochsner, K., Pilkonis, P., … Stanley, B. (2018). Trustworthiness appraisal deficits in borderline personality disorder are associated with prefrontal cortex, not amygdala, impairment. *NeuroImage: Clinical*, 101616. https://doi.org/10.1016/j.nicl.2018.101616

Fertuck, E. A., Grinband, J., & Stanley, B. (2013). Facial trust appraisal negatively biased in borderline personality disorder. *Psychiatry Research*, *207*(3), 195–202. https://doi.org/10.1016/j.psychres.2013.01.004

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., … Corlett, P. R. (2018). Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biological Psychiatry*, *84*(11), 838–845. https://doi.org/10.1016/j.biopsych.2018.05.020

Finkelhor, D., Turner, H. A., Shattuck, A., & Hamby, S. L. (2013). Violence, Crime, and Abuse Exposure in a National Sample of Children and Youth: An Update. *JAMA Pediatrics*, *167*(7), 614–621. https://doi.org/10.1001/jamapediatrics.2013.42

Finkelhor, D., Turner, H. A., Shattuck, A., & Hamby, S. L. (2015). Prevalence of Childhood Exposure to Violence, Crime, and Abuse: Results From the National Survey of Children's Exposure to Violence. *JAMA Pediatrics*, *169*(8), 746–754. https://doi.org/10.1001/jamapediatrics.2015.0676

Finkelhor, D., Turner, H., Ormrod, R., & Hamby, S. L. (2010). Trends in childhood violence and abuse exposure: Evidence from 2 national surveys. *Archives of Pediatrics & Adolescent Medicine*, *164*(3), 238–242. https://doi.org/10.1001/archpediatrics.2009.283

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

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

Fitzpatrick, K. M., & Boldizar, J. P. (1993). The prevalence and consequences of exposure to violence among African-American youth. *Journal of the American Academy of Child and Adolescent Psychiatry*, *32*(2), 424–430. https://doi.org/10.1097/00004583-199303000-00026

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personality Disorder and Emotion Dysregulation*, *4*(1), 9. https://doi.org/10.1186/s40479-017-0062-8

Fouragnan, E. (2013). *The Neural Computation of Trust and Reputation* (Phd, University of Trento). Retrieved from http://eprints-phd.biblio.unitn.it/970/

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *The Journal of Neuroscience*, *33*(8), 3602–3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013

Fowler, P. J., Tompsett, C. J., Braciszewski, J. M., Jacques-Tiura, A. J., & Baltes, B. B. (2009). Community violence: A meta-analysis on the effect of exposure and mental health outcomes of children and adolescents. *Development and Psychopathology*, *21*(1), 227–259. https://doi.org/10.1017/S0954579409000145

Freedman, J. L., & Steinbruner, J. D. (1964). Perceived choice and resistance to persuasion. *The Journal of Abnormal and Social Psychology*, *68*(6), 678–681.

Frick, C., Lang, S., Kotchoubey, B., Sieswerda, S., Dinu-Biringer, R., Berger, M., … Barnow, S. (2012). Hypersensitivity in Borderline Personality Disorder during Mindreading. *PLOS ONE*, *7*(8), e41650. https://doi.org/10.1371/journal.pone.0041650

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, *43*(5), 349–358. https://doi.org/10.1037/0003-066X.43.5.349

Frijda, N. H., Manstead, A. S. R., & Bem, S. (2000). *Emotions and Beliefs: How Feelings Influence Thoughts*. Cambridge University Press.

Frith, C. D., & Frith, U. (2012). *Mechanisms of Social Cognintion*. *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *The American Economic Review*, *102*(2), 720–749. https://doi.org/10.1257/aer.102.2.720

Gantman, A. P., Bavel, V., & J, J. (2015). *Moral Perception* (SSRN Scholarly Paper No. ID 2647767). Retrieved from Social Science Research Network website: http://papers.ssrn.com/abstract=2647767

Gantman, A. P., & Van Bavel, J. J. (2016). Exposure to justice diminishes moral perception. *Journal of Experimental Psychology: General*, *145*(12), 1728–1739. https://doi.org/10.1037/xge0000241

Garbarino, J., & Sherman, D. (1980). High-Risk Neighborhoods and High-Risk Families: The Human Ecology of Child Maltreatment. *Child Development*, *51*(1), 188–198. https://doi.org/10.2307/1129606

Gartner, J. (1988). The capacity to forgive: An object relations perspective. *Journal of Religion and Health*, *27*(4), 313–320. https://doi.org/10.1007/BF01533199

Gert, B. (2004). *Common morality : Deciding what to do*. Oxford ; New York: Oxford University Press.

Giesen-Bloo, J., Dyck, R. van, Spinhoven, P., Tilburg, W. van, Dirksen, C., Asselt, T. van, … Arntz, A. (2006). Outpatient Psychotherapy for Borderline Personality Disorder: Randomized Trial of Schema-Focused Therapy vs Transference-Focused Psychotherapy. *Archives of General Psychiatry*, *63*(6), 649–658. https://doi.org/10.1001/archpsyc.63.6.649

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–179. https://doi.org/10.1006/jtbi.2000.2111

Goffman, A. (2009). On the Run: Wanted Men in a Philadelphia Ghetto. *American Sociological Review*, *74*(3), 339–357. https://doi.org/10.1177/000312240907400301

Goffman, A. (2015). *On the Run: Fugitive Life in an American City*. Picador.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Gratz, K. L., Rosenthal, M. Z., Tull, M. T., Lejuez, C. W., & Gunderson, J. G. (2010). An experimental investigation of emotional reactivity and delayed emotional recovery in borderline personality disorder: The role of shame. *Comprehensive Psychiatry*, *51*(3), 275–285. https://doi.org/10.1016/j.comppsych.2009.08.005

Grim, P. (1995). The greater generosity of the spatialized prisoner's dilemma. *Journal of Theoretical Biology*, *173*(4), 353–359. https://doi.org/10.1006/jtbi.1995.0068

Guerra, N. G., Huesmann, L. R., & Spindler, A. (2003). Community Violence Exposure, Social Cognition, and Aggression Among Urban Elementary School Children. *Child Development*, *74*(5), 1561–1576. https://doi.org/10.1111/1467-8624.00623

Gunderson, J. G., Stout, R. L., McGlashan, T. H., Shea, M. T., Morey, L. C., Grilo, C. M., … Skodol, A. E. (2011). Ten-Year Course of Borderline Personality Disorder: Psychopathology and Function From the Collaborative Longitudinal Personality Disorders Study. *Archives of General Psychiatry*, *68*(8), 827–837. https://doi.org/10.1001/archgenpsychiatry.2011.37

Guo, X., Zheng, L., Wang, H., Zhu, L., Li, J., Wang, Q., … Yang, Z. (2013). Exposure to violence reduces empathetic responses to other's pain. *Brain and Cognition*, *82*(2), 187–191. https://doi.org/10.1016/j.bandc.2013.04.005

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233. https://doi.org/10.1038/nn.4080

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology*, *52*, 15–23. https://doi.org/10.1016/j.jesp.2013.12.001

Hackjin Kim, L. H. S. (2005). Contextual Modulation of Amygdala Responsivity to Surprised Faces. *Journal of Cognitive Neuroscience*, *16*(10), 1730–1745. https://doi.org/10.1162/0898929042947865

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66. https://doi.org/10.1162/0011526042365555

Hare, R. (2003). *Manual for the Revised Psychopathy Checklist (2 ed.)*. Toronto, Ontario, Canada: Multi-Health System.

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The revised Psychopathy Checklist: Reliability and factor structure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(3), 338–341. https://doi.org/10.1037/1040-3590.2.3.338

Harris, L. T., McClure, S. M., Bos, W. van den, Cohen, J. D., & Fiske, S. T. (2007). Regions of the MPFC differentially tuned to social and nonsocial affective evaluation. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 309–316. https://doi.org/10.3758/CABN.7.4.309

Hawkins, J. D., Herrenkohl, T. I., Farrington, D. P., Brewer, D., Catalano, R. F., Harachi, T. W., & Cothern, L. (2000). *Predictors of Youth Violence. Juvenile Justice Bulletin*. Retrieved from https://eric.ed.gov/?id=ED440196

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling.*

Henry, C., Mitropoulou, V., New, A. S., Koenigsberg, H. W., Silverman, J., & Siever, L. J. (2001). Affective instability and impulsivity in borderline personality and bipolar II disorders: Similarities and differences. *Journal of Psychiatric Research*, *35*(6), 307–312. https://doi.org/10.1016/S0022-3956(01)00038-3

Hetey, R. C., & Eberhardt, J. L. (2018). The Numbers Don't Speak for Themselves: Racial Disparities and the Persistence of Inequality in the Criminal Justice System , The Numbers Don't Speak for Themselves: Racial Disparities and the Persistence of Inequality in the Criminal Justice System. *Current Directions in Psychological Science*, *27*(3), 183–187. https://doi.org/10.1177/0963721418763931

Heuer, F., & Reisberg, D. (1992). Emotion, arousal, and memory for detail. In *The handbook of emotion and memory: Research and theory* (pp. 151–180). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Holm, A. L., Berg, A., & Severinsson, E. (2009). Longing for Reconciliation: A Challenge for Women with Borderline Personality Disorder. *Issues in Mental Health Nursing*, *30*(9), 560–568. https://doi.org/10.1080/01612840902838579

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. https://doi.org/10.1007/s10683-011-9273-9

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

Huesmann, L. R., & Kirwil, L. (2007). Why observing violence increases the risk of violent behavior by the observer. In *The Cambridge handbook of violent behavior and aggression* (pp. 545–570). https://doi.org/10.1017/CBO9780511816840.029

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting From Misfortune When Harmless Actions Are Judged to Be Morally Blameworthy. *Personality and Social Psychology Bulletin*, *38*(1), 52–62. https://doi.org/10.1177/0146167211430232

Irwin, W., Davidson, R. J., Lowe, M. J., Mock, B. J., Sorenson, J. A., & Turski, P. A. (1996). Human amygdala activation detected with echo-planar functional magnetic resonance imaging. *Neuroreport*, *7*(11), 1765–1769.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Javdani, S., Abdul-Adil, J., Suarez, L., Nichols, S. R., & Farmer, A. D. (2014). Gender Differences in the Effects of Community Violence on Mental Health Outcomes in a Sample of Low-Income Youth Receiving Psychiatric Care. *American Journal of Community Psychology*, *53*(3–4), 235–248. https://doi.org/10.1007/s10464-014-9638-2

Jepma, M., & Nieuwenhuis, S. (2010). Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, *142*, 12–38. https://doi.org/10.1016/j.cognition.2015.05.006

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, *28*(8), 474–481. https://doi.org/10.1016/j.tree.2013.05.014

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*(6), 899.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (2000). *Choices, Values, and Frames*. Cambridge University Press.

Kaletsch, M., Krüger, B., Pilgramm, S., Stark, R., Lis, S., Gallhofer, B., … Sammer, G. (2014). Borderline personality disorder is associated with lower confidence in perception of emotional body movements. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01262

Kapp, B. S., Supple, W. F., & Whalen, P. J. (1994). Effects of electrical stimulation of the amygdaloid central nucleus on neocortical arousal in the rabbit. *Behavioral Neuroscience*, *108*(1), 81–93. https://doi.org/10.1037/0735-7044.108.1.81

Kapp, B. S., Whalen, P. J., Supple, W. F., & Pascoe, J. P. (1992). Amygdaloid contributions to conditioned arousal and sensory information processing. In *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 229–254). New York, NY, US: Wiley-Liss.

Keil, A., Gruber, T., Müller, M. M., Moratti, S., Stolarova, M., Bradley, M. M., & Lang, P. J. (2003). Early modulation of visual perception by emotional arousal: Evidence from steady-state visual evoked brain potentials. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(3), 195–206. https://doi.org/10.3758/CABN.3.3.195

Kim, H., Somerville, L. H., Johnstone, T., Alexander, A. L., & Whalen, P. J. (2003). Inverse amygdala and medial prefrontal cortex responses to surprised faces.

Neuroreport, *14*(18), 2317–2322. https://doi.org/10.1097/01.wnr.0000101520.44335.20

Kimonis, E. R., Frick, P. J., Munoz, L. C., & Aucoin, K. J. (2008). Callous-unemotional traits and the emotional processing of distress cues in detained boys: Testing the moderating role of aggression, exposure to community violence, and histories of abuse. *Development and Psychopathology*, *20*(2), 569–589. https://doi.org/10.1017/S095457940800028X

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The Rupture and Repair of Cooperation in Borderline Personality Disorder. *Science*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Kjær, J. N., Biskin, R., Vestergaard, C. H., & Munk-Jørgensen, P. (2015). A Nationwide Study of Mortality in Patients with Borderline Personality Disorder. *European Psychiatry*, *30*, 202. https://doi.org/10.1016/S0924-9338(15)30162-0

Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, *112*(4), 1251–1288. https://doi.org/10.1162/003355300555475

Koenigsberg, H. W., Siever, L. J., Lee, H., Pizzarello, S., New, A. S., Goodman, M., … Prohovnik, I. (2009). Neural Correlates of Emotion Processing in Borderline Personality Disorder. *Psychiatry Research*, *172*(3), 192–199. https://doi.org/10.1016/j.pscychresns.2008.07.010

Krause, N., & Ellison, C. G. (2003). Forgiveness by God, Forgiveness of Others, and Psychological Well–Being in Late Life. *Journal for the Scientific Study of Religion*, *42*(1), 77–93. https://doi.org/10.1111/1468-5906.00162

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Krettek, J. E., & Price, J. L. (1978). Amygdaloid projections to subcortical structures within the basal forebrain and brainstem in the rat and cat. *The Journal of Comparative Neurology*, *178*(2), 225–253. https://doi.org/10.1002/cne.901780204

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial Construction of a Maladaptive Personality Trait Model and Inventory for DSM-5. *Psychological Medicine*, *42*(9), 1879–1890. https://doi.org/10.1017/S0033291711002674

LaPorta, R., Lopez-de-Silane, F., Shleifer, A., & Vishny, R. W. (1996). *Trust in Large Organizations* (Working Paper No. 5864). https://doi.org/10.3386/w5864

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Lee, V. K., & Harris, L. T. (2014). Sticking with the nice guy: Trait warmth information impairs learning and modulates person perception brain network activity. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(4), 1420–1437. https://doi.org/10.3758/s13415-014-0284-9

Leifer, A. D. (1971). *Children's Responses to Television Violence.* Retrieved from https://eric.ed.gov/?id=ED054596

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect Theory of Mind and Moral Judgment. *Psychological Science*, *17*(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

Linehan, M. M. (1993). *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Guilford Publications.

Lojowska, M., Mulckhuyse, M., Hermans, E. J., & Roelofs, K. (2019). Unconscious processing of coarse visual information during anticipatory threat. *Consciousness and Cognition*, *70*, 50–56. https://doi.org/10.1016/j.concog.2019.01.018

Lowyck, B., Luyten, P., Vanwalleghem, D., Vermote, R., Mayes, L. C., & Crowley, M. J. (2016). What's in a face? Mentalizing in borderline personality disorder based on dynamically changing facial expressions. *Personality Disorders: Theory, Research, and Treatment*, *7*(1), 72–79. https://doi.org/10.1037/per0000144

Malle, B. F. (2011). Attribution Theories: How People Make Sense of Behavior. In *Theories in social psychology: 72-95*.

Maltby, J., Macaskill, A., & Day, L. (2001). Failure to forgive self and others: A replication and extension of the relationship between forgiveness, personality, social

desirability and general health. *Personality and Individual Differences*, *30*(5), 881–885. https://doi.org/10.1016/S0191-8869(00)00080-5

Martijn, C., Spears, R., Van Der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology*, *22*(5), 453–463. https://doi.org/10.1002/ejsp.2420220504

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. https://doi.org/10.3389/fnhum.2011.00039

Mathys, C., Lomakina, E., Daunizeau, J., Iglesias, S., Brodersen, K., Friston, K., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825. https://doi.org/10.3389/fnhum.2014.00825

McCrory, E., Brito, D., A, S., & Viding, E. (2011). The Impact of Childhood Maltreatment: A Review of Neurobiological and Genetic Factors. *Frontiers in Psychiatry*, *2*. https://doi.org/10.3389/fpsyt.2011.00048

McCrory, E., Brito, S. A. D., & Viding, E. (2010). Research Review: The neurobiology and genetics of maltreatment and adversity. *Journal of Child Psychology and Psychiatry*, *51*(10), 1079–1095. https://doi.org/10.1111/j.1469-7610.2010.02271.x

McCullough, M. E. (2000). Forgiveness as Human Strength: Theory, Measurement, and Links to Well-Being. *Journal of Social and Clinical Psychology*, *19*(1), 43–55. https://doi.org/10.1521/jscp.2000.19.1.43

McCullough, M. E. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. John Wiley & Sons.

McCullough, M. E., Pargament, K. I., & Thoresen, C. E. (2000). *Forgiveness: Theory, Research, and Practice*. Guilford Press.

McDonald, M. M., Defever, A. M., & Navarrete, C. D. (2017). Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas. *Evolution and Human Behavior*, *38*(6), 770–778. https://doi.org/10.1016/j.evolhumbehav.2017.06.001

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., … McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *The Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2012). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, nss040. https://doi.org/10.1093/scan/nss040

Moffitt, T. E., & Tank, T. K.-G. 2012 T. (2013). Childhood exposure to violence and lifelong health: Clinical intervention science and stress-biology research join forces. *Development and Psychopathology*, *25*(4pt2), 1619–1634. https://doi.org/10.1017/S0954579413000801

Molander, P. (1985). The Optimal Level of Generosity in a Selfish, Uncertain Environment. *The Journal of Conflict Resolution*, *29*(4), 611–618.

Monahan, K. C., King, K. M., Shulman, E. P., Cauffman, E., & Chassin, L. (2015). The effects of violence exposure on the development of impulse control and future orientation across adolescence and early adulthood: Time-specific and generalized effects in a sample of juvenile offenders. *Development and Psychopathology*, *27*(4pt1), 1267–1283. https://doi.org/10.1017/S0954579414001394

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature*, *383*(6603), 812–815. https://doi.org/10.1038/383812a0

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nelson, R. J., & Trainor, B. C. (2007). Neural mechanisms of aggression. *Nature Reviews Neuroscience*, *8*(7), 536–546. https://doi.org/10.1038/nrn2174

Neumann, C., & Pardini, D. (2012). Factor Structure and Construct Validity of the Self-Report Psychopathy (SRP) Scale and the Youth Psychopathic Traits Inventory (YPI) in Young Men. *Journal of Personality Disorders*, *28*(3), 419–433. https://doi.org/10.1521/pedi_2012_26_063

Neumann, D. L., Fitzgerald, Z. T., Furedy, J. J., & Boyle, G. J. (2007). Sexually dimorphic effects of acute nicotine administration on arousal and visual-spatial ability in non-smoking human volunteers. *Pharmacology Biochemistry and Behavior*, *86*(4), 758–765. https://doi.org/10.1016/j.pbb.2007.03.001

Ng-Mak, D. S., Salzinger, S., Feldman, R. S., & Stueve, C. A. (2004). Pathologic Adaptation to Community Violence Among Inner-City Youth. *American Journal of Orthopsychiatry*, *74*(2), 196–208. https://doi.org/10.1037/0002-9432.74.2.196

Ng-Mak, D. S., Stueve, A., Salzinger, S., & Feldman, R. (2002). Normalization of Violence Among Inner-City Youth: A Formulation for Research. *American Journal of Orthopsychiatry*, *72*(1), 92–101. https://doi.org/10.1037/0002-9432.72.1.92

Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social Judgement in Borderline Personality Disorder. *PLOS ONE*, *8*(11), e73440. https://doi.org/10.1371/journal.pone.0073440

Niedtfeld, I. (2017). Experimental investigation of cognitive and affective empathy in borderline personality disorder: Effects of ambiguity in multimodal social information processing. *Psychiatry Research*, *253*, 58–63. https://doi.org/10.1016/j.psychres.2017.03.037

Noonan, M. P., Kolling, N., Walton, M. E., & Rushworth, M. F. S. (2012). Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. *European Journal of Neuroscience*, *35*(7), 997–1010. https://doi.org/10.1111/j.1460-9568.2012.08023.x

Noonan, M. P., Walton, M. E., Behrens, T. E. J., Sallet, J., Buckley, M. J., & Rushworth, M. F. S. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, 201012246. https://doi.org/10.1073/pnas.1012246107

Noordewier, M. K., Scheepers, D. T., & Hilbert, L. P. (2019). Freezing in response to social threat: A replication. *Psychological Research*. https://doi.org/10.1007/s00426-019-01203-4

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357), 250–253. https://doi.org/10.1038/355250a0

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, *9*(5), 242–249. https://doi.org/10.1016/j.tics.2005.03.010

Ochsner, K. N., Ray, R. R., Hughes, B., McRae, K., Cooper, J. C., Weber, J., … Gross, J. J. (2009). Bottom-Up and Top-Down Processes in Emotion Generation Common and Distinct Neural Mechanisms. *Psychological Science*, *20*(11), 1322–1331. https://doi.org/10.1111/j.1467-9280.2009.02459.x

Öhman, A. (1986). Face the Beast and Fear the Face: Animal and Social Fears as Prototypes for Evolutionary Analyses of Emotion. *Psychophysiology*, *23*(2), 123–145. https://doi.org/10.1111/j.1469-8986.1986.tb00608.x

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

P J Whalen, S. L. R. (1998). Masked Presentations of Emotional Facial Expressions Modulate Amygdala Activity Without Explicit Knowledge. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *18*(1), 411–418.

Pagliaro, S., Brambilla, M., Sacchi, S., D'Angelo, M., & Ellemers, N. (2013). Initial Impressions Determine Behaviours: Morality Predicts the Willingness to Help Newcomers. *Journal of Business Ethics*, *117*(1), 37–44. https://doi.org/10.1007/s10551-012-1508-y

Pearce, S., Scott, L., Attwood, G., Saunders, K., Dean, M., Ridder, R. D., … Crawford, M. (2017). Democratic therapeutic community treatment for personality disorder: Randomised controlled trial. *The British Journal of Psychiatry*, *210*(2), 149–156. https://doi.org/10.1192/bjp.bp.116.184366

Pfohl, B., Blum, N., St. John, D., McCormick, B., Allen, J., & Black, D. W. (2009). Reliability and validity of the borderline evaluation of severity over time (BEST): a self-rated scale to measure severity and change in persons with borderline personality disorder. *Journal of Personality Disorders*, *23*(3), 281–293. https://doi.org/10.1521/pedi.2009.23.3.281

Phelps, E. A. (2006). Emotion and Cognition: Insights from Studies of the Human Amygdala. *Annual Review of Psychology*, *57*(1), 27–53. https://doi.org/10.1146/annurev.psych.56.091103.070234

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion Facilitates Perception and Potentiates the Perceptual Benefits of Attention. *Psychological Science*, *17*(4), 292–299. https://doi.org/10.1111/j.1467-9280.2006.01701.x

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391.

Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social Cognition in Borderline Personality Disorder: Evidence for Disturbed Recognition of the Emotions, Thoughts, and Intentions of others. *Frontiers in Behavioral Neuroscience*, *4*. https://doi.org/10.3389/fnbeh.2010.00182

Quirk, G. J., & Beer, J. S. (2006). Prefrontal involvement in the regulation of emotion: Convergence of rat and human studies. *Current Opinion in Neurobiology*, *16*(6), 723–727. https://doi.org/10.1016/j.conb.2006.07.004

Quirk, G. J., Likhtik, E., Pelletier, J. G., & Paré, D. (2003). Stimulation of Medial Prefrontal Cortex Decreases the Responsiveness of Central Amygdala Output Neurons. *The Journal of Neuroscience*, *23*(25), 8800–8807.

Rand, D. G. (2016). *Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation* (SSRN Scholarly Paper No. ID 2783800). Retrieved from Social Science Research Network website: http://papers.ssrn.com/abstract=2783800

Rand, D. G., Dreber, A., Haque, O. S., Kane, R. J., Nowak, M. A., & Coakley, S. (2014). Religious motivations for cooperation: An experimental investigation using explicit

primes. *Religion, Brain & Behavior*, *4*(1), 31–48. https://doi.org/10.1080/2153599X.2013.775664

Rand, D. G., Ohtsuki, H., & Nowak, M. A. (2009). Direct reciprocity with costly punishment: Generous tit-for-tat prevails. *Journal of Theoretical Biology*, *256*(1), 45–57. https://doi.org/10.1016/j.jtbi.2008.09.015

Raudenbush, D. (2016). *"I Stay by Myself": Social Support, Distrust, and Selective Solidarity Among the Urban Poor*. Presented at the Paper presented at the Sociological Forum. Paper presented at the Sociological Forum.

Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition*, *4*(1), 1–17. https://doi.org/10.1521/soco.1986.4.1.1

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*, 13289. https://doi.org/10.1038/ncomms13289

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology*, *103*(1), 171–173. https://doi.org/10.1037/h0036892

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: Perspectives from human threat of shock studies. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00203

Roelofs, K., Hagenaars, M. A., & Stins, J. (2010). Facing Freeze: Social Threat Induces Bodily Freeze in Humans. *Psychological Science*, *21*(11), 1575–1581. https://doi.org/10.1177/0956797610384746

Rosenberg, S., Nelson, C., & S, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294. https://doi.org/10.1037/h0026086

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. https://doi.org/10.1146/annurev.psych.52.1.59

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Salehi, B., Cordero, M. I., & Sandi, C. (2010). Learning under stress: The inverted-U-shape function revisited. *Learning & Memory*, *17*(10), 522–530. https://doi.org/10.1101/lm.1914110

Sandage, S. J., Long, B., Moen, R., Jankowski, P. J., Worthington, E. L., Wade, N. G., & Rye, M. S. (2015). Forgiveness in the Treatment of Borderline Personality Disorder: A Quasi-Experimental Study. *Journal of Clinical Psychology*, *71*(7), 625–640. https://doi.org/10.1002/jclp.22185

Sansone, R. A., Kelley, A. R., & Forbis, J. S. (2013). The Relationship Between Forgiveness and Borderline Personality Symptomatology. *Journal of Religion and Health*, *52*(3), 974–980. https://doi.org/10.1007/s10943-013-9704-3

Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Schilling, L., Wingenfeld, K., Löwe, B., Moritz, S., Terfehr, K., Köther, U., & Spitzer, C. (2012). Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. *Psychiatry and Clinical Neurosciences*, *66*(4), 322–327. https://doi.org/10.1111/j.1440-1819.2012.02334.x

Schupp, H. T., Ohman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: An ERP analysis. *Emotion (Washington, D.C.)*, *4*(2), 189–200. https://doi.org/10.1037/1528-3542.4.2.189

Selner-O'Hagan, M. B., Kindlon, D. J., Buka, S. L., Raudenbush, S. W., & Earls, F. J. (1998). Assessing Exposure to Violence in Urban Youth. *Journal of Child Psychology and Psychiatry*, *39*(2), 215–224. https://doi.org/10.1111/1469-7610.00315

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. https://doi.org/10.1016/j.neuron.2010.07.020

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750. https://doi.org/10.1038/s41562-018-0425-1

Skowronski, J. J., & Carlston, D. E. (1987). Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology*, *22*(5), 435–452. https://doi.org/10.1002/ejsp.2420220503

Snyder, C. R., & Lopez, S. J. (2001). *Handbook of Positive Psychology*. Oxford University Press.

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stiglmayr, C. E., Grathwol, T., Linehan, M. M., Ihorst, G., Fahrenberg, J., & Bohus, M. (2005). Aversive tension in patients with borderline personality disorder: A computer-based controlled field study. *Acta Psychiatrica Scandinavica*, *111*(5), 372–379. https://doi.org/10.1111/j.1600-0447.2004.00466.x

Stillman, P. E., Van Bavel, J. J., & Cunningham, W. A. (2015). Valence asymmetries in the human amygdala: Task relevance modulates amygdala responses to positive

more than negative affective cues. *Journal of Cognitive Neuroscience*, *27*(4), 842–851. https://doi.org/10.1162/jocn_a_00756

Storbeck, J., & Clore, G. L. (2008). Affective Arousal as Information: How Affective Arousal Influences Judgments, Learning, and Memory. *Social and Personality Psychology Compass*, *2*(5), 1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

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

Thielmann, I., Hilbig, B. E., & Niedtfeld, I. (2014). Willing to Give but Not to Forgive: Borderline Personality Features and Cooperative Behavior. *Journal of Personality Disorders*, *28*(6), 778–795. https://doi.org/10.1521/pedi_2014_28_135

Thome, J., Liebke, L., Bungert, M., Schmahl, C., Domes, G., Bohus, M., & Lis, S. (2016). Confidence in facial emotion recognition in borderline personality disorder. *Personality Disorders*, *7*(2), 159–168. https://doi.org/10.1037/per0000142

Tiedens, L. Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology*, *81*(6), 973–988. https://doi.org/10.1037/0022-3514.81.6.973

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, *39*(6), 549–562. https://doi.org/10.1016/S0022-1031(03)00059-3

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The Prevalence of Personality Disorders in a Community Sample. *Archives of General Psychiatry*, *58*(6), 590–596. https://doi.org/10.1001/archpsyc.58.6.590

Tormala, Z. L., & Rucker, D. D. (2007). Attitude Certainty: A Review of Past Findings and Emerging Perspectives. *Social and Personality Psychology Compass*, *1*(1), 469–492. https://doi.org/10.1111/j.1751-9004.2007.00025.x

Tyrer, P., Nur, U., Crawford, M., Karlsen, S., McLean, C., Rao, B., & Johnson, T. (2005). The Social Functioning Questionnaire: A rapid and robust measure of perceived functioning. *The International Journal of Social Psychiatry*, *51*(3), 265–275.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Unoka, Z., Fogd, D., Füzy, M., & Csukly, G. (2011). Misreading the facial signs: Specific impairments and error patterns in recognition of facial emotions with negative valence in borderline personality disorder. *Psychiatry Research*, *189*(3), 419–425. https://doi.org/10.1016/j.psychres.2011.02.010

Unoka, Z., Seres, I., Áspán, N., Bódi, N., & Kéri, S. (2009). Trust Game Reveals Restricted Interpersonal Transactions in Patients With Borderline Personality Disorder. *Journal of Personality Disorders*, *23*(4), 399–409. https://doi.org/10.1521/pedi.2009.23.4.399

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*, 14637. https://doi.org/10.1038/ncomms14637

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

Valentino, N. A., Hutchings, V. L., Banks, A. J., & Davis, A. K. (2008). Is a Worried Citizen a Good Citizen? Emotions, Political Information Seeking, and Learning via the Internet. *Political Psychology*, *29*(2), 247–273. https://doi.org/10.1111/j.1467-9221.2008.00625.x

van Asselt, A. D. I., Dirksen, C. D., Arntz, A., & Severens, J. L. (2007). The cost of borderline personality disorder: Societal cost of illness in BPD-patients. *European Psychiatry*, *22*(6), 354–361. https://doi.org/10.1016/j.eurpsy.2007.04.001

van ’t Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2014). Spatial Attention, Precision, and Bayesian Inference: A Study of Saccadic Response Speed. *Cerebral Cortex*, *24*(6), 1436–1450. https://doi.org/10.1093/cercor/bhs418

Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., Wright, C. I., & Rauch, S. L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. *Emotion*, *1*(1), 70–83. https://doi.org/10.1037/1528-3542.1.1.70

Whiteley, S. (2004). The Evolution of the Therapeutic Community. *Psychiatric Quarterly*, *75*(3), 233–248. https://doi.org/10.1023/B:PSAQ.0000031794.82674.e8

Wilkinson, G. S. (1993). *WRAT-3: Wide range achievement test administration manual*. Wilmington, Del.: Wide Range, Inc.

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wilson, W. J. (2009). *More Than Just Race: Being Black and Poor in the Inner City (Issues of Our Time)*. W. W. Norton & Company.

Winston, J. S., Strange, B. A., O’Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–283. https://doi.org/10.1038/nn816

Witvliet, C. van O., Ludwig, T. E., & Laan, K. L. V. (2001). Granting Forgiveness or Harboring Grudges: Implications for Emotion, Physiology, and Health. *Psychological Science*, *12*(2), 117–123. https://doi.org/10.1111/1467-9280.00320

Wojciszke, B. (2005). Morality and competence in person- and self-perception. *European Review of Social Psychology*, *16*(1), 155–188. https://doi.org/10.1080/10463280500229619

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the Dominance of Moral Categories in Impression Formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Wojciszke, B., Brycz, H., & Borkenau, P. (1993). Effects of information content and evaluative extremity on positivity and negativity biases. *Journal of Personality and Social Psychology*, *64*(3), 327–335. https://doi.org/10.1037/0022-3514.64.3.327

Wong, R. Y.-M., & Hong, Y.-Y. (2005). Dynamic influences of culture on cooperation in the prisoner's dilemma. *Psychological Science*, *16*(6), 429–434. https://doi.org/10.1111/j.0956-7976.2005.01552.x

Worthington, E. L., & Scherer, M. (2004). Forgiveness is an emotion-focused coping strategy that can reduce health risks and promote health resilience: Theory, review, and hypotheses. *Psychology & Health*, *19*(3), 385–405. https://doi.org/10.1080/0887044042000196674

Wu, J., & Axelrod, R. (1995). How to Cope with Noise in the Iterated Prisoner's Dilemma. *The Journal of Conflict Resolution*, *39*(1), 183–189.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., … Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes*, *120*(2), 260–271. https://doi.org/10.1016/j.obhdp.2012.06.002

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166. https://doi.org/10.1007/BF02249397

Ybarra, O., Chan, E., & Park, D. (2001). Young and Old Adults' Concerns About Morality and Competence. *Motivation and Emotion*, *25*(2), 85–100. https://doi.org/10.1023/A:1010633908298

Yerkes, R., & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482.

Yu, A., & Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. *Advances in Neural Information Processing Systems*. Retrieved from http://discovery.ucl.ac.uk/185399/

Zachary, R. A. (1991). *Shipley Institute of Living Scale*. Los Angeles, CA: WPS, Western Psychological Services.

Zanarini, M. C. (2000). Childhood experiences associated with the development of Borderline Personality Disorder. *Psychiatric Clinics of North America*, *23*(1), 89–101. https://doi.org/10.1016/S0193-953X(05)70145-3

Zanarini, M. C., Frankenburg, F. R., Reich, D. B., & Fitzmaurice, G. (2010). Time to Attainment of Recovery From Borderline Personality Disorder and Stability of Recovery: A 10-year Prospective Follow-Up Study. *American Journal of Psychiatry*, *167*(6), 663–667. https://doi.org/10.1176/appi.ajp.2009.09081130

Zanarini, M. C., Vujanovic, A. A., Parachini, E. A., Boulanger, J. L., Frankenburg, F. R., & Hennen, J. (2003). A screening measure for BPD: The McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD). *Journal of Personality Disorders*, *17*(6), 568–573.

Zapparoli, L., Seghezzi, S., Scifo, P., Zerbi, A., Banfi, G., Tettamanti, M., & Paulesu, E. (2018). Dissecting the neurofunctional bases of intentional action. *Proceedings of the National Academy of Sciences*, 201718891. https://doi.org/10.1073/pnas.1718891115

APPENDIX

Appendix A: MODEL ESTIMATED FINAL HARM AVERSION

		Excluding low accuracy participants				Including all participants			
Study	agent	mean ± SEM	Test Statistic	p-value	effect size (*r*)	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	bad agent	0.322 ± 0.004*	-5.373	<0.001	0.870	0.322 ± 0.004*	-5.442	<0.001	0.871
	good agent	0.681 ± 0.004*				0.681 ± 0.004*			
Study 2	bad agent	0.301 ± 0.004*	-11.074	<0.001	0.867	0.297 ± 0.003*	-13.789	<0.001	0.867
	good agent	0.707 ± 0.003*				0.710 ± 0.002*			
Study 2s	bad agent	0.288 ± 0.002*	-9.347	<0.001	0.868	0.288 ± 0.002*	-9.712	<0.001	0.869
	good agent	0.713 ± 0.002*				0.714 ± 0.002*			
Study 3	bad agent	0.292 ± 0.001*	-10.081	<0.001	0.868	0.293 ± 0.002*	-11.040	<0.001	0.867
	good agent	0.726 ± 0.002*				0.724 ± 0.002*			
Study 4	bad agent	0.288 ± 0.002*	-9.062	<0.001	0.868	0.288 ± 0.002*	-10.192	<0.001	0.858
	good agent	0.715 ± 0.002*				0.714 ± 0.002*			
	low-skill agent	0.714 ± 0.002†	9.186	<0.001	0.868	0.288 ± 0.002†	-10.302	<0.001	0.877
	high-skill agent	0.288 ± 0.002†				0.714 ± 0.002†			
Study 5	bad agent moral	0.290 ± 0.002*	-11.922	<0.001	0.867	0.288 ± 0.002*	-13.233	<0.001	0.867
	good agent moral	0.706 ± 0.002*				0.708 ± 0.002*			
	bad agent skill	0.500 ± 0.004†	0.001	0.991	N.S.	0.502 ± 0.059†	0.528	0.598	N.S.
	good agent skill	0.499 ± 0.004†				0.498 ± 0.004†			
Study 6‡	bad agent	0.329 ± 0.002*	-16.500	<0.001	0.865	0.330 ± 0.002*	-17.471	<0.001	0.916
	good agent	0.667 ± 0.002*				0.667 ± 0.002*			

* higher values denote more money to inflict each additional shock (higher morality)

† higher values denote more time required to score each additional point (lower-skill)

‡ between-subjects, model was only fit for phase 1 (trials before the agent shifted preferences)

Appendix B: MODEL ESTIMATED Ω

		Excluding low accuracy participants				Including all participants			
Study	agent	mean ± SEM	Test Statistic	p-value	effect size (*r*)	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	bad agent	-3.779 ± 0.102	3.212	0.001	0.521	-3.321 ± 0.100	3.321	<0.001	0.532
	good agent	-4.212 ± 0.104				-4.218 ± 0.101			
Study 2	bad agent	-3.411 ± 0.051	6.831	<0.001	0.527	-3.299 ± 0.039	8.367	<0.001	0.526
	good agent	-3.877 ± 0.051				-3.747 ± 0.040			
Study 2s	bad agent	-4.303 ± 0.064	4.416	<0.001	0.410	-4.219 ± 0.067	4.803	<0.001	0.430
	good agent	-4.608 ± 0.060				-4.561 ± 0.061			
Study 3	bad agent	-3.468 ± 0.042	7.296	<0.001	0.628	-3.438 ± 0.038	8.264	<0.001	0.649
	good agent	-3.974 ± 0.043				-3.969 ± 0.038			
Study 4	bad agent	-4.390 ± 0.064	4.219	<0.001	0.404	-4.319 ± 0.059	3.890	<0.001	0.331
	good agent	-4.714 ± 0.048				-4.594 ± 0.050			
	low-skill agent	-4.726 ± 0.047	-0.574	0.566	N.S.	-4.559 ± 0.053	-0.448	0.654	N.S.
	high-skill agent	-4.665 ± 0.057				-4.497 ± 0.063			
Study 5	bad agent moral	-4.116 ± 0.046	5.079	<0.001	0.369	-4.101 ± 0.039	5.937	<0.001	0.369
	good agent moral	-4.428 ± 0.039				-4.401 ± 0.034			
	bad agent skill	-4.224 ± 0.039	3.03	0.002	0.22	-4.192 ± 0.032	2.412	0.016	0.150
	good agent skill	-4.327 ± 0.034				-4.218 ± 0.030			
Study 6‡	bad agent	-3.559 ± 0.042	6.577	<0.001	0.345	-3.563 ± 0.039	7.097	<0.001	0.372
	good agent	-3.928 ± 0.036				-3.391 ± 0.033			

‡ between-subjects, model was only fit for phase 1 (trials before the agent shifted preferences)

N.S. = not significant, SEM = standard error of the mean

Appendix C: MODEL ESTIMATED DECISION NOISE

		Excluding low accuracy participants				Including all participants			
Study	agent	mean ± SEM	Test Statistic	p-value	effect size (*r*)	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	bad agent	0.849 ± 0.062	1.109	0.267	N.S.	0.829 ± 0.064	1.047	0.295	N.S.
	good agent	0.782 ± 0.112				0.766 ± 0.110			
Study 2	bad agent	0.759 ± 0.037	4.651	<0.001	0.364	0.692 ± 0.026	5.439	<0.001	0.342
	good agent	0.574 ± 0.028				0.536 ± 0.022			
Study 2s	bad agent	1.101 ± 0.059	-2.391	.017	0.222	1.073 ± 0.056	-2.044	0.041	0.183
	good agent	1.465 ± 0.119				1.395 ± 0.113			
Study 3	bad agent	0.584 ± 0.034	4.230	<0.001	0.364	0.561 ± 0.030	5.351	<0.001	0.420
	good agent	0.456 ± 0.033				0.408 ± 0.029			
Study 4	bad agent	1.128 ± 0.060	-2.332	0.020	0.223	1.025 ± 0.056	-2.027	0.043	0.173
	good agent	1.482 ± 0.111				1.297 ± 0.095			
	low-skill agent	1.738 ± 0.137	3.852	<0.001	0.364	1.454 ± 0.119	2.831	0.005	0.238
	high-skill agent	1.179 ± 0.061				1.072 ± 0.059			
Study 5	bad agent moral	1.134 ± 0.056	-3.905	<0.001	0.284	1.003 ± 0.046	-3.770	<0.001	0.234
	good agent moral	1.598 ± 0.098				1.359 ± 0.079			
	bad agent skill	1.192 ± 0.056	-0.473	0.636	N.S.	1.065 ± 0.047	-0.578	0.563	N.S.
	good agent skill	1.181 ± 0.050				1.065 ± 0.043			
Study 6‡	bad agent	0.816 ± 0.042	6.178	<0.001	0.324	0.754 ± 0.040	5.825	<0.001	0.288
	good agent	0.536 ± 0.025				0.496 ± 0.024			

‡ between-subjects, model was only fit for phase 1 (trials before the agent shifted preferences)

N.S. = not significant, SEM = standard error of the mean

Appendix D: FINAL SUBJECTIVE RATING

		Excluding low accuracy participants				Including all participants			
Study	agent	mean ± SEM	Test Statistic	p-value	effect size (*r*)	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	bad agent	42.663 ± 4.021*	-5.303	<0.001	0.86	42.702 ± 3.917*	-5.373	<0.001	0.860
	good agent	78.831 ± 2.869*				78.172 ± 2.871*			
Study 2	bad agent	42.227 ± 1.962*	-10.295	<0.001	0.843	40.672 ± 1.609*	-11.892	<0.001	0.748
	good agent	80.706 ± 1.444*				75.688 ± 1.322*			
Study 2s	bad agent	44.112 ± 2.591‡	-7.007	<0.001	0.651	45.096 ± 2.514‡	-6.894	<0.001	0.617
	good agent	72.310 ± 2.220‡				71.464 ± 2.160‡			
Study 3	bad agent	41.104 ± 2.085 *	-9.187	<0.001	0.791	40.877 ± 1.855*	-9.612	<0.001	0.755
	good agent	77.741 ± 1.675*				74.574 ± 1.693*			
Study 4	bad agent	36.688 ± 2.371*	-8.006	<0.001	0.775	40.181 ± 2.226*	-8.128	<0.001	0.692
	good agent	76.211 ± 2.050*				73.515 ± 1.923*			
	low-skill agent	15.231 ± 1.639†	-9.154	<0.001	0.865	19.567 ± 1.813†	-10.135	<0.001	0.854
	high-skill agent	78.330 ± 1.390†				76.773 ± 1.403†			
Study 5	bad agent moral	34.614 ± 1.474*	-11.755	<0.001	0.855	42.216 ± 1.497*	-10.139	<0.011	0.630
	good agent moral	74.111 ± 1.257*				68.220 ± 1.367*			
	bad agent skill	42.593 ± 1.559†	-3.061	0.002	0.223	44.027 ± 1.417†	-2.845	0.004	0.177
	good agent skill	51.042 ± 1.722†				50.761 ± 1.484†			
Study 6§	bad agent	51.169 ± 1.884*	-10.343	<0.001	0.521	50.087 ± 1.779*	-10.659	<0.001	0.559
	good agent	79.094 ± 1.500*				77.649 ± 1.436*			

* final character rating (0 = *nasty*, 1 = *nice*)

† final competence rating (0 = *beginner*, 1 = *expert*)

‡final character rating (0 = *bad*, 1 = *good*)

§ between-subjects

Appendix E: MEAN UNCERTAINTY RATING

		Excluding low accuracy participants				Including all participants			
Study	agent	mean ± SEM	Test Statistic	p-value	effect size (*r*)	mean ± SEM	Test Statistic	p-value	effect size (*r*)
Study 1	bad agent	28.623 ± 2.428	3.444	<0.001	0.559	29.186 ± 2.431	3.586	<0.001	0.574
	good agent	20.612 ± 2.367				20.809 ± 2.314			
Study 2	bad agent	33.078 ± 1.330	7.213	<0.001	0.556	32.592 ± 1.045	5.409	<0.001	0.340
	good agent	24.087 ± 1.371				27.323 ± 1.148			
Study 2s	bad agent	29.209 ± 1.485	3.207	0.001	0.298	29.495 ± 1.426	3.471	<0.001	0.311
	good agent	24.602 ± 1.474				24.783 ± 1.412			
Study 3	bad agent	35.609 ± 1.432	3.896	<0.001	0.335	36.304 ± 1.299	3.927	<0.001	0.309
	good agent	30.858 ± 1.665				32.257 ± 1.518			
Study 4	bad agent	29.335 ± 1.598	3.649	<0.001	0.350	30.269 ± 1.443	3.050	0.002	0.260
	good agent	24.166 ± 1.607				17.776 ± 1.513			
	low-skill agent	18.457 ± 1.227	-1.775	0.076	N.S.	22.514 ± 1.370	-0.872	0.384	N.S.
	high-skill agent	20.653 ± 1.274				23.652 ± 1.412			
Study 5	bad agent moral	27.880 ± 1.019	4.127	<0.001	0.300	29.158 ± 0.940	2.332	0.020	0.145
	good agent moral	24.209 ± 1.027				27.293 ± 0.990			
	bad agent skill	28.875 ± 0.955	2.323	0.020	0.169	30.153 ± 0.892	1.516	0.130	N.S.
	good agent skill	27.277 ± 0.992				29.206 ± 0.932			
Study 6[§]	bad agent	33.584 ± 1.164	4.362	<0.001	0.229	33.942 ± 1.077	3.958	<0.001	0.208
	good agent	26.945 ± 1.347				28.223 ± 1.286			

[§] between-subjects

N.S. = not significant, SEM = standard error of the mean

Appendix F: EXPOSURE TO VIOLENCE (ETV), SUBJECTIVE CHARACTER REGRESSIONS

Subjective impression ratings

	estimate	SEM	t-statistic	p-value
Intercept	0.649	0.059	11.009	<0.001
Trial	0.003	0.003	0.955	0.340
Agent	-1.300	0.075	-17.284	<0.001
ETV	-0.018	0.006	-2.902	0.004
ETV*Agent	0.037	0.009	4.222	<0.001

Subjective impression ratings controlling for age

	estimate	SEM	t-statistic	p-value
Intercept	0.658	0.078	8.459	<0.001
Trial	0.003	0.003	0.955	0.339
Agent	-1.299	0.075	-17.281	<0.001
ETV	-0.018	0.006	-2.904	0.004
Age	0.000	0.001	-0.178	0.859
ETV*Agent	0.037	0.009	4.223	<0.001

Subjective impression ratings controlling for education

	estimate	SEM	t-statistic	p-value
Intercept	0.714	0.110	6.465	<0.001
Trial	0.003	0.003	0.951	0.342
Agent	-1.300	0.075	-17.294	<0.001
ETV	-0.018	0.006	-2.963	0.003
Education	-0.005	0.008	-0.674	0.500
ETV*Agent	0.037	0.009	4.226	<0.001

Subjective impression ratings controlling for psychopathy

	estimate	SEM	t-statistic	p-value
Intercept	0.658	0.068	9.640	<0.001
Trial	0.003	0.003	0.956	0.339
Agent	-1.299	0.075	-17.272	<0.001
ETV	-0.017	0.006	-2.658	0.008
Psychopathy	-0.001	0.002	-0.256	0.798
ETV*Agent	0.037	0.009	4.217	<0.001

Subjective impression ratings controlling for APD

	estimate	SEM	t-statistic	p-value
Intercept	0.646	0.059	10.890	<0.001
Trial	0.003	0.003	0.953	0.341
Agent	-1.299	0.075	-17.275	<0.001
ETV	-0.016	0.007	-2.423	0.015
APD	-0.019	0.033	-0.578	0.563
ETV*Agent	0.037	0.009	4.213	<0.001

Subjective impression ratings controlling for CTQ

	estimate	SEM	t-statistic	p-value
Intercept	0.651	0.066	9.883	<0.001
Trial	0.003	0.003	0.955	0.340
Agent	-1.299	0.075	-17.277	<0.001
ETV	-0.018	0.006	-2.841	0.005
CTQ	0.000	0.001	-0.048	0.962
ETV*Agent	0.037	0.009	4.221	<0.001

Subjective impression ratings controlling for years of incarceration

	estimate	SEM	t-statistic	p-value
Intercept	0.643	0.059	10.885	<0.001
Trial	0.003	0.003	1.169	0.243
Agent	-1.295	0.075	-17.211	<0.001
ETV	-0.017	0.006	-2.641	0.008
Years in prison	-0.001	0.002	-0.430	0.667
ETV*Agent	0.036	0.009	4.084	<0.001

Appendix G: EXPOSURE TO VIOLENCE (ETV), UNCERTAINTY RATING REGRESSIONS

Subjective uncertainty ratings

	estimate	SEM	t-statistic	p-value
Intercept	-0.151	0.061	-2.483	0.013
Trial	-0.018	0.003	-5.969	<0.001
Agent	0.513	0.078	6.605	<0.001
ETV	0.019	0.006	2.982	0.003
ETV*Agent	-0.045	0.009	-4.973	<0.001

Subjective uncertainty ratings controlling for age

	estimate	SEM	t-statistic	p-value
Intercept	-0.194	0.080	-2.410	0.016
Trial	-0.018	0.003	-5.977	<0.001
Agent	0.514	0.078	6.611	<0.001
ETV	0.019	0.006	2.998	0.003
Age	0.001	0.001	0.822	0.411
ETV*Agent	-0.045	0.009	-4.976	<0.001

Subjective uncertainty ratings controlling for education

	estimate	SEM	t-statistic	p-value
Intercept	-0.005	0.114	-0.048	0.962
Trial	-0.017	0.003	-5.961	<0.001
Agent	0.513	0.078	6.599	<0.001
ETV	0.018	0.006	2.838	0.005
Education	-0.012	0.008	-1.506	0.132
ETV*Agent	-0.045	0.009	-4.968	<0.001

Subjective uncertainty ratings controlling for psychopathy

	estimate	SEM	t-statistic	p-value
Intercept	-0.177	0.071	-2.504	0.012
Trial	-0.018	0.003	-5.964	<0.001
Agent	0.513	0.078	6.600	<0.001
ETV	0.017	0.007	2.588	0.010
Psychopathy	0.002	0.002	0.704	0.482
ETV*Agent	-0.045	0.009	-4.969	<0.001

Subjective uncertainty ratings controlling for APD

	estimate	SEM	t-statistic	p-value
Intercept	-0.150	0.061	-2.453	0.014
Trial	-0.018	0.003	-5.972	0.000
Agent	0.513	0.078	6.601	0.000
ETV	0.019	0.007	2.687	0.007
APD	0.004	0.034	0.118	0.906
ETV*Agent	-0.045	0.009	-4.969	0.000

Subjective uncertainty ratings controlling for CTQ

	estimate	SEM	t-statistic	p-value
Intercept	-0.161	0.068	-2.370	0.018
Trial	-0.018	0.003	-5.969	<0.001
Agent	0.514	0.078	6.604	<0.001
ETV	0.019	0.006	2.866	0.004
CTQ	0.000	0.001	0.335	0.737
ETV*Agent	-0.045	0.009	-4.972	<0.001

Subjective uncertainty ratings controlling for years of incarceration on current bid

	estimate	SEM	t-statistic	p-value
Intercept	-0.152	0.061	-2.486	0.013
Trial	-0.018	0.003	-6.087	<0.001
Agent	0.517	0.078	6.632	<0.001
ETV	0.019	0.007	2.829	0.005
Years in prison	0.001	0.002	0.624	0.533
ETV*Agent	-0.045	0.009	-5.007	<0.001

Appendix H: EXPOSURE TO VIOLENCE (ETV), TRUST GAME REGRESSIONS

Trust

	estimate	SEM	t-statistic	p-value
Intercept	69.65048	7.096815	9.814329	<0.001
Agent	-40.1468	10.03641	-4.00012	<0.001
ETV	-1.89661	0.819505	-2.31433	0.022
ETV*Agent	3.079798	1.158956	2.657391	0.008

Trust controlling for age

	estimate	SEM	t-statistic	p-value
Intercept	61.987	9.804	6.323	<0.001
Agent	-40.171	10.033	-4.004	<0.001
ETV	-1.858	0.820	-2.267	0.024
Age	0.210	0.188	1.119	0.264
ETV*Agent	3.087	1.159	2.665	0.008

Trust controlling for education

	estimate	SEM	t-statistic	p-value
Intercept	90.93495	14.24324	6.384429	<0.001
Agent	-39.902	9.972	-4.001	<0.001
ETV	-2.037	0.817	-2.493	0.013
Education	-1.738	1.020	-1.703	0.090
ETV*Agent	3.098	1.152	2.691	0.008

Trust controlling for psychopathy

	estimate	SEM	t-statistic	p-value
Intercept	62.29476	8.487528	7.339565	<0.001
Agent	-40.348	10.087	-4.000	<0.001
ETV	-2.354	0.871	-2.702	0.007
Psychopathy	0.474	0.295	1.611	0.109
ETV*Agent	3.073	1.165	2.638	0.009

Trust controlling for APD

	estimate	SEM	t-statistic	p-value
Intercept	71.098	7.128	9.975	<0.001
Agent	-40.419	10.021	-4.034	<0.001
ETV	-2.526	0.892	-2.831	0.005
APD	7.535	4.357	1.729	0.085
ETV*Agent	3.099	1.157	2.678	0.008

Trust controlling for CTQ

	estimate	SEM	t-statistic	p-value
Intercept	60.349	8.054	7.493	<0.001
Agent	-39.775	9.983	-3.984	<0.001
ETV	-2.296	0.830	-2.767	0.006
CTQ	0.289	0.118	2.444	0.015
ETV*Agent	3.110	1.153	2.698	0.007

Trust controlling for years of incarceration

	estimate	SEM	t-statistic	p-value
Intercept	69.74759	7.168304	9.729999	<0.001
Agent	-0.007	0.278	-0.024	0.981
ETV	-39.988	10.130	-3.947	<0.001
Years in prison	-1.908	0.855	-2.233	0.027
ETV*Agent	3.070	1.174	2.614	0.010

Appendix I: EFFECT OF BPD ON SUBJECTIVE UNCERTAINTY, REGRESSIONS

Subjective uncertainty rating

	Estimate	SEM	t-stat	p-value
Intercept	0.027	0.034	0.795	0.427
Group	0.098	0.057	1.738	0.082
Trial	-0.011	0.001	-10.643	0.000
Agent	0.418	0.032	13.099	0.000
Group*Agent	-0.263	0.080	-3.284	0.001

Subjective uncertainty rating controlling for medication

	Estimate	SEM	t-stat	p-value
Intercept	0.038	0.035	1.074	0.283
Group	0.055	0.070	0.784	0.433
Trial	-0.011	0.001	-10.644	0.000
Agent	0.418	0.032	13.105	0.000
Medication	0.080	0.074	1.084	0.278
Group*Agent	-0.262	0.080	-3.280	0.001

Subjective uncertainty rating controlling for age

	Estimate	SEM	t-stat	p-value
Intercept	0.058	0.064	0.907	0.364
Group	0.100	0.057	1.769	0.077
Trial	-0.011	0.001	-11.040	0.000
Agent	0.421	0.033	12.597	0.000
Age	0.000	0.001	-0.270	0.787
Group*Agent	-0.266	0.080	-3.314	0.001

Subjective uncertainty rating controlling for education

	Estimate	SEM	t-stat	p-value
Intercept	0.036	0.057	0.629	0.529
Group	0.110	0.061	1.801	0.072
Trial	-0.011	0.001	-10.571	0.000
Agent	0.420	0.034	12.512	0.000
Education	0.001	0.017	0.045	0.964
Group*Agent	-0.266	0.086	-3.085	0.002

Subjective uncertainty rating controlling for gender

	Estimate	SEM	t-stat	p-value
Intercept	-0.003	0.109	-0.031	0.976
Group	0.100	0.057	1.754	0.080
Trial	-0.011	0.001	-11.040	0.000
Agent	0.421	0.033	12.596	0.000
Gender	0.025	0.054	0.460	0.646
Group*Agent	-0.266	0.080	-3.316	0.001

Subjective uncertainty rating controlling for prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	0.050	0.048	1.050	0.294
Group	0.091	0.057	1.587	0.113
Trial	-0.011	0.001	-10.645	0.000
Agent	0.418	0.032	13.092	0.000
Prior expectation	0.000	0.001	-0.680	0.497
Group*Agent	-0.261	0.080	-3.265	0.001

Appendix J: EFFECT OF BPD ON LEARNING RATES, REGRESSIONS

Learning rate

	Estimate	SEM	t-stat	p-value
Intercept	0.077	0.018	4.231	<0.001
Group	0.058	0.031	1.880	0.060
Trial	-0.014	0.001	-25.987	<0.001
Agent	0.323	0.017	18.601	<0.001
Group*Agent	-0.167	0.044	-3.827	<0.001

Learning rate controlling for medication

	Estimate	SEM	t-stat	p-value
Intercept	0.077	0.018	4.223	<0.001
Group	0.181	0.038	4.781	<0.001
Trial	-0.014	0.001	-26.053	<0.001
Agent	0.323	0.017	18.638	<0.001
Medication	-0.236	0.040	-5.884	<0.001
Group*Agent	-0.147	0.043	-3.375	0.001

Learning rate controlling for age

	Estimate	SEM	t-stat	p-value
Intercept	0.058	0.064	0.907	0.364
Group	0.100	0.057	1.769	0.077
Trial	-0.011	0.001	-11.040	<0.001
Agent	0.421	0.033	12.597	<0.001
Age	0.000	0.001	-0.270	0.787
Group*Agent	-0.266	0.080	-3.314	0.001

Learning rate controlling for education

	Estimate	SEM	t-stat	p-value
Intercept	0.036	0.057	0.629	0.529
Group	0.110	0.061	1.801	0.072
Trial	-0.011	0.001	-10.571	<0.001
Agent	0.420	0.034	12.512	<0.001
Education	0.001	0.017	0.045	0.964
Group*Agent	-0.266	0.086	-3.085	0.002

Learning rate controlling for gender

	Estimate	SEM	t-stat	p-value
Intercept	-0.003	0.109	-0.031	0.976
Group	0.100	0.057	1.754	0.080
Trial	-0.011	0.001	-11.040	0.000
Agent	0.421	0.033	12.596	0.000
Gender	0.025	0.054	0.460	0.646
Group*Agent	-0.266	0.080	-3.316	0.001

Learning rate controlling for prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	0.156	0.026	6.122	<0.001
Group	0.033	0.031	1.059	0.289
Trial	-0.014	0.001	-25.940	<0.001
Agent	0.320	0.017	18.465	<0.001
Prior expectation	-0.001	0.000	-4.371	<0.001
Group*Agent	-0.161	0.044	-3.692	<0.001

Learning rate fixing prior beliefs to participants' prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	-0.929	0.011	-87.491	<0.001
Group	0.000	0.018	-0.015	0.988
Trial	0.011	0.000	34.932	<0.001
Agent	0.944	0.010	93.534	<0.001
Group*Agent	-0.107	0.025	-4.233	<0.001

Appendix K: EFFECT OF DTC ON SUBJECTIVE UNCERTAINTY RATINGS, REGRESSIONS

Subjective uncertainty rating

	Estimate	SEM	t-stat	p-value
Intercept	0.194	0.065	2.994	0.003
Group	-0.085	0.067	-1.265	0.206
Trial	-0.014	0.002	-8.534	0.000
Agent	0.156	0.070	2.240	0.025
Group*Agent	0.277	0.095	2.904	0.004

Subjective uncertainty rating controlling for medication

	Estimate	SEM	t-stat	p-value
Intercept	0.195	0.071	2.734	0.006
Group	-0.082	0.068	-1.201	0.230
Trial	-0.014	0.002	-8.531	0.000
Agent	0.156	0.070	2.241	0.025
Medication	0.027	0.048	0.551	0.582
Group*Agent	0.277	0.095	2.898	0.004

Subjective uncertainty rating controlling for age

	Estimate	SEM	t-stat	p-value
Intercept	0.264	0.109	2.419	0.016
Group	-0.081	0.068	-1.192	0.233
Trial	-0.014	0.002	-8.511	0.000
Agent	0.158	0.070	2.259	0.024
Age	-0.002	0.002	-0.805	0.421
Group*Agent	0.274	0.095	2.877	0.004

Subjective uncertainty rating controlling for education

	Estimate	SEM	t-stat	p-value
Intercept	0.106	0.101	1.052	0.293
Group	-0.099	0.076	-1.310	0.191
Trial	-0.013	0.002	-6.965	0.000
Agent	0.157	0.078	2.029	0.043
Education	0.030	0.029	1.048	0.295
Group*Agent	0.273	0.107	2.559	0.011

Subjective uncertainty rating controlling for gender

	Estimate	SEM	t-stat	p-value
Intercept	0.042	0.174	0.243	0.808
Group	-0.080	0.068	-1.176	0.240
Trial	-0.014	0.002	-8.516	0.000
Agent	0.156	0.070	2.243	0.025
Gender	0.078	0.083	0.939	0.348
Group*Agent	0.277	0.095	2.909	0.004

Subjective uncertainty rating controlling for prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	0.207	0.075	2.750	0.006
Group	-0.083	0.068	-1.229	0.219
Trial	-0.014	0.002	-8.522	0.000
Agent	0.158	0.070	2.256	0.024
Prior expectation	0.000	0.001	-0.321	0.748
Group*Agent	0.274	0.095	2.872	0.004

Appendix L: EFFECT OF DTC ON LEARNING RATES, REGRESSIONS

Learning rates

	Estimate	SEM	t-stat	p-value
Intercept	-0.016	0.034	-0.483	0.629
Group	-0.031	0.036	-0.870	0.384
Trial	-0.013	0.001	-14.897	<0.001
Agent	0.153	0.037	4.115	<0.001
Group*Agent	0.589	0.051	11.588	<0.001

Learning rates controlling for medication

	Estimate	SEM	t-stat	p-value
Intercept	0.075	0.036	2.053	0.040
Group	-0.051	0.036	-1.427	0.154
Trial	-0.013	0.001	-15.136	<0.001
Agent	0.169	0.037	4.586	<0.001
Medication	-0.180	0.026	-7.050	<0.001
Group*Agent	0.577	0.050	11.441	<0.001

Learning rates controlling for age

	Estimate	SEM	t-stat	p-value
Intercept	0.264	0.109	2.419	0.016
Group	-0.081	0.068	-1.192	0.233
Trial	-0.014	0.002	-8.511	<0.001
Agent	0.158	0.070	2.259	0.024
Age	-0.002	0.002	-0.805	0.421
Group*Agent	0.274	0.095	2.877	0.004

Learning rates controlling for education

	Estimate	SEM	t-stat	p-value
Intercept	0.106	0.101	1.052	0.293
Group	-0.099	0.076	-1.310	0.191
Trial	-0.013	0.002	-6.965	<0.001
Agent	0.157	0.078	2.029	0.043
Education	0.030	0.029	1.048	0.295
Group*Agent	0.273	0.107	2.559	0.011

Learning rates controlling for gender

	Estimate	SEM	t-stat	p-value
Intercept	0.042	0.174	0.243	0.808
Group	-0.080	0.068	-1.176	0.240
Trial	-0.014	0.002	-8.516	<0.001
Agent	0.156	0.070	2.243	0.025
Gender	0.078	0.083	0.939	0.348
Group*Agent	0.277	0.095	2.909	0.004

Learning rates controlling for prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	-0.117	0.039	-2.954	0.003
Group	-0.038	0.036	-1.047	0.295
Trial	-0.013	0.001	-15.005	<0.001
Agent	0.146	0.037	3.953	<0.001
Prior expectation	0.003	0.001	4.968	<0.001
Group*Agent	0.589	0.051	11.632	<0.001

Learning rates fixing prior beliefs to participants' prior expectations

	Estimate	SEM	t-stat	p-value
Intercept	-0.973	0.020	-49.344	<0.001
Group	0.023	0.021	1.122	0.262
Trial	0.011	0.001	21.812	<0.001
Agent	0.875	0.022	40.530	<0.001
Group*Agent	0.363	0.030	12.285	<0.001