WHOM?

THE EFFECT OF TARGET AGENT ON MOTIVATION TO HELP, HURT, OR UNDERSTAND



Cody Kommers

Jesus College

Department of Experimental Psychology

University of Oxford

DPhil
2019 - 2022

A dissertation submitted in partial fulfilment for the degree of…

DOCTOR OF PHILOSOPHY

Prepared under the supervision of…

Dr Matthew J. Apps, Ph.D.

and

Professor Geoffrey Bird, Ph.D.

To My Parents

"'I do feel a distance between myself and those people. I can be lively at times, and they like it in moderation. But I'm not, at least in the areas they respect, a properly serious person…'

—J.I.M. Stewart"

—Jerry Fodor, *Representations*

ABSTRACT

This thesis is about the social psychology of motivation: When do we go out of our way to help an individual other than ourselves? In the studies that follow, I explore how what we know about that other person—whether they are a member of an in-group or out-group, whether we perceive them to be similar to ourselves—modulates our willingness to exert effort on their behalf (i.e., prosocial motivation). I investigate these effects of different target agents on motivation not only for helping behaviour, but also for one's willingness to harm others or to better understand the contents of their minds. I use a combination of experimental and computational methods to address these questions in twelve studies across five empirical chapters.

The studies presented in this thesis can be divided into two parts. In the first part, I investigate the role of prosocial motivation in reward-seeking behaviour. How do we value rewards for others, depending on what we know about who that other person is? How do those valuations change our willingness to exert effort on another's behalf? This line of research represents an advancement on an existing paradigm, the Prosocial Effort Task (Lockwood et al., 2017; Lockwood et al., 2021), to extend findings about prosocial motivation to help an anonymous other—about whom no social identifying information is provided—to various kinds of target agents, such as members of one's in- or out-group. In the second part, I develop my own novel paradigm designed to assess individual differences in accuracy and perceived effort for performing mental state inferences by neurotypical adults. I investigate the possibility of these stimuli for detecting differences in theory of mind ability, depending on what social information is given about the target mind. Are people less willing to exert effort when making inferences about out-group minds? I end with a discussion of possible future directions for a research program investigating this kind of intergroup theory of mind.

Table of Contents

Abstract **5**
Acknowledgments **7**
Chapter 1: General Introduction **9**
Question I: How does social identity affect the cost-benefit analysis of motivation? 11
Question II: What effort costs might be associated with mental state inference? 25
Thesis Outline 35
Chapter 2: Prosocial Motivation as a Function of Group Membership **38**
Study 2.1: Intergroup Prosocial Effort Task 47
Study 2.2: Intergroup Prosocial Effort Task (Replication & Extension) 57
Chapter 3: Prosocial Motivation as a Function of Perceived Similarity **73**
Study 3.1: Perceived Similarity Correlations in an Independent Dataset 80
Study 3.2: Manipulation of Perceived Similarity in Prosocial Effort Task 88
Study 3.3: Manipulation of Perceived Similarity (Increased Difference) 99
Chapter 4: Antisocial Motivation as a Function of Group Membership **108**
Study 4.1: Inverse-Rewards Prosocial Effort Task 116
Study 4.2: Inverse-Rewards Prosocial Effort Task (Replication & Extension) 125
Chapter 5: A Novel Assessment of Effort and Accuracy in Theory of Mind **133**
Study 5.1: Theory of Mind Stimuli Assessment 144
Study 5.2: Theory of Mind Stimuli Test-Retest Reliability Assessment 155
Chapter 6: Effortful Theory of Mind as a Function of Group Membership **167**
Study 6.1: Prosocial Effort Task with Theory of Mind Stimuli 175
Study 6.2: Theory of Mind Assessment with "Intergroup" Stimuli 183
Study 6.3: Prosocial Effort Task with "Intergroup" Theory of Mind Stimuli 190
Chapter 7: General Discussion **205**
Summary of Findings and Contributions 205
Evaluating Overall Questions of Interest 208
The Intuitive Anthropologist Framework 225
Future Directions 229
References **236**

ACKNOWLEDGMENTS

Thank you, Matt and Geoff. I couldn't have done this without you. I know because I tried. I didn't get very far. Despite the fact that I insinuated myself into both of your labs under circumstances which elicited less than your full enthusiasm, you lent your full advisorial support to my research. That you did so speaks both to your own good nature, as well as the good nature of the British people.

Thanks to my friends and colleagues at Oxford, especially those at Jesus College. Special thanks to Maja and Jonas; you were my colleagues and friends in the department since day one. Aleksandra and Trinh, you agreed to care for Zora in a capacity that afforded me many happy months of procrastination on this work. Thank you for being a part of our household.

Thanks to Jesse Rissman at UCLA and Samuel Gershman at Harvard. Both of you supported my career in ways which profoundly impacted the direction of my life and work. I'll always be grateful for the opportunities you gave me.

Thanks to MG, BB, AB, OS, BR, and LM. You taught me that that ideas and words matter. Not necessarily these ideas, or these words. But some do. I hope one day I'll find some that mean half as much to someone else as yours did to me.

Thanks to CG, GA, MG, JB, TP, and everyone at the former Department of Social Relations. Your legacy kept me going when I'd forgotten why I started. Thanks also to everyone who took the time to talk with me on my podcast, Cognitive Revolution. You gave me a sense of community when community was hard to come by.

Thank you to my parents, Kim and Ed—this thesis is dedicated to you. Without your emotional, genetic, financial, and every other conceivable kind of support, I would never have

been able to pursue my interests and passions this far. The opportunity to have done so is a greater gift than any individual deserves.

Thank you, above all, to my partner, Haily. No one should have to listen to every thought that comes into my head. But you do, and I'm thankful for that every day. Without you, I'd never have been able to do any of this. Without you, I doubt I would've wanted to. I look forward to what life will bring us, especially once there are no more degrees to be had.

Speaking of degrees: thanks, finally, to everyone who told me not to do a PhD.

I get it now.

CHAPTER 1: GENERAL INTRODUCTION

In the 1960s, the neurologist Oliver Sacks treated a population of patients who had suffered from *encephalitis lethargica*, a sleeping-sickness epidemic of the early twentieth century. These patients, described in Sacks's (1973) *Awakenings*, exhibited a curious slate of symptoms in the aftermath of their encephalitic state. For instance, it was not uncommon for many of Sacks's patients to stand in the hall, completely motionless, unable to move for hours on end. Characterised by an inability to initiate motor action, Sacks described their illness as a sort of Parkinsonism. It was not precisely the same disease illustrated in the 1800s by Parkinson and Charcot—for instance, the post-encephalitic Parkinsonism could affect patients of any age—but the overlap in symptoms was conspicuous. As each of these physicians noted in turn, their patients were left not only with an inability to move but without even the impulse to do so in the first place. Parkinson's Disease is, in this sense, fundamentally a pathology of motivation.

The story of Sacks's patients is that upon treatment with the dopamine precursor levodopa the patients experienced a shocking reanimation of their impulse to engage in activity. They went from stillness to alacrity—exuberant caricatures of their pre-encephalitic selves. It is an illustration of the importance of motivation in human behaviour. Not only is motivation necessary for successful reward-seeking behaviour, but it is necessary for any behaviour at all, from getting out of bed in the morning, to speaking, to walking down the hall. Sacks's post-encephalitic patients are an example of what behaviour looks like devoid of motivation at a fundamental, neural level.

Motivation is the impetus for an agent to perform one course of goal-directed action over another (Daw & Shohamy, 2008; Harmon-Jones & van Honk, 2012). It is typically defined as the

result of a cost-benefit analysis: an agent is motivated to take an action when the benefits of that action outweigh the associated costs (Lockwood et al., 2017). In this standard account, sufficiently well-motivated actions will result in an agent exerting the requisite amount of effort necessary to perform the action. Research investigating an egocentric conception of motivation is premised on the idea that the benefits of an action consist primarily of rewards obtained by the acting agent (Simpson & Balsam, 2015). But humans also engage in behaviour motivated by an allocentric desire to benefit others, often known as prosociality or altruism (Batson & Powell, 2003; Pfattheicher, Nielsen, & Thielmann, 2022). In the face of an opportunity to help one another out, humans do not stand inert and motionless, like Sacks's patients, unable to summon the motivation to act. Humans can, for example, be seen to hold doors open for passing strangers, to shovel snow from the walkway of an elderly neighbour, to prepare meals for conspecifics even when they do not share an especially large amount of the cook's genetic material, and to devote precious cognitive resources to editing one another's manuscripts. Even when there is no ostensible benefit to the acting agent, the motivation to enact a particular behaviour can still be present. People attribute some value to rewards earned by agents other than themselves. This observation leaves us with the problem of *prosocial motivation*: Under what conditions are people willing to exert effort on behalf of another agent?

This thesis is about an individual's motivation to help, hurt, or understand another person, and how that motivation changes as a function of the information the individual has about who that other person is. In this Introduction, I will describe existing paradigms and findings which frame the two overarching questions which I believe are at the core of the research presented within this thesis. I will end with a thesis outline, briefly summarising the contributions of each chapter and the connections between them.

Question I: How does social identity affect the cost-benefit analysis of motivation?

One of the most common forms of cost associated with any action is the exertion of effort (Kool & Botvinick, 2018; Lockwood et al., 2017). It has long been hypothesised that humans are naturally effort averse, an idea often referred to as the law of least effort (Hull, 1943; Zipf, 1949). This "law" states that if all other costs and benefits are held constant between two possible actions, people will tend to choose the one that requires less effort.

From this basic insight comes the contemporary psychological and neuroscientific study of motivation (Fiske & Taylor, 1991; Daw & Shohamy, 2008; Harmon-Jones & van Honk, 2012; Yee & Braver, 2018; Shenhav, Prater Fahey, & Grahek, 2021). If human actions minimise cost, our psychological mechanisms must have some way of calculating the estimated cost of a potential action and comparing it to the estimated value of anticipated rewards. Motivation, then, is the impetus to engage in one course of goal-directed action over another, based on this kind of cost-benefit analysis of enacting a particular course of behaviour (Daw & Shohamy, 2008; Harmon-Jones & van Honk, 2012; Lockwood et al., 2017). Accordingly, a key question of motivation is: how exactly are the costs of an action—in particular its effort costs—weighed against its potential benefits?

The cost of effort in motivation.

Motivation to exert physical effort has two critical components: the result of the cost-benefit analysis of the potential action and the actual carrying out of the intended sequence of motor movements (Manohar et al., 2015; Ang, Manohar, & Apps, 2015; Lockwood et al., 2021; Dunn, Inzlicht, & Risko, 2019). In this thesis, I will measure motivation by presenting participants with a choice between two possible courses of actions, each with a different set of a

costs and rewards. In each trial a participant will choose whether to engage in the high- or low-effort option, then will perform the associated effort task of their choosing. By analysing how participants' choices change across the different permutations of associated costs and rewards, I will be able to measure how a participant's motivation changes as a function of the choices they are presented with. These two stages fulfil both requirements in the definition of motivation provided above: participants perform the cost-benefit analysis, then carry out the associated motor action.

This dissociation between the choice to engage in an action and actually seeing it through is a standard approach in the scientific literature on motivation (e.g., Dunn, Inzlicht, & Risko, 2019; Lockwood et al., 2021) and is a key distinction between paradigms meant to measure motivation versus other, more general decision-making paradigms. While many experiments present participants with a cost-benefit trade-off, most paradigms do not require an effortful response from the participant. Rather, the costs associated with a choice are often monetary—for example, in economic games such as the Ultimatum Game (Güth, Schmittberger, & Schwarze, 1982; Thaler, 1988). In an Ultimatum Game, an individual makes a proposal about how to split a sum of money between themselves and another person. The other person then chooses whether to accept the proposed split or to decline the proposal, in which case each agent gets nothing. From the proposer's perspective, the benefit is monetary: How much can the proposer allocate to themselves while expecting the decider to accept the split? For the proposer, the cost is either monetary (proposing a split that the decider rejects, resulting in no monetary reward) or social (proposing a split that is unfair or otherwise socially undesirable; Nowak, Page, & Sigmund, 2000). While these kind of cost-benefit analyses have their own strengths, they cannot tell us about how the decision-making dynamics underlying people's willingness to exert effort to earn

reward—either for themselves or another agent. If costs and benefits are in the same currency, it is difficult to distinguish between the cost and benefit in any meaningful sense; there is only the outcome. Furthermore, such economic games rarely, if ever, require the participant to exert effortful motor movement, beyond what is required to indicate their choice.

There are three primary reasons why the study of prosocial behaviour can benefit from consideration of physical effort costs in addition to economic costs. First, while financial costs are objective in the sense that they map value onto concrete numbers, the perceived value of, say, £20 depends on the person who is in possession of it. The amount may be trivial to someone from a wealthy background, but important to someone just trying to make ends meet. Effort, by contrast, can more readily be calibrated to individual capacity (Lockwood et al., 2017). Accounting for a potential discrepancy in willingness to exert effort versus give up money may be important because many every day prosocial acts come at an effortful, but not a financial, cost (Inzlicht & Hutcherson, 2017). Second, when costs and benefits are in the same currency (i.e., money), it difficult to distinguish between self- and other-regarding preferences. That is, more money for the other person often (though not always) equates to less money for oneself. By separating the currency of benefit (financial) and cost (effort), experiments can show when earning rewards for others varies as a function of different levels of cost to the actor. And third, it is plausible that participants might endorse a prosocial decision but not follow through on it. Explanations for this could range from virtue signalling to insufficiently high valuation of the expected outcome. The study of effortful costs does not need to supersede the study of economic costs. But it does add a crucial additional element to the story of prosociality.

Another kind of cost often studied in decision-making experiments is moral (e.g., Crockett et al., 2014; Crockett et al., 2015). While research based on economic games tends to

present humans as egocentric reward-maximisers, these morality-based studies are often consistent with a more prosocial disposition. In one experiment, participants were presented with an opportunity to either receive a painful electric shock or to pay a given amount of money to forgo the shock (Crockett et al., 2014). In different trials, the shock would either be applied to the participant or to another person. People showed a consistent willingness to pay more to forgo electric shocks applied to others than shocks applied to themselves. That is, they valued the cost of other's pain as higher than their own. Crockett and colleagues argued that this effect was due to participants' disinclination to violate moral norms associated with harming others. In short, people evaluate moral and monetary costs according to very different rubrics. Any other kind of cost—such as effort—would thus require its own study of its effect on motivation.

In this thesis, I will use effort-based paradigms to study the cost-benefit analysis of motivation, where the result of this analysis is the subjective value of a reward. I define subjective value as the value of the reward itself minus the cost of obtaining it (Lockwood et al., 2017; Lockwood et al., 2021). I will formalise this definition with a computational model in Chapter 2, but for the purposes of this Introduction I will simply describe the relevant concepts as they are related in the existing scientific literature.

A key concept in the literature on subjective value is "discounting" (Ainslie & Haslam, 1992). For example, if a reward of £1 is presented to an agent unexpectedly, immediately, and without any cost, then the subjective value of that reward would correspond closely to its objective value of one pound. However, if the effort cost of obtaining this reward is high, then the subjective value of the reward diminishes as a function of the effort required; this is known as effort discounting (Botvinick, Huffstetler, & McGuire, 2009; Hartmann et al., 2015; Westbrook, Kester, & Braver, 2013). Another common form of diminished subjective value is

temporal discounting: the subjective value of a future reward is perceived as less than a reward obtained in the present (Kobayashi & Schultz, 2008; Hamid et al., 2016; Ruggeri et al., 2022). Therefore, in assessing the perceived effort costs associated with earning a reward, I will define these effort costs as effort discounting rates.

Though the work presented in this thesis does not extend into brain imaging, there is a large literature showing that these kind of discounting rates and the resulting output of subjective valuation calculation are tracked in specific brain regions. The neural substrate primarily associated with carrying out such a cost-benefit analysis is the dopamine system, specifically via dopamine neurons in the midbrain (Schultz, Dayan, Montague, 1997). Dopamine plays many roles in neural functions related to motivation (Gershman, 2017), which include the encoding of both prediction errors associated with earned rewards (Daw & Shohamy, 2003) and the predicted effort cost associated with a potential action (Varazzani, et al., 2015). Dopamine also appears to be crucial in motivating the response vigour of an action (Niv et al., 2007; Beierholm et al., 2013) and in overcoming the predicted effort costs of future behaviours (Salamone & Correa, 2002). In other words, dopamine plays a key role both in computing the cost-benefit trade-off associated with a potential action, as well as translating the result of that analysis into concrete action. Many psychiatric conditions—such as schizophrenia (McCutcheon, Abi-Dargham, & Howes, 2019), depression (Whitton et al., 2020), and Parkinson's (Meder et al., 2019; McGuigan et al., 2019)—can be considered pathologies of dopaminergic, and thereby motivational, function.

It is also well-established that the anterior cingulate cortex plays a critical role in computing the cost-benefit integration associated with an action (Hillman and Bilkey, 2010; Hosokawa et al., 2013; Shenhav, Botvinick, & Cohen, 2013), and that lesions to this area impair

an individual's ability to make choices which require evaluating trade-offs between reward and effortful action (Walton et al., 2003; Walton et al., 2006). There is also evidence that, prior to this integration, there are dissociable neural signals for calculating effort costs and potential rewards (Klein-Flügge et al., 2016). Klein-Flügge and colleagues presented participants with decisions in which each trial varied in effort and reward level. Using functional magnetic resonance imaging, they showed that engagement in reward-motivated behaviour (choosing to earn rewards regardless of the required effort level) recruited a different brain area than effort-aversive behaviour (choosing to forgo high effort levels regardless of potential reward). Reward seeking was associated with activation in the ventromedial prefrontal cortex, while effort avoidance as associated with the supplementary motor area and putamen. Taken together, these studies show that the evaluation of effort costs is clearly distinct from other elements in the brain's cost-benefit analysis of potential behaviours, and that costs associated with effort are dissociable, in both behavioural and neural terms, from other kinds of costs.

Physical effort associated with motor actions is not the only kind of effort that needs to be considered in a cost-benefit analysis. Cognition is also often thought of as effortful, though it is an empirical question to what extent the costs of cognitive and physical effort have the same effects on motivating behaviour (Kool & Botvinick, 2018). Reward devaluation by both cognitive and physical effort has been shown to recruit a common network of brain areas (Chong et al., 2017), and in theoretical work it is often useful to treat cognitive effort costs in the same manner as economic costs (Westbrook & Braver., 2015). However, a recent meta-analytic study has opened up these assertions for debate (Lopez-Gamundi et al., 2021). It may be the case that physical and cognitive effort are closely linked—both by association with the same brain regions, and by being treated as a cost—but that they rely on distinct computations of value.

Preliminary evidence suggests that cognitive effort is computed as hyperbolical discount rate, while physical effort is computed as a parabolic discount rate; furthermore, some people may treat one type of effort as more costly than another (Chong et al., 2018).

While much work has supported the claim that cognitive effort is aversive (Kurzban, 2016; Yee et al., 2021), other studies have shown that cognitive effort can be intrinsically rewarding (Tanaka, O'Doherty, & Sakagami, 2019; Clay et al., 2022; Müller, Husain, & Apps, 2022). Inzlicht, Shenhav, and Olivola (2018) refer to this as the effort paradox: the observation that effort can be both costly and valued in different contexts. This observation is echoed in the need-for-cognition literature (Cacioppo & Petty, 1982; Petty et al., 2009).

In summary, successfully motivating human behaviour often requires overcoming an effort cost. The costs associated with effort are dissociable—behaviourally and neurally—from other kinds of costs, such as economic costs. While effort costs are, in general, a crucial component of motivational cost-benefit analysis, the specific calculation of those costs depends on many factors: whether the effort is cognitive or physical, whether the exertion is perceived as a cost, and how it is integrated with other value signals, such as reward. These considerations can all be subsumed under the concept of "discounting"—with different discount rates quantifying the effort costs associated with different scenarios. One kind of discounting, which will be essential to the work presented in this thesis, is devaluation of rewards by effort depending on the agent to whom that reward is going, known as social discounting (Jones & Rachlin, 2006). For example, people have been shown to more heavily discounting rewards for others over oneself (Lockwood et al., 2017) and strangers over family or close friends (Ostaszewski & Osiński, 2011). How does the subjective value of a reward change based on the agent to whom it is going?

The value of rewards for other agents.

In this thesis, I will seek to make contact between the literature on motivation outlined above and the literature on intergroup processes and interactions (Allport, 1954; Hewstone, Rubin, & Willis, 2002; Cikara & Van Bavel, 2014). In particular, I will focus on the flexible representations of "us" and "them" associated with social identity theory (Tajfel et al., 1979; Turner et al., 1987). Social identity theory suggests that people can flexibly view others either as individuals or as representatives of a social group, depending on the circumstances of the interaction. How these groups are segmented is also flexible; people can be led to see someone as a member of either their social in-group or out-group, depending on contextual clues. As I will outline the relevant chapter introductions, most existing paradigms for studying prosocial motivation have yet to look at the effects of social identity on motivation. In the experiments throughout Chapters 2 through 4, I will investigate how different aspects of social identity (e.g., whether another agent is perceived as an in- or out-group member) affect the willingness to exert effort for another individual in different circumstances (e.g., whether rewards are framed as gains or losses for the other individual). In the work presented throughout this thesis, I rely primarily on the minimal groups paradigm to elicit intergroup effects (Tajfel et al., 1971)—for which I will discuss the strengths and weaknesses of this approach in each of the relevant chapters, as the trade-off dynamics may vary based the specific question addressed in a given study.

There is a large scientific literature showing that humans are more likely to help members of their own social groups than they are members of other groups, an effect usually known as parochial altruism (Bernhard, Fischbacher, & Fehr, 2006; Everett, Faber, & Crockett, 2015). Parochial altruism has been argued to be central to human cooperation, as it incentivises us to work with social groups among whom we likely share common goals and interests (Bernhard,

Fischbacher, & Fehr, 2006). Indeed, a recent study showed that parochial altruism was observable at the national level across 42 countries around the world (Romano et al., 2021). At a neurochemical level, the neuropeptide oxytocin has been linked with altruistic behaviour specifically for in-group, but not out-group, members (De Dreu et al., 2010). Theoretical work has shown how parochial altruism can arise among agents operating under basic interaction principles, most notably with increased opportunity to engage in out-group punishment (Gao et al., 2015). Meanwhile, highly ecologically valid work has shown how supporters of one football team are more motivated to earn rewards for supporters of that same team versus individuals who did not support that team—with distinct hemodynamic responses for in- and out-group choices (Bortolini et al., 2017).

Though the roots of parochial altruism run deep, viewed in another light its effects are also quite fragile. Parochial altruism appears to be sensitive to intruding information about other groups to which people might belong, or even information about the individual themselves. For example, a recent experimental paradigm showed that most participants were willing to "rob" a large group of victims to earn rewards for themselves (Alós-Ferrer, García-Segarra, & Ritschel, 2021). However, participants were reluctant to perform a similar form of robbery on a single individual from that group. The authors argue that this is evidence of how humans can be both prosocially disposed toward individuals, while harbouring malevolent feelings toward the abstract group of which they are a part (a related phenomenon is subtyping, in which people consider an individual a subtype of their stereotyped group and therefore do not attribute positive traits of the individual to the group as a whole; Richards & Hewstone, 2001). Individuating information can also be used to reduce prosociality toward in-group members (Lee et al., 2021). The authors of this series of experiments show how as people come to learn more about in-group

members as individuals, they are less likely to apply the same decision strategy that they would for generic members of their in-group. Synthesising from these and similar studies, De Dreu and colleagues (2022) argue that the neurochemical and sociocultural mechanisms underlying prosocial behaviour for one's in-group also lead, as a by-product, to intergroup conflict. Together, these studies show how the subjective value of a reward for another agent can vary widely depending on how the social identity of that other agent is perceived by the participant, with parochial altruism being a signal example of when people place a higher relative value on positive outcomes for in-group members, an effect typically known as in-group favouritism (Everett, Faber, & Crockett, 2015; Moscatelli & Rubini, 2017; Amira, Wright, & Goya-Tocchetto, 2021).

A key question concerning in-group favouritism is the mechanism that drives it. Do people favour in-group over out-group members because they are increasing their valuation of rewards for in-group members, or because they are decreasing their valuation of rewards for out-group members? It is reasonable to imagine that both possibilities can be brought out to various degrees depending on the situation, and so it is of potential importance to understand the factors under which each is likely to occur. More generally, the effects of in-group favouritism can be ameliorated by a community's specific social norms (Kok et al., 2020). Similarly, moral suasion can disrupt in-group favouritism by appealing to participants to behave in a way they believe to be right (Bilancini et al., 2020). Introducing conflicting incentives within a group can also ameliorate in-group prosociality (Zuo, Chen, & Zhao, 2015). Recent neuroimaging evidence also showed that participants exhibited different hemodynamic responses in the face of violations of prosocial expectations by in-group versus out-group members (Mei, Zhang, & Yin, 2020). This line of empirical work is consistent with theoretical work showing how varying parameters of

social interactions, such as inequality between groups, can lead to increases or decreases in in-group favouritism (Durrheim et al., 2016). What drives these fluctuations in overall in-group favouritism? The variation in subjective valuation of rewards for in- vs out-group members in different situations has proved difficult to pin down.

One notable program of research exploring the increased valuation of negative outcomes for out-group members is intergroup schadenfreude (Cikara, Bruneau, & Saxe, 2011; Cikara, 2015; Cikara & Fiske, 2012; Cikara & Fiske, 2013; Cikara et al., 2014). Cikara and Fiske (2012; 2013) argue that while prosocial behaviour toward in-group members is motivated largely by empathic concern (in which we share in the affective responses to their positive or negative rewards), people's empathic concern for out-group members is highly attenuated, or in some cases non-existent. Under certain circumstances, such as competition for limited resources, this creates a counter-empathic response in which the valence of one's own subjective value of rewards received by out-group is flipped, with negative rewards for out-group members perceived as positive. In particular, Cikara and colleagues investigate these effects of intergroup schadenfreude from the perspective of the stereotype content model (Cuddy, Fiske, & Glick, 2008)—which predicts that stereotypes high in the dimension of competence (and are thereby enviable rather than pitiable) are most likely to elicit schadenfreude. In one series of experiments, the authors' explicit goal was to investigate intergroup schadenfreude in novel minimal groups, rather than existing real-world groups (Cikara et al., 2014). I will describe and analyse two of these experiments in-depth, as they are concerned with many of the same questions I intend to address in the work presented in this thesis.

In the first experiment, Cikara and colleagues (2014) investigated the extent to which competitive, cooperative, or neutral scenarios influenced feelings of schadenfreude. Participants

were presented with a series of brief vignettes ("Liane ate a really good sandwich" or "Andrew sat in gum on a park bench") and asked to rate the magnitude of their positive or negative response either "How bad does this make you feel?" or "How good does this make you feel?" Participants were assigned to teams randomly to either the Rattlers or the Eagles, though they were told that it was based on the results of a five-item personality questionnaire. In each trial, participants were presented with: a vignette, the positive or negative response magnitude question (measured via a slider), and a group label for the individual in the vignette (e.g., "Andrew is on the RATTLERS"). Participants were told that the experiment was a teams-based problem-solving challenge. They were told they would be given 17 of these vignettes—which were of interest because "Scientific evidence suggests that the more people know about other players' personal experiences, the better people perform in these particular problem solving challenges"—before doing the problem solving challenge (which they were never, in fact, asked to complete). In the competitive scenario, participants were told that whichever team completed one-hundred problem solving tasks (with each participant, recruited via Amazon Mechanical Turk, only completing a single one) would double the money they earned for participating, from $1 to $2. In the cooperative scenario, participants were told when both teams completed two hundred tasks in total, both teams would double their money. In the neutral scenario, whenever a team reached one hundred, the members of those teams would double their money (independent of the other team's performance). Cikara and colleagues found that across all conditions, the competitive scenario elicited the largest intergroup counter-empathic responses. In the competitive scenario, participants reported more positive feelings associated with negative outcomes for out-group members—as well as more negative feelings associated with positive outcomes for out-groups—than in the cooperative or neutral conditions. Across all conditions,

participants showed no difference in their feelings toward in-group outcomes between competitive and non-competitive scenarios.

In a second experiment, Cikara and colleagues (2014) introduce a set of six extra vignettes, which describe outcomes for individuals who are associated with neither the Eagles nor the Rattlers. The presence of this non-group-identified control allows the authors to ask whether differences in participants' responses are due to especially empathic responses for in-group members or especially counter-empathic responses for out-group members, when compared with a baseline. They found that participants' responses to the non-group-identified agents resembled their responses for in-group members, with significantly less counter-empathic feelings than those directed toward out-group members (though only for schadenfreude, and not for its inverse: negative feelings toward an out-group's positive outcome). This offers one means of investigating the mechanism driving in-group favouritism, posed earlier in this section.

These experiments are aligned with the work I will present in this thesis (particularly in Chapters 2 and 4), in that they seek to investigate the circumstances under which minimal groups can lead to differential valuations of outcomes experienced by members of one's in-group versus one's out-group. However, there are several potential limitations to this experiment, which I will seek to address in my own experimental paradigms. One limitation is that there is no objective value associated with the other agent's experience, as presented in the vignette. The stories are surely representative of stereotypically enjoyable or unenjoyable experiences, but the magnitude of the experience of the hypothetical agent remains vaguely defined; it is therefore difficult to say precisely how changes in these outcomes would affect the participants valuations of them. For example, what is the relative value of eating a good sandwich versus sitting in gum? Does it depend on whether Liane especially likes sandwiches or whether Andrew was wearing his

favourite pair of trousers? Without a more concrete measure of the hypothetical agents' own affective states, it is difficult to know. A second limitation is that there are no costs, effortful or otherwise, associated with these self-reported value judgments. Participants self-reported the magnitude of the feelings associated with a particular outcome, but what does that endorsement mean? Would participants give up some of their money to actually cause such an outcome to happen? Would they willingly exert effort to make the outcome in question more likely to occur? Would they be willing to incur the social cost of others judging them for their antagonistic feelings of schadenfreude? Understanding how these potentially counter-empathic feelings of schadenfreude motivate actual behaviour, it would be necessary to investigate them in the context of a particular cost. A third, and slightly more nuanced, limitation is the indirect connection between the competitive scenarios and the potentially counter-empathic response. Participants are ostensibly competing in a problem-solving task and are told that considering the other agents' feelings will help them solve the relevant problems. The authors use this to argue that the mere implication of a competitive scenario is enough to elicit counter-empathic responses to negative out-group outcomes, and that in the absence of these implications such responses are less likely. While the implication does appear to make a difference, the tenuous connection makes it difficult to generalise from: in many real-life competitive team scenarios, one would feel positive emotions at an out-group's loss of reward because it implies an attendant gain in reward for one's own team. The matter is further complicated by the fact that a full quarter of the 320 participants initially sampled failed an attention check asking whether or not the teams were in direct competition, with almost all of the incorrect answers coming from participants in the non-competitive scenarios mistakenly believing they were in a competition. The authors do not report whether those participants who mistakenly believed they were in a

competition also showed increases in counter-empathic responses. In sum, it is clear from this study that different intergroup circumstances can elicit counter-empathic responses toward out-groups—but it is difficult to understand the nuanced dynamics of subjective value associated with others' outcomes, how these feelings might motivate actual behaviour, and the precise role that competition might play in intergroup schadenfreude. Finally, one possible alternative interpretation of these data—and one that I will investigate a version of in my own experiments—is that empathy itself incurs an effort cost (Cameron et al., 2019) and people are less willing engage in it because they are less willing to devote prosocial effort more generally toward out-group members.

In this thesis, I will seek to draw a similar connection between prosociality, intergroup processes, and one's understanding about the content of a target agent's mind. Instead of focusing on empathy, as in studies by Cikara and colleagues described above, I will focus on theory of mind. Many mental processes can be cognitively effortful, and while previous research has primarily investigated the mental effort costs associated with cognitive control (Kool & Botvinick, 2018), I will explore the possible effort costs associated with theory of mind. In the next section, I will introduce existing research on theory of mind, with a focus on why it might be hypothesised that mental state inference incurs an effort cost, and in particular why this effort cost might be greatest for members of one's out-group.

Question II: What effort costs might be associated with mental state inference?

Human social behaviour requires us to understand the mental states and intentional actions of our conspecifics (Tomasello, 1999). Much research over the past half century has framed this process as a form of "intuitive psychology"—suggesting that when people make

everyday inferences about the mental states of others, they are doing an informal version of what academic psychologists do formally (Ross, 1977). The mental process underlying this capacity is often referred to as "theory of mind," defined as the ability to impute mental states to oneself and others. It is a concept which has played an important role in understanding the social-cognitive abilities of chimpanzees (Premack & Woodruff, 1978), individuals with Autism Spectrum Disorder (Baron-Cohen, Leslie, & Frith, 1985), infants (Gopnik & Wellmann, 1992), and neurotypical adults (Frith & Frith, 2005).

The specific definition of theory of mind I will rely on in this thesis comes from a computational formalisation of Dennett's (1987) notion of the "intentional stance" (Baker, Saxe, & Tenenbaum, 2009; Baker et al., 2017). Dennett's intentional stance framework suggests that when humans make inferences about another agent's mental states, they assume that the agent is behaving in an approximately rational way to achieve their goals based on their beliefs. Baker and colleagues formalise this framework in an inverse planning model, where the computational mechanisms of reinforcement learning are used to make probabilistic inferences about an agent's unobservable goals and beliefs based on their observable actions. This formalisation allows Baker and colleagues to take advantage of the computational models developed for the vast field of reinforcement learning and Bayesian modelling in artificial intelligence and cognitive science (Sutton & Barto, 1998; Lake et al., 2017). Across two separate sets of studies, Baker and colleagues show that their model predicts participants' experimentally observed mentalising judgments. The mathematical specifics of this computational formalism are not germane to the work presented within this thesis. However, their delineation of the boundary conditions for what constitutes theory of mind is important. When I present my theory of mind work (in Chapters 5 and 6), one possible criticism will be: is this *really* theory of mind? I will use this widely-cited

and computationally-formalised definition to argue that it is. My goal when discussing theory of mind in the work that follows is not to make a claim about the set of all possible aspects of mind theoretic processes, but rather the specific subset that has to do with inferring the beliefs and desires of intentional agents based on their observable behaviour. I will use "theory of mind" to refer to this specific capacity.

Effort in theory of mind.

In the work presented in the second half of this thesis (Chapter 5 and 6), I will explore the circumstances under which theory of mind might be cognitively effortful. For as long as mental state attribution has been studied, the process underlying it has been thought of as automatic (Heider & Simmel, 1944; Coricelli, 2005). As shown in the famous Heider and Simmel demonstrations, people cannot help but attribute intentionality to even the simplest of agents. Accordingly, few experiments have investigated a possible effortful basis for processes of mental state inference, with most previous studies existing piecemeal rather than as part of a cumulative literature. One group of researchers has argued that mental state inference requires individuals to challenge their assumption in a "shared reality," and subsequently that having to acknowledge another individual's version of reality is cognitively effortful (Hodges, Denning, & Lieber, 2018). One survey provided evidence that people self-report consistent levels of "mind-reading motivation"—willingness or desire to consider others' mental states—across time (Carpenter, Green, and Vacharkulksemsuk, 2016). Another study demonstrated that investors showed increased motivation to take the perspective of their auditors in order to convince the auditor to endorse an outcome desired by the investor (Altiero, Kang, & Peecher, 2022). There is also evidence that games—and, by extension, behavioural tasks—can be designed to elicit more or

less perspective-taking (Dishon & Kafai, 2020). Other evidence suggests that while aspects of theory of mind, such as instruction-following, are automatic, it is the intentionality-tracking component which is effortful (Lin, Keysar, & Epley, 2010).

One influential account of theory of mind characterises it as a process of "egocentric anchoring and adjustment" (Epley et al., 2004). In this account, taking another individual's perspective begins with one's own perspective as a starting point and understanding the other person's in the extent to which it deviates from this baseline. This suggests how one kind of motivated account of theory of mind might work—assuming that the more significant the deviation between the target perspective and one's own, the more cognitively effortful it is to construct. One intriguing response to this initial idea has been the concept of "perspective-getting" (Eyal, Steffel, & Epley, 2018). While perspective-taking relies on the unreliable mental gymnastics of "putting oneself in another person's shoes," the idea behind perspective-getting is that the most effective way of accurately eliciting another person's perspective is to ask them for it. This conception of perspective-getting constitutes a process even more reliant on motivation. However much effort is required to deliberately engage in theory of mind, it is almost certainly more difficult to go out of one's way to request, consider, and analyse another individual's self-report data on the contents of their own mental states. However, none of these accounts has explicitly explore theory of mind or related phenomena in a way that both measures the relationship between perceived effort and one's ability to successfully infer another agent's mental states.

A notable exception to the lack of contact between motivation and theory mind research is the social motivation theory of Autism Spectrum Disorder (Chevallier et al., 2012). This account argues that Autism can be understood as a deficit not in capacity to engage in theory of

mind, but in the motivation to do so. Though—as with any theory of Autism Spectrum Disorder—this account has been heavily disputed (e.g., Jaswal & Akhtar, 2019). To the best of my understanding, this theory has not spawn a comparable debate in the theory of mind abilities of neurotypical populations.

In the rest of this section, I will outline three reasons why theory of mind might be hypothesised to be an effortful process in a way that has not yet been fully explored by existing research.

(1) Recent evidence suggests empathy is effortful, and theory of mind may prove to be effortful for a similar set of reasons. The term "theory of mind" is used to mean different things by different authors, and other terms may be used by yet other authors to refer to similar mental processes (Schaafsma et al., 2015). For example, Batson (2009) identified no less than eight entirely distinct (though related) mental phenomena referred to by the word "empathy" within the broader psychological literature. One of these eight processes is "cognitive empathy," which Batson defines as, essentially, theory of mind. For the purposes of this discussion, I will define empathy as a mental process related to the target agent's affective states, and theory of mind as one related to the target's cognitive states (Schurz et al., 2020).

While theory of mind is often treated as an automatic process (Coricelli, 2005), more research has investigated empathy as a motivated phenomenon—and therefore a process requiring mental effort (Keysers & Gazzola, 2014; Zaki, 2014). In Zaki's influential account, empathic concern is an automatic response, and because such emotional responses can be intrusive and therefore mentally costly, individuals are motivated to suppress this automatic response. Subsequent experimental evidence showed that participants' scores in a self-report measure of apathy and motivation predicted their scores in a measure of empathy (Lockwood et

al., 2017). In line with these studies, recent evidence suggests that an individual's propensity to engage in empathy can be increased with motivation-targeted interventions (Weisz et al., 2020).

A key experimental paradigm in this line of research has been the empathy selection task (Cameron et al., 2019; Ferguson & Inzlicht, 2022). In each trial of the task, participants choose instruction cards from one of two decks: one deck in which the participant is instructed to "share in the experiences" and indicate the internal experiences of a person in a photo, and one in which they are asked to "remain detached" and indicate the person's external features. In a series of variations on this paradigm, it was found that participants consistently chose the objective deck over the empathy deck (Cameron et al., 2019). In one variation, participants incurred a hypothetical cost of $2 to select the empathy deck (which they were not actually required to pay) while the objective deck featured varying costs across different trials. It was found that the average indifference point was $1.61, indicating a (hypothetical) willingness to pay an amount of $0.39 to avoid empathising. A further variation provided evidence that this effect was not due to aversion to experiencing emotional states in general or to verbalising emotions, and a recent study suggests that individual performance in this task is somewhat stable over time (with an average intraclass correlation coefficient of 0.66; Ferguson & Inzlicht, 2022). Yet, it is possible to raise several concerns about this study, the primary one being that it is difficult to say with certainty what exactly participants are doing when they select the empathy deck. For example, one reason why participants might not show the same avoidance effect for the version in which they are asked to imagine experiencing a given emotional state (rather than empathising) is that what participants find aversive is the process of attempting to understand another individual's mental states, rather than any process of affective simulation. This would be closer to an account in which theory of mind—the cognitive process underlying the understanding of another

person's mental states—is itself effortful. Ultimately, it has proved difficult to experimentally isolate the process of empathy, which is partially what has led to the terminological confusion on the topic. More nuanced paradigms will be required to distinguish precisely between the effortful components of theory of mind and empathy.

(2) Individual differences in theory of mind might be related to either motivation or capacity to engage in mental state inferences—and we do not have good paradigms for measuring either. One current issue in theory of mind research is that the vast majority of paradigms designed to study theory of mind can be successfully completed with near perfect accuracy by neurotypical adults (Conway & Bird, 2018; Apperly, 2012). Theory of mind is typically viewed as an all-or-none capacity, and it has proved difficult to develop measures that are able to distinguish between two individuals' respective abilities to successfully perform mental state inference (Conway et al., 2020). When existing paradigms are employed in an attempt to identify individual differences, any such differences can often be attributed to inferential abilities generally and not necessarily those having to do with social cognition (Conway et al., 2020). The extent to which certain kinds of cognition are intrinsically and exclusively social is still contested (Lockwood, Apps, & Chang, 2020). Thus, very few studies have investigated individual differences in capacity or motivation to perform mental state inference in any compelling way. Yet, as with any inference problem, mental state inferences can be more or less accurate—and therefore one would expect to find that individuals vary in their mind theoretic abilities.

One recent program of research which has sought to rectify this oversight is the Mind Space framework (Conway & Bird, 2018; Conway, Catmur, & Bird, 2019; Conway et al., 2020). The core thesis of the Mind Space framework is that mental states arise within specific minds;

therefore, a better understanding of the target mind could conceivably lead to more accurate inferences about what kind of mental states are likely to arise from it. Individual minds, then, can be represented as points in an arbitrarily high-dimensional space of mind-related traits. This framework applies a similar approach as the face-space literature, which seeks to understand the cognitive and neural processes underlying facial recognition by treating each individual face as an individual instance of a common template which can vary in an arbitrarily large number of dimensions (Valentine, 1991). Any individual differences in theory of mind ability may stem from more effective mapping of mind dimensions to resulting mental states. Conway and colleagues (2020) provide experimental evidence that a measure of perceived covariation of mind-related traits can predict a participant's more or less accurate inferences about the mental states of a target agent. Theory of mind accuracy was measured using the Movie for the Assessment of Social Cognition (Dziobek et al., 2006) along with a task designed to teach participants about the personality dimensions of the individuals in the video. As Conway and colleagues report, participants were more likely to make the correct mental state inference when they scored higher in the personality learning task. One limitation of this study—which I will outline in detail in Chapter 5 as a general problem in theory of mind tasks—is that there is not necessarily a causal link between mental states and any observed behaviour. The video stimuli feature actors, who may or may not be experiencing the mental states they are supposed to demonstrate.

The Mind Space framework suggests a potential theoretical account of why mental state inference might be perceived as difficult. A key claim is that accuracy of mental state inferences is dependent on accurately situating a given mind within a space of relevant mind-related traits. This implies that a target mind might be closer or further away to one's own within this space. If

theory of mind is, at least partly, a process of "egocentric anchoring and adjustment" (Epley et al., 2004; Thornton et al., 2019), then the inference problem might prove more difficult the further away in conceptual space a target mind is from one's own. Additional work suggests that people understand mental states—mental states, not minds—in terms of rationality, social impact, and valence (Thornton & Tamir, 2020; Thornton et al., 2022). The possibility that the difficulty of mental state inference is a function of distance between minds is also suggested by experimental evidence—for example, that social working memory training is linked with improved perspective-taking accuracy (Meyer & Lieberman, 2016), and that higher-order intentionality tasks have been shown to be more cognitively demanding (Lewis et al., 2017). In order to explore this account fully, experimental paradigms for teasing apart individual differences in theory of mind will need to be developed.

(3) People are less motivated to act prosocially toward out-group members, and might therefore perceive an attempt to understand out-group minds as requiring more effort. An important consequence of the idea of empathy as a motivated phenomenon is that people are motivated to engage in empathic concern more for members of our social in-groups than those of an out-group (Galinsky et al., 2006). Indeed, it is well-accepted that intergroup dynamics can lead to attenuated empathic concern, observable even in hemodynamic response (Cikara, Bruneau, & Saxe, 2011; Gutsell & Inzlicht, 2012). A related set of research explores intergroup discrepancies in meta-perceptions—whether people look down on out-groups because they believe those out-groups look down on them (Lees & Cikara, 2019; Lees, 2022). Though such meta-perceptions are comprised of beliefs about others' beliefs as a function of group membership, they are not necessarily driven by any process resembling theory of mind. Similar studies have showed how theory of mind employed in a competitive context to more effectively

counter one's competitors (Galinsky et al., 2008), how group membership can affect performance in a visual perspective-taking task (Simpson & Todd, 2017), and how the threshold for mind perception can change for in- versus out-group members (Hackel, Looser, & Van Bavel, 2014). Another intriguing finding is the potential connection between children's prosocial tendencies and their faculties of theory of mind (Stengelin, Hepach, & Haun, 2019; Hepach, Hedley, & Nuske, 2020). But to my knowledge, existing research cannot adequately characterise the extent to which we might be better or worse at, or more or less motivated to engage in, theory of mind for out-group members.

In summary, I hypothesise that there are two plausible accounts of why inferences about the mental states of out-group targets might be perceived as more difficult than those about in-group targets:

(1) *People are averse to engaging in any active consideration of an out-group target's mental lives—whether in affective states via empathy, or cognitive states via theory of mind*. If empathy itself is effortful (Zaki, 2014), and people are particularly averse to engaging in empathy directed at out-groups (Cikara, Bruneau, & Saxe, 2011)—then if, as I have argued above, theory of mind is effortful for similar reasons as empathy, we might expect to find a similar aversion toward mental state inference directed at out-group targets. This is a hypothesis about the *motivation* to engage in mental state inference. It is not that people are fundamentally worse at making inferences about out-group minds, but that they are less inclined to do so.

(2) *Out-group minds are more difficult to make inferences about, and therefore represent a substantively more difficult theory of mind problem*. This is a hypothesis about the *capacity* to engage in mental state inference, rather than the motivation to do so. If people understand minds and mental states as points in space (Conway & Bird, 2018; Thornton & Tamir, 2022), and

people also rely on an understanding of their own minds and mental states to interpret those of others (Epley et al., 2004; Thornton et al., 2019)—then making an inference about a mind sufficiently different from one's own may represent a more difficult inference problem. While there are many different ways of dividing up social groups, members of an out-group who share vastly different beliefs or customs than one's own might be harder to make inferences about simply because it requires taking into account these differences in worldview.

I contend that the reason that no existing experimental evidence directly tests or supports these claims is that our present paradigms for studying theory of mind cannot pick up on reliable signals of individual performance in mental state inference (I will expand on this claim in Chapter 5). In short, current theory of mind paradigms were initially designed to test the ability to impute mental states in general (e.g., Premack & Woodruff, 1978) rather than to distinguish between individual abilities in making more challenging inferences about another agent's specific beliefs about a situation conditional on differences in their more general beliefs about the world—which is where I hypothesise greater difficulties in mental inferences about out-group members will be found. This "intergroup theory of mind" framework is the general research program in which I intend the work in this thesis to represent the first step.

Thesis Outline

The conceptual throughline of this thesis lies at the intersection of the cognitive psychology of motivation and the social psychology of group membership. In the studies reported in each chapter, I will present participants with an opportunity to engage in an effortful action, vary the amount of effort required or the magnitude of reward promised, and observe how

participants' willingness to engage in those effortful actions changes based on the person on whose behalf they are acting.

In Chapter 2, I will extend a core paradigm in the study of prosocial motivation (the Prosocial Effort Task; Lockwood et al., 2017) to an intergroup setting. Previous versions of the Prosocial Effort Task have tended to look at an anonymous other—where no socially identifying information is provided about the other agent for whom the participant is working. I will use my Intergroup Prosocial Effort Task to investigate how the decisions dynamics of this task change when the other target is presented as an in- or out-group member.

In Chapter 3, I will explore the extent to which the intergroup effects identified in Chapter 2 can be explained by a participant's perceived similarity with the other agent. I will investigate this both in correlational studies from multiple datasets, as well as by experimentally manipulating perceived similarity directly.

In Chapter 4, I will invert the paradigm from Chapter 2. Instead of using the Intergroup Prosocial Effort Task to look at prosocial behaviour, I will investigate antisocial behaviour. In other words, how much effort are participants willing to put in not for the benefit of another agent, but for their detriment?

In Chapter 5, I will shift from studies on the role of effort in prosocial decision-making to studies on the role of effort in theory of mind. This remains prosocially-focused in that it often is necessary to understand another agent's beliefs and desires before deciding how best to exert effort on their behalf. In this chapter, I will present a novel stimuli set for evaluating effort and accuracy in theory of mind. Though theory of mind is a well-studied topic, most previous paradigms fall short of providing strong measures of variation in accuracy and perceived effort in neurotypical adults' engagement in mental state inference.

In Chapter 6, I will seek to use my novel theory of mind stimuli to demonstrate differences in motivation to perform mental state inference as a function of group membership—will participants put in less effort to understand an out-group mind simply because the agent is a member of their out-group? I will present several different studies investigating this question.

In Chapter 7, the General Discussion, I will describe how the present empirical work extends our knowledge of the existing research presented in this Introduction. I will discuss the limitations of the intergroup theory of mind studies presented in Chapter 6. Then I will sketch out what a program of research might look like to more adequately address questions of intergroup theory of mind.

Note on author contributions. All of the following material—experimental design, data collection, analysis, drafting of written material, etc.—is based solely on the contributions of Cody Kommers, the author of this thesis. Therefore, the pronoun "I" (instead of the standard scientific practice of "we") is used throughout. None of the following work was featured in published articles at the time of submission of this thesis. There is one exception: the data from Study 3.1 come from Lockwood and colleagues (2021), with the relevant author contributions detailed in their original paper. In Study 3.1, it is indicated where conclusions are based on the original author's published analyses; all other analyses were novel and conducted by the author of this thesis. In Chapter 3, Figures 1 and 2 were created by Lockwood and colleagues (2021). All other figures were created by the author.

CHAPTER 2:

PROSOCIAL MOTIVATION AS A FUNCTION OF GROUP MEMBERSHIP

Abstract. *This chapter investigates prosocial motivation—a person's willingness to put in effort on another person's behalf—as a function of group membership. How does whether a target agent is an in-group or out-group member affect an individual's motivation to act prosocially toward them? In particular, I look at whether intergroup effects on prosocial motivation are driven by an "in-group bump" (an increased valuation of rewards for in-group members similarly to valuation of rewards for oneself) or "out-group dip" (a decreased valuation of rewards for out-groups, below the valuation of rewards for an in-group). In two studies, I use a behavioural task and suite of computational models based on Lockwood et al.'s (2017) "Prosocial Effort Task" to investigate the conditions under which participants chose to behave prosocially toward in-group and out-group members. In the first study, I demonstrate a basic effect consistent with both an in-group bump and out-group dip account—indicating that participants show significant discounting of rewards for out-group members, but value in-group rewards similarly to rewards to oneself. In the second study, I replicate this finding and extend it to show that these decisions were based on the amount of reward at stake rather than the effort required to earn it. Although this finding is aligned with the general trend of decades of experiments on group behaviour, I discuss how an investigation of these effects in the context of willingness to exert effort adds to our understanding of the topic. Finally, I present correlational data suggesting these group effects might be driven by variation in perceived similarity.*

Introduction

Given the choice between two actions leading to equivalent outcomes, humans tend to choose the easier one (Hull, 1943; Zipf, 1949). However, in real life situations it is rarely straightforward to compare the value of potential outcomes and the costs required to attain them. In particular, it is difficult to evaluate an outcome when multiple agents are involved. How do we weigh the value of outcomes experienced not just from our own point of view, but from the point of view of other agents? When a person chooses to go to out of their way to help an individual other than themselves, we describe this as an act of prosocial behaviour. Such actions contradict a simple self-interested utility-maximising account of behaviour and require further explanation.

Recent inquiry has begun to interrogate the nuances of prosocial motivation in adult humans (Lockwood et al., 2017; Inzlicht & Hutcherson, 2017; Le Heron et al., 2018; Inzlicht, Shenhav, & Olivola, 2018; Müller & Apps, 2019; Lockwood et al., 2021; Umemoto, Lin, & Inzlicht, 2021; for a related literature in infants and chimpanzees, see Hepach, Vaish, & Tomasello, 2013; Hepach et al., 2021; Hepach et al., 2016). Specifically, this work has sought to characterise how people judge the costs and benefits of their actions when it is the acting individual who incurs the cost, and another agent who receives the benefits. Prior to this recent line of work, most research into prosociality investigated costs such as financial donations to other individuals (e.g., Engel, 2011) or moral decisions to profit from others' harm (e.g., Crockett et al., 2014). But in daily life such financial and moral considerations often do not come into play. For example, the decisions to help a friend prepare for an upcoming deadline, to hold the door for a stranger, or to shovel snow off an elderly neighbour's sidewalk do not come at a moral or financial cost. Instead, prosocial acts often depend on our motivation to exert either physical or mental effort (Contreras-Huerta, Pisauro, & Apps, 2020).

The upshot of this nascent literature is that people are less willing to put in effort to earn rewards for others than for themselves (Lockwood et al., 2017; Lockwood et al., 2021). This can be viewed as a type of discounting: one can identify the extent to which an individual devalues a reward depending on the person to whom that reward is given (Jones & Rachlin, 2006). For example, a five pound sterling note may have an abstract value as a monetary instrument, but the value a person derives from the existence of such a note declines severely when it moves from their own wallet to someone else's. Lockwood and colleagues (2017) have described this kind of discounting as a form of apathy, which they define as one's lack of willingness to enact goal-directed behaviours. The inverse of apathy within this framework is motivation: the extent to which an individual is willing to perform such a goal-directed action. Crucially, motivation can be dissociated from the execution of a goal-directed action. For example, a person could choose to attempt a particular action without actually successfully executing it. This is the difference between motivation and capacity to execute an action (Contreras-Huerta, Pisauro, & Apps, 2020). In the studies that follow, I will explore the specific concept of prosocial motivation: when faced with the opportunity to either perform or forgo an action, what are the conditions under which people are willing to put in effort to help others?

Most of the studies investigating this question have sought to outline the contours of prosocial motivation toward an anonymous other agent—one about whom relatively little is known by the agent making the decision. This has been done in previous experiments (e.g., Lockwood et al., 2017) in order to assess prosocial motivation without any confounding effects of reputation or reciprocity. However, several recent experiments have begun to look at how these effects are mediated by information about the other agent. For example, one line of research has investigated how people make social decisions concerning close others (e.g., family

and friends) and showed in both risk-taking and probabilistic learning tasks that participants tended to make decisions to benefit their parents over their close friends (Guassi Moreira et al., 2020; Guassi Moreira et al., 2021). Another experiment demonstrated that in a reinforcement learning task participants showed different neural responses when they were observed by an out-group versus in-group member while receiving rewards (Hobson & Inzlicht, 2016). When they were observed by an out-group member, participants showed a stronger feedback-related negativity signal versus when observed by an in-group member. This neural marker tracked with self-reported intergroup bias. Other recent tasks looked at choices to engage in effort to benefit oneself, a stranger, or a charity (Depow, Lin, & Inzlicht, 2022). In this study, participants were more willing to engage in effort to benefit a charity than a stranger, but less willing to help a charity than themselves. While these studies allude to the impact of group-based considerations on prosocial motivation, to my knowledge no studies have directly investigated how prosocial motivation varies as a function of social group.

While there is a long tradition of studying intergroup biases in psychological research (e.g., Tajfel et al., 1971; Tajfel et al., 1978; Tajfel, Turner, & Austin, 1979), such considerations have not yet been brought to bear in this emerging literature on prosocial motivation. In this chapter, I will use a combination of computational and experimental methods to study how people discount rewards as a function of group membership: how does the value of a reward change depending on whether it is going to oneself, an in-group member, or an out-group member?

Intergroup bias in prosocial motivation.

Batson (2022) defines prosocial motivation according to four constituent components, each of which represents a different set of people who are the beneficiary of one's actions. These components are: egoism (increasing one's own welfare), altruism (increasing the welfare of individuals who are not oneself), collectivism (increasing the welfare of a group), and "principlism" (defining welfare in terms of some a priori moral principle, standard, or ideal). Dovidio (2022) notes in a commentary on Batson that this definition prioritises group dynamics and social identity to a greater degree than most existing research on prosocial motivation. Indeed, most studies in prosocial motivation are concerned with motivation to help an anonymous other person—an agent about whom no identifying social information is provided (e.g., Lockwood et al., 2017).

While the nascent prosocial motivation literature has not yet broached the topic of social groups, one more developed literature which has looked at prosociality in an intergroup context is based on parochial altruism (an increased likelihood to behave altruistically toward members of one's own group; cf. Everett et al., 2015; Gao et al., 2015; Turner, Brown, & Tajfel, 1979; Brewer, 1979; Hewstone, Rubin, & Willis, 2002). While it has been shown that people tend to behave more prosocially toward in-group members—an effect usually known as "in-group favouritism" (Everett et al., 2015)—it has often proved difficult to disentangle the mechanism behind this effect. I will explore two competing explanations, framed in the decision-scientific terms of "social discounting" (Jones & Rachlin, 2006; Margittai et al., 2015; Vekaria et al., 2017; Tamir & Hughes, 2018): how much is a reward devalued when it goes to someone other than oneself? These two competing mechanisms are: an "in-group bump" (in which valuation of rewards for in-group members is increased, such that rewards for in-groups are not discounted

more heavily than those for self) versus an "out-group dip" (in which valuation of rewards for out-groups is decreased, such that rewards for out-groups are more heavily discounted than those for in-group). Either pattern of reward valuation can lead to an effect that resembles standard in-group favouritism (Everett et al., 2015). What I define as the "in-group bump" is consistent with other effects observed in the social psychological literature, such as depersonalisation (Tajfel and Turner, 1979; Turner et al., 1987) and in-group love (Everett et al., 2015). The "out-group dip" is consistent with effects such as out-group derogation or out-group hate (Everett et al., 2015). I use these novel terms in order to reflect the fact that I am specifically looking at reward valuation (as well as willingness to exert effort) as a function of group membership. Furthermore, these forces of differential reward valuation are dissociable (Moscatelli & Rubini, 2017). Moscatelli and Rubini (2017) argue that, from the perspective of social and evolutionary psychology, the in-group bump (which they call in-group love), which is based on the need for in-group protection and affiliation, should be more prevalent than an out-group dip (which they call out-group hate), which is based on hostility and competitive attitudes. This argument is aligned with the conclusion of a meta-analysis providing evidence that intergroup cooperation tends to derive from preference for positive in-group outcomes rather than negative out-group outcomes (Balliett, Wu, & De Dreu, 2014). On the other hand, a recent naturalistic study by Abbink and Harris (2019) shows that both increased valuation of positive outcomes for in-group members and decreased valuation for out-groups occur in supporters of two rival political factions in Thailand. Yet the authors find when they employ an isomorphic paradigm in artificial groups, the only effects are due solely to increased valuation of positive in-group outcomes. In short, experimental and theoretical evidence suggests that in-group favouritism is more often motivated

by a desire to achieve positive outcomes for in-group members, than to achieve negative ones for out-group members—but, under the right conditions, both can occur.

When they do occur, both an in-group bump and out-group dip can be caused by a constellation of factors, for example: out-group threat (Lang et al., 2021), moral suasion (Bilancini et al., 2020), and perception of out-group homogeneity (Simon et al., 1990). But the dynamics that govern when one of these forces wins out over the other depend not just on psychological factors, but cultural ones. A recent naturalistic study showed that participants in a Tanzanian village with social norms emphasising out-group cooperation showed the inverse of the usual intergroup effect, with higher rates of prosociality toward out-group members than in-group members (Kok et al, 2020). Similarly, while victims of genocide in Rwanda show strong in-group prosociality bias, Rwandan perpetrators of genocide actually show greater prosociality toward out-group members (Pech, Gishoma, & Caspar, 2022). And while Fiedler and colleagues (2018) found consistent in-group favouritism in a multinational study, participants from different nationalities showed significant differences in the extent to which they endorsed favouritism. Ultimately, the extent to which an individual engages in in-group favouritism affects their valuation of rewards that go to members of those respective groups—which can be seen not only in behavioural measures, but in the intensity of response in the ventral striatum, an area typically associated with reward signalling (Hackel, Zaki, & Van Bavel, 2017).

While useful in elucidating the circumstances in which these forces of intergroup bias emerge, none of the above studies distinguishes between the two dissociable facets studied in the literature on prosocial effort: capacity and motivation. Similarly, most of the above studies rely on paradigms based in distributing monetary rewards or choosing whether to cooperate or defect rather than willingness to exert physical effort. Finally, it is also known from existing prosocial

motivation studies that people exhibit a self-bias: we value our own rewards more than strangers, and we put more effort in to benefit ourselves than strangers (Lockwood et al., 2017). This is important because most studies on in-group favouritism do not distinguish between in-group benefit and self-benefit. Thus, we do not know whether people will be more motivated for themselves than for an anonymous in-group member.

The Prosocial Effort Task.

In the studies that follow, I employ a version of the Prosocial Effort Task used by Lockwood and colleagues (2017; 2021). In this paradigm, participants choose whether to exert effort in order to earn rewards. On some trials, the rewards go to the participant themselves, while on other trials the rewards go to another individual. The paradigm is designed to separate the participants' decisions to exert effort (motivation) from their successful execution of their effort (capacity). This allows us to investigate whether any intergroup bias exhibited by the participants is due to their choice to help others versus their ability to follow through on that intention. As mentioned above, a key limitation of most studies on prosocial behaviour is that they give participants the opportunity to unevenly distribute monetary rewards (e.g., Duclos & Barasch, 2014). Yet in the real world, prosocial behaviour often does not require an investment of money—but of effort. In the studies that follow, I will be able to disentangle the effects of group membership on prosocial motivation as it varies throughout different levels of effort and reward.

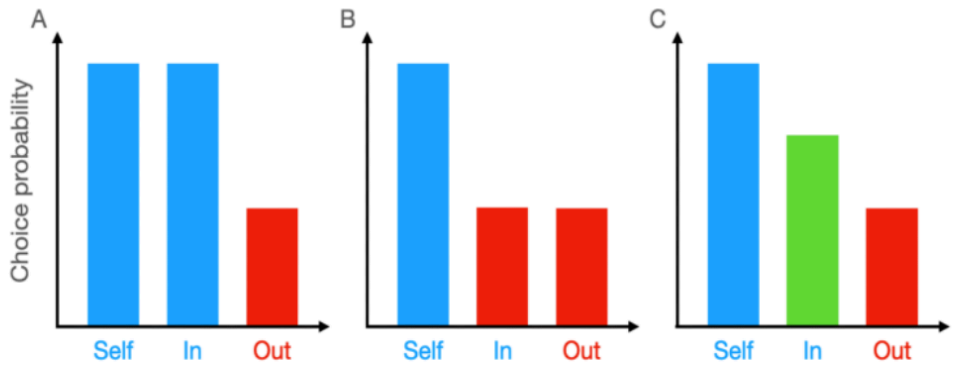


Figure 1 — Intergroup hypotheses. These panels depict three broad patterns of plausible social discounting which could be born out in these data. (**A**) Only decisions for out-groups are significantly discounted, with participants employing roughly the same pattern of choice for in-group members as themselves, while choosing to earn less for out-group members. (**B**) Decisions for in- and out-groups are discounted to the same extent, in line with the standard divisions of "self" and "other." (**C**) Participants could also use a distinct discounting strategy for each agent, with rewards for out-groups being substantively more discounted than for in-group agent.

Hypotheses and summary of studies.

In the two experiments described below, I will measure participants' willingness to exert effort, as defined by how often participants choose a high-effort option which earns more reward over a low-effort option which earns less reward. In each trial these rewards will go to the participant themselves, a member of their in-group, or a member of their out-group—with the groupings defined by a minimal groups paradigm. In the first experiment, I will demonstrate a preliminary intergroup effect. In the second experiment, I will replicate and extend this initial finding. Figure 1 shows a schematic of three plausible hypotheses for general patterns that can be found in these data. (A) Participants could show significant discounting of rewards going to out-group members (as in Lockwood et al., 2017), while making choices for in-group members with a discount function that resembles choices for themselves; (B) Participants could make similar choices for both in-group and out-group members, according to the same discount function; or

(C) Participant choices could reflect discounting of rewards for in-group, while discounting out-group rewards even more heavily. These hypotheses will be tested by both behavioural measures (whether there are significant differences in choice probabilities between the different conditions) and computational models (based on the parameters employed by the winning model).

Study 2.1: Intergroup Prosocial Effort Task

Participants. Thirty participants were recruited on the Prolific subject recruitment platform (8 female; mean age = 26.1). All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately fifty minutes to complete.

Method.

Design. This implementation of the Prosocial Effort Task was similar the one described by Lockwood and colleagues (2017; 2021) with two main differences: it was adapted to be performed online (and thus required a different effort operationalisation), and included an additional condition for the intergroup manipulation (Figure 2). In previous versions of the task, effort was operationalised as grip force, measured by a hand-held dynamometer, with maximum voluntary contraction measured as a baseline to allow for individual differences in grip strength. In the present version, I used button presses (as completed on a computer mouse or trackpad) as an effort manipulation. This approach has been used in previous effort-based decision tasks (Treadway et al., 2009; Treadway et al., 2012; Suzuki et al., 2021) and has been shown to elicit

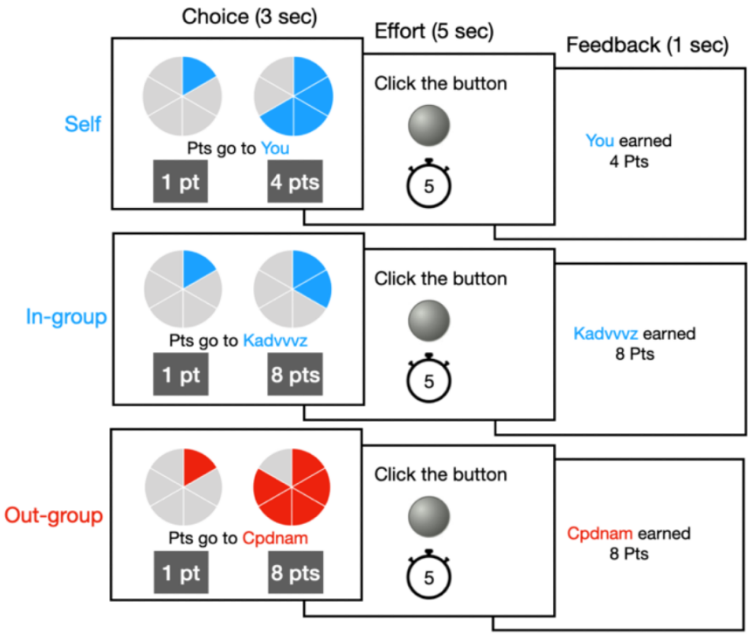


Figure 2 — The Intergroup Prosocial Effort Task. Before the main task, participants underwent a minimal groups induction procedure. They were separated into two teams based on colour preference. In each trial of the main task, participants were presented with the option to earn points for one of three agents: self, in-group, or out-group. Participants could choose to earn one point in exchange for no effort, or a greater number of points for more effort. Number of points was indicated on the choice button (first screen, bottom right). Effort levels were represented by a pie chart (first screen, top right)—more slices indicated a greater number of required button clicks during the five second effort screen in order to earn the available points.

standard effort discounting effects. Furthermore, it requires no special equipment, such a dynamometer, making button-pressure a useful measure of effort for online research.

Participants completed 225 total trials across three conditions: 75 trials earning points for themselves, 75 for an in-group member, and 75 for an out-group member. These trials were distributed across three blocks. In each block, participants completed each possible reward-effort combination (five reward levels: 2, 4, 6, 8, 10; five effort levels: 2, 3, 4, 5, 6) in each of the three conditions (self, in-group, out-group). In each trial, the participant was given a choice between a baseline option of 1 point requiring no effort versus an option to earn a greater number of points requiring a specified level of effort. Participants were informed whether the points would go to

themselves, the individual on their "team," or the individual on the other "team." Each participant had three seconds to make a choice; if they failed to do so, they earned zero points.

If a participant chose to exert effort for more points, they were shown a subsequent screen with an image of a button labelled with the injunction to "press the button." Exerting the required amount of effort consisted in pressing the button (i.e., clicking on the button with their mouse) a specified number of times. (Participants were informed in the instructions what to expect when presented with the button screen.) Participants were given exactly five seconds to press the button the required number of times. They were given feedback about whether they succeeded, which lasted an additional 1 second. If a participant chose the baseline option, they were shown a screen stating that they had earned a single point, lasting for 6 seconds—equivalent to amount of time they would have spent had they chosen the effort condition, controlling for any effects of temporal discounting. If a participant failed to make a choice, the subsequent feedback screen also lasted for 6 seconds. No matter which actions were taken by the participant, each trial lasted 9 seconds in total.

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Prolific subject recruitment platform.

Calibration procedure. Before the main task, participants underwent a calibration phase to gauge an appropriately difficult rate of clicking for the main effort task. The calibration phase employed a staircase design. On the first level, participants were instructed to click the button 40 times within an eight second period. Participants who succeeded on at least one of the three attempts were then moved to the second level: 50 clicks within an eight second period. If participants succeeded on at least two of these attempts, they were assigned to the calibration requiring the greatest number of clicks (i.e., effort level 2 required 5 clicks in five seconds,

increasing by 5 up to 25 clicks at level 5). Otherwise they were placed in the medium calibration (effort level 2 required 4 clicks in five seconds, increasing by 4 up to 20 clicks at level 5). If they failed the first level, they were given three chances to click the button 30 times in eight seconds. Participants were only excluded if they failed to succeed in every attempt; otherwise they were assigned to the calibration requiring fewest clicks (effort level 2 required 3 clicks in five seconds, increasing by 3 up to 16 clicks at level 5).

Group-induction procedure. Following the calibration procedure, participants underwent a minimal groups induction procedure. The procedure was based on recommendations from Pinter & Greenwald (2010), who compared different minimal groups induction procedures. To begin, participants were asked which colour they preferred: blue or red. Participants were told that they would be paired on a team with other Prolific workers who shared the same preference (Pinter & Greenwald also provided evidence that defining groups based on an arbitrary differences, such as colour preference, is more effective than random assignment, such as a coin flip).

Participants were then presented with a list of five names of their "teammates" drawn from a "previous version of this experiment" (the strongest effects from Pinter & Greenwald were produced by memorising the names of in-group members). The participants were given thirty seconds to memorise the names. The list contained the same names for each participant, though they were presented in the participant's team colour. These teammate names also consisted of usernames instead of given names—in an effort to avoid other group-identifying information, such as gender or ethnicity, and to give the impression of being drawn from an online group. The usernames were created by the experimenter to resemble actual usernames found on crowd-working online forums, such as Reddit. Participants were asked to pick a

username of their own (without identifying information) to give further credibility to the cover story. The participants then completed a short memory task to test whether they had successfully memorised the usernames (based on Dunham, 2013), requiring participants to categorise a set of ten usernames as "red team" or "blue team" with each name appearing, in random order, four times.

Questionnaire assessments and demographics. Each participant completed three questionnaires immediately after providing consent, performing the calibration, and undergoing the group induction: the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011), the Apathy Motivation Index (Ang et al., 2017), and the 20-item Toronto Alexithymia Scale (Bagby, Parker, & Taylor, 1994; Leising, Grande, & Faber, 2009).

Post-task self-report judgments. After completing the main experiment, each participant was ask to report how "similar" they perceive themselves to be to each of the two others players for whom they were ostensibly earning points ("How similar do you feel to [name]?" rated on a slider scale from "completely dissimilar" to "completely similar"). Participants were also asked about their beliefs of whether an agent would earn points for them ("Given the chance, how many points do you think [name] would have earned for you?" on a slider scale "as few as possible" to "as many as possible"). Participants were also asked to rate on the same scale: "In general, how many points do you think people earn for others in this task?" The order of these five questions was randomised.

Computational Models.

A series of computational models of effort discounting were fit to each participant's choice behaviour in the self, in-group, and out-group trials. These models were designed to tease

apart the different rates at which people discount rewards by effort, depending on the agent to whom the points are going. In other words, this parameter measures how quickly the effort becomes "not worth it" when earning points for oneself versus others. In line with previous modelling on this task (Lockwood et al., 2017; 2020) the model used various combinations of effort discounting functions, discounting parameters, and noise parameters.

Models were fitted to behavioural data using a softmax function. Each model contained two individually-fit parameters: one which reflects the degree to which a reward was devalued by effort (k), and a noise parameters which reflects the stochasticity of choices (β). The different combinations of models were based on varying two features. The first was the mathematical function characterised the form of discounting (i.e., whether in the course of exerting effort rewards are devalued linearly, hyperbolically, or parabolically; Hartmann et al., 2015). The second was whether participants' discounting of reward by effort was different for self, in-group, or out-group trials. This resulted in four classes of models: one discounting parameter for all trials; two separate discount parameters for self versus other (including in-group and out-group); two separate discounting parameters for in-group (including self and in-group) and out-group; or three separate discounting parameters for self, in-group, and out-group. Within these models, I also tested a further division of models, featuring separate noise parameters in the softmax function, according to the same structure: self; self vs other; in-group vs out-group; and self vs in-group vs out-group. With three possible discounting functions, four possible combinations of discounting parameters, and four possible combinations of noise parameters, this yielded 48 possible combinations of potential models.

In line with results from previous versions of the Prosocial Effort Task (Lockwood et al., 2017), I predicted that the winning model employed the following parabolic discounting formula:

$$V = R - (k * E^2)$$

where V is the subjective value of an option, R is the reward level, k is the discount parameter (fit to individual participants), and E is the effort level. The noise parameter, β, was built into the softmax decision function. In particular, the discounting parameters are a key means of distinguishing between the three hypotheses posed in Figure 1: an in-group would be consistent with a winning model featuring discounting parameters for in- versus out-group, while an out-group dip would be consistent with a model featuring three separate discounting parameters.

Results.

Throughout this thesis, I use a suite of statistical tests identical to those used by Lockwood and colleagues (2021)—namely, a combination of a linear mixed-effects regression model, generalised linear mixed-effects model, and a computational model fitted to individual participants' choices. I take this to represent the current standard practice in the field, as previous work on the Prosocial Effort Task (Lockwood et al., 2017) used a similar analysis suite, but then was updated in subsequent work (Lockwood et al., 2021) to the statistical techniques used in the analyses throughout this thesis. While there may not yet be consensus on the best practices for these relatively novel statistical techniques (Blanca, Alarcón, & Bono, 2018), they each represent techniques which rely on varying assumptions about the underlying data—suggesting that the analyses in this thesis are robust against the specific statistical assumptions of any single test.

To test hypotheses of whether participants' willingness to exert effort for self, in-group agent, or out-group agents were different, I ran a linear mixed-effects regression using the lmer function from the lme4 package for R (Bates et al., 2015). I defined choice probability as an outcome variable (the percentage of high-effort choices made by each participant) and agent as a predictor variable (self, in-group, or out-group), including a subject-level random intercept. While the regression did show a significant difference between self and in-group choice probabilities (b = 0.146 [0.017 – 0.275], z = 2.225, p = 0.026; Figure 3A), there was a much larger difference between in-group and out-group choice probabilities (b = -0.487 [-0.616 – -0.358], z = -7.413, p < 0.001; Figure 3A).

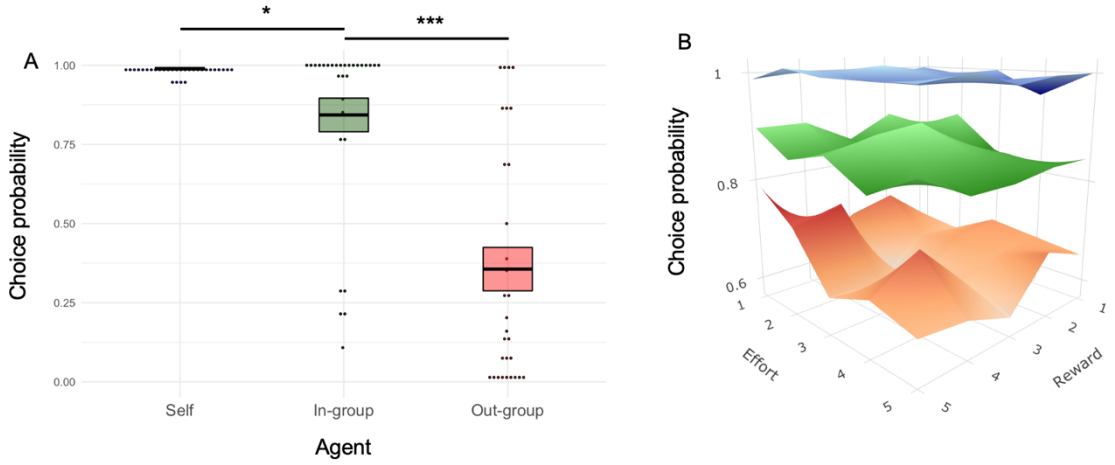


Figure 3 — Choice to exert effort varied by group. (**A**) Significant differences in average choice probability were observed between different group conditions. However, these data exhibit a ceiling effect in the 'self' condition. This suggests the effort levels were not sufficiently difficult to disincentive the participants from choosing to do the higher effort levels. (**B**) The three planes indicate self (blue), in-group (green), and out-group (red) choice probabilities as a function of reward and effort levels. Participants' choices tended not to be discounted by effort (i.e., the slope of the plane is relatively flat). This also suggests the effort levels were not properly calibrated, though these data still show significant effects of group. I recalibrated the effort levels in Study 2.2 to better understand the effect of effort on participants' choices.

Models were compared probabilistically according to their Bayesian Information Criterion, with models being penalised for additional parameters (Figure 4). The intergroup model with two discounting parameters (in-group and out-group; where self and in-group other are both treated as "in-group") consistently outperformed the discounting combinations, when

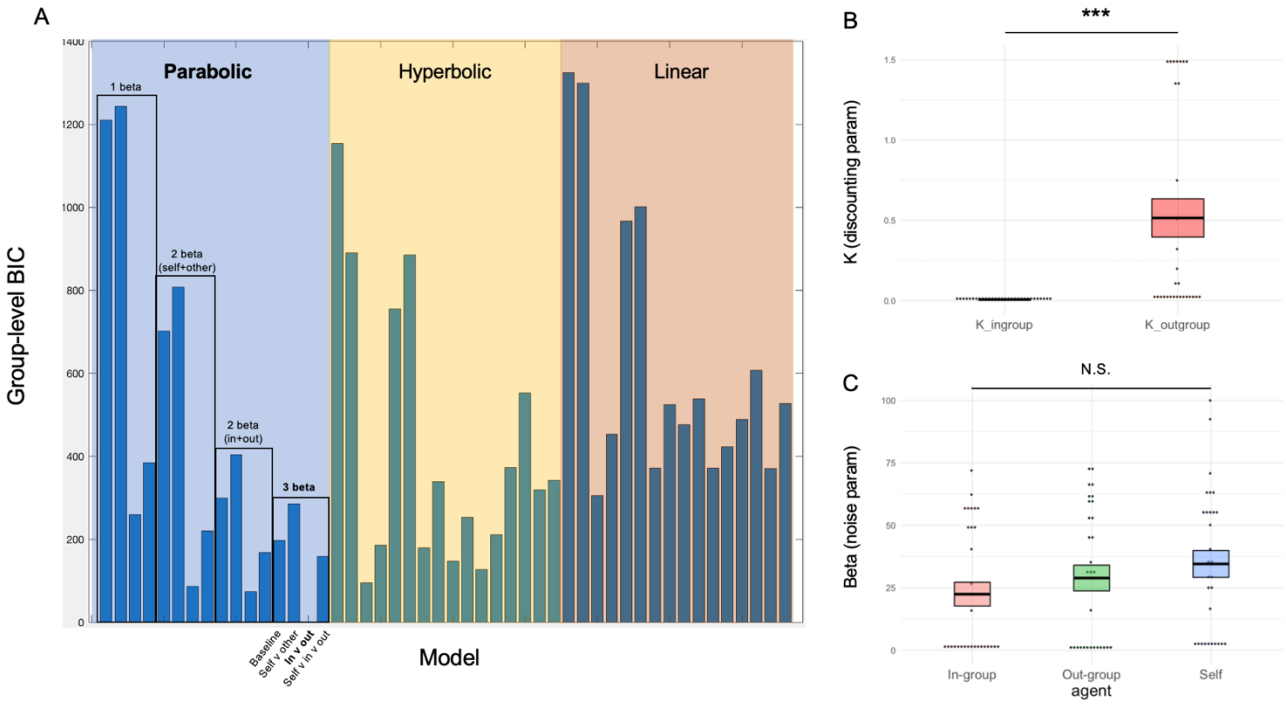


Figure 4 — **Model comparison.** These model results suggest that participants treated in-group members more like 'self' than 'other'—an initial result which was used to preregister hypotheses for the modelling in Study 2.2. The results replicated successfully. (**A**) Models varied across (1) the type of function used for effort discounting (parabolic, hyperbolic, or linear); (2) the number of noise parameters β (self; self, other; in-group, out-group; self, in-group, out-group); and (3) the number of discounting parameters k (self; self, other; in-group, out-group; self, in-group, out-group). Models were compared by group-level accumulated BIC (lower values are better). The winning model featured two parabolic discounting parameters (in-group, out-group) and three noise parameters (self, in-group, out-group. (**B**) Mean k values for in-group and out-group discounting parameters in winning model. Lower values indicate less discounting effect (i.e., greater willingness to exert effort). The difference in these mean values indicate that exerting effort for one's in-group (self or in-group other) remains relatively worthwhile across all effort and reward levels, while exerting for one's out-group shows greater discounting effects, as well as greater variability. (**C**) Mean β values for in-group, out-group, and self in winning model. A higher β value indicates the softmax choice function concentrates on fewer options. A lower β value indicates equiprobability.

holding other parameters constant. The winning model employed a parabolic discounting function, two discounting parameters (in-group and out-group), and three noise parameters (self, in-group, and out-group). This suggests that the difference between self and in-group other was not large enough to justify an extra discounting parameter in the model—consistent with the prediction of the in-group bump hypothesis (depicted in Figure 1A).

I also ran a generalised linear mixed-effects model to look for possible interactions (Figure 5). This was done using the glmer function of the lme4 package in R (Bates et al., 2015). There were significant main effects of agent ($X^2 = 1116.64$, $p < 0.001$) and effort ($X^2 = 17.62$, $p < 0.001$), but not reward ($X^2 = 3.16$, $p = 0.075$). I did not find interactions among agent, reward, and effort—which had been expected based on previous versions of this task. As with the ceiling effects of the "self" condition in Figure 3, this was likely because the highest effort level conditions were still trivially easy for most participants.

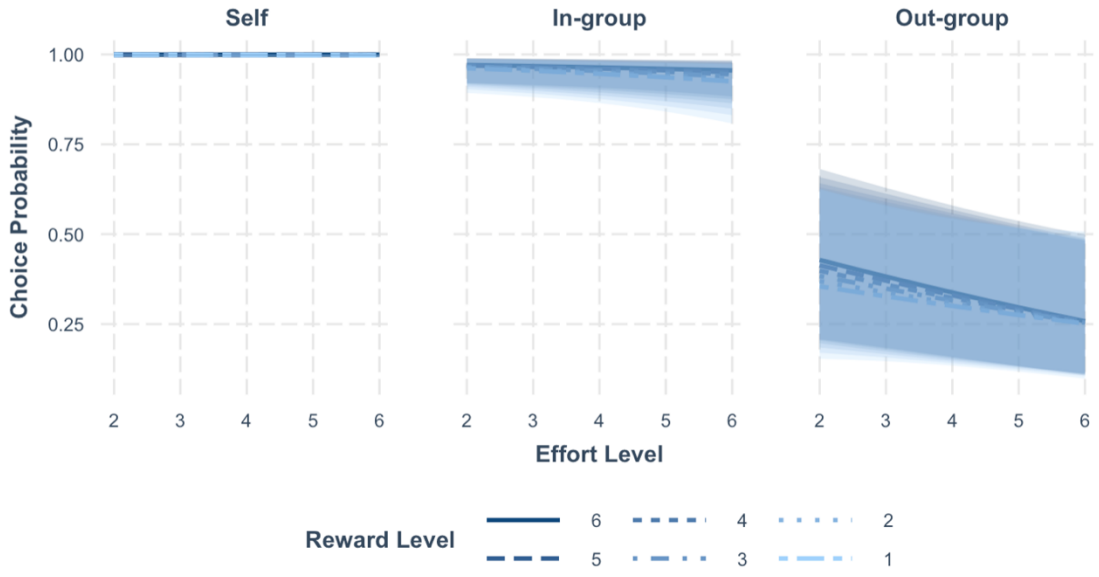


Figure 5 — Generalised linear model. I used a GLM predicting choice probability on individual trials with effort and reward level as fixed effects and participant as a random effect. This analysis supports the same conclusion as the average choice probabilities in Figure 3: significant group differences, with insufficiently difficult high effort levels.

Discussion.

The present findings show a clear difference between self and in-group motivation from that for out-group agents. Participants chose to exert effort on fewer trials for the out-groups than in both other conditions. However, whether self and in-group were different is less clear in these data. While the modelling results showed the same parameter for the discounting effect, different noise parameters were needed to characterise choices. In contrast, the statistical analysis revealed a small but significant difference between self and in-group. The mixed pattern of results could be explained by a ceiling effects in the self and in-group conditions, with choices to exert more effort close to 100% in many participants in both conditions. One reason for this may be that the task was simply perceived as not particularly effortful. In order to further test the same hypotheses in the next study I recalibrated the effort levels to make them more difficult and collected a new sample for a second study.

Study 2.2: Intergroup Prosocial Effort Task (Replication & Extension)

Power analysis. A full power analysis could not be conducted prior to this experiment, as not all of the effects of interest were present in the Study 2.1 data. Nonetheless, I conducted a power analysis for mixed-effects models based on Kumle, Võ, & Draschkow (2020) for the intergroup main effects, which were present in Study 2.1. This analysis showed that 80% power could be achieved to find an equivalently-sized main effect of group in only 20 participants (as the group effect is strong and each participant does many trials). Based on this analysis, the expected interaction effects, as well as power analyses of other versions of the Prosocial Effort Task (e.g., Lockwood et al., 2017; Lockwood et al., 2021), I settled on a sample of forty participants for the main version of the task.

Pre-registration. Methods, recruitment, analyses, and hypotheses were pre-registered for this experiment at https://osf.io/u2brs.

Participants. Forty participants were recruited on the Prolific subject recruitment platform (6 female; mean age = 26.7). Conditions for the participants' participation were identical to those in Study 2.1.

Method.

The methods for Study 2.2 were identical to Study 2.1, except that each effort level was made comparatively harder (i.e., requiring more button clicks) than in the initial version.

The structure of the calibration phase was similar to the one used in Study 1, though this one began with a practice round, to make sure participants understood the task, then followed a staircase design as in Study 2.1. On the first level, participants were instructed to click the button 45 times within an eight second period. Participants who succeeded on at least one of three attempts were then moved to the second level: 50 clicks within an eight second period. If participants succeeded on at least one of these attempts, they were assigned to the hardest calibration ($N = 9$; effort levels: 12, 17, 22, 27, 31 clicks within the five second window). Otherwise they were placed in the medium calibration ($N = 26$; effort levels: 8, 13, 18, 23, 28 clicks within the five second window). If they failed the first level, they were given three chances to click the button 40 times in eight seconds. Participants who completed at least one of these were assigned to the easiest calibration ($N = 5$; effort levels: 5, 10, 15, 20, 25 clicks within the five second window). Given the roughly normal distribution of calibration results (with most participants ending up in the medium condition), this calibration seemed to better reflect the true

distribution of participants' clicking abilities. Two participants were automatically disqualified because they failed to succeed in every clicking attempt.

Results.

Participants treated in-group members more like 'self' than 'other'.

Whereas the data from Study 2.1 showed a small main effect of mean choice probabilities between self and in-group, the same linear mixed-effects model showed no significant difference between self and in-group for the data in Study 2.2 (b = 0.052 [-0.046 – 0.151], z = 1.048, p = 0.294; Figure 6). There was still a significant difference between in-group and out-group (b = -0.364 [-0.463 – -0.266], z = -7.275, p < 0.001).

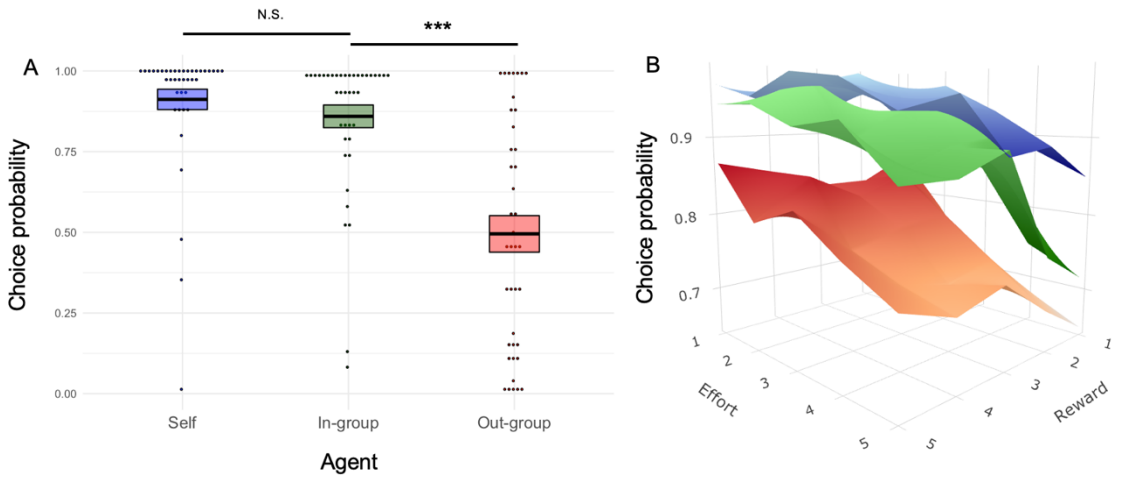


Figure 6 — Replication: Choice to exert effort varied by group, effort, and reward, with minimal difference between self and in-group choices. (**A**) Participants' choices for in-group members more closely resembled choices for "self" than for an anonymous "other." There was no significant difference between "self" and "in-group" conditions. Box plots reflect standard error. (**B**) Choice probability changed as a function of reward and effort levels. Colors reflect the same agent condition as in the 2D plot.

The computational model of effort discounting—defined identically to the model in Study 2.1—provided further evidence for this hypothesis. In line with my pre-registered hypotheses, the winning model featured a parabolic discounting function and three β noise parameters (Figure 7). I did not have a strong a priori hypothesis as to whether the winning model would feature two k discounting parameters (in-group and out-group) or three (self, in-group, and out-group), though I predicted one of these would win out over the non-group parameters (including a single discounting parameters across all conditions, or separate parameters for self and other). The winning model featured two discounting parameters: in-group and out-group. This was the same model that won in Study 2.1.

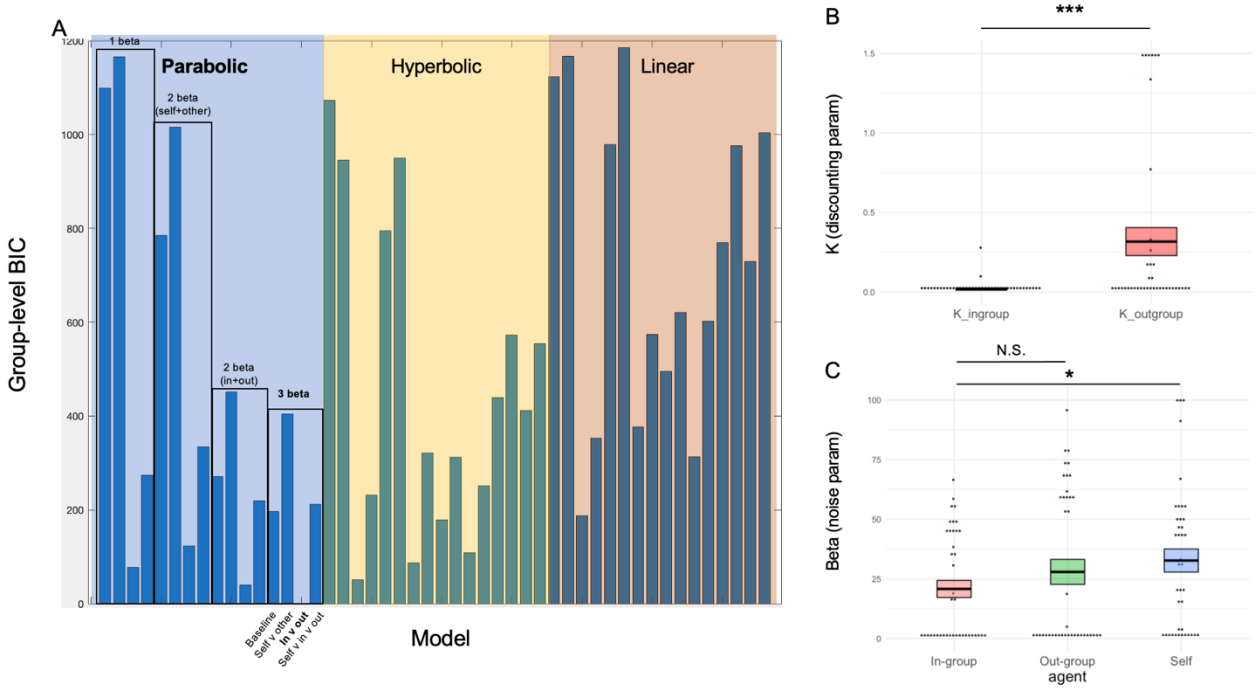


Figure 7 — Model comparison replication: participants treated in-group members more like 'self' than 'other'. The same structure of model was used on these data as in Figure 8. (**A**) The winning model featured two parabolic discounting parameters (in-group, out-group) and three noise parameters (self, in-group, out-group), in line with our pre-registered hypothesis. (**B**) Mean k values for in-group and out-group discounting parameters in winning model. (**C**) Mean β values for in-group, out-group, and self in winning model.

I also ran a linear mixed-effects model using individually-fit parameters from the computational model, which predicted the model's learned *k* values with the agent conditions defined as fixed effects and a subject-level random intercept. This model showed a significant difference between *k* values fit for in-group (when choosing for self or in-group other) and out-group ($b = 0.298$ [0.126 – 0.471], $z = 3.392$, $p = 0.001$; Figure 7B).

Participants earned for their in-group based on reward but not effort.

To support these model-based analyses, I analysed the choice data with a generalised linear mixed-effects model of the choice behaviour (Figure 8). With choice coded as a binary outcome variable, I defined agent, effort level, reward level, and their interactions as fixed effects, including a subject-level random intercept. Statistical significance of the fixed effects was assessed using a Type II Wald chi-square test.

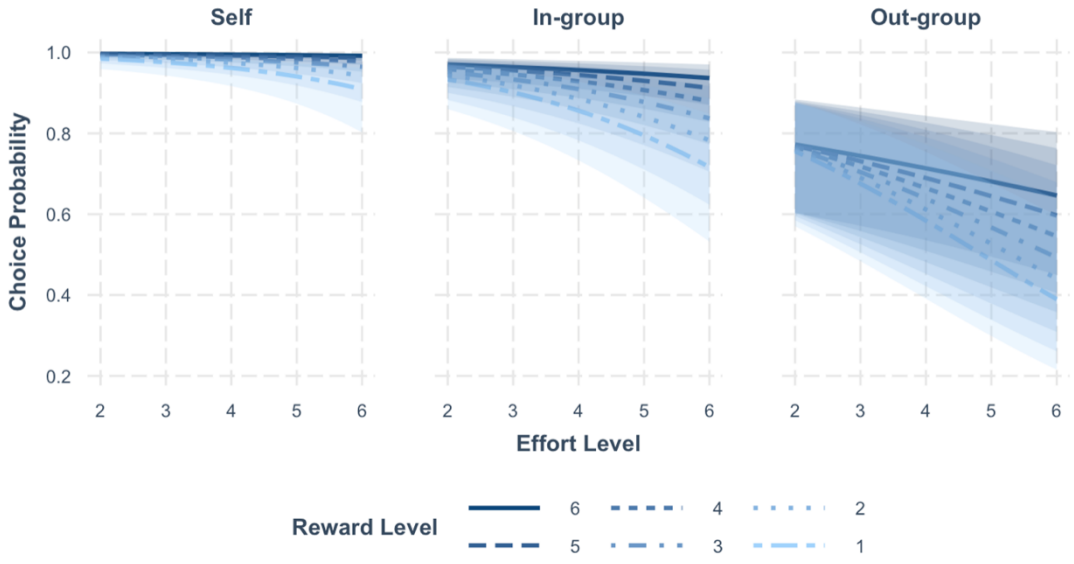


Figure 8 — Generalised linear model. This is the same structure for the GLM model used in Figure 4 on data from Study 2.1. These data reflect the recalibrated effort levels. Differences in discounting as a function of effort and reward can be seen between groups.

The model yielded significant main effects of agent (X^2= 1106.18, $p < 0.001$), effort (X^2= 117.13, $p < 0.001$), and reward (X^2= 66.86, $p < 0.001$). There was an interaction between agent and reward (X^2= 15.82, $p < 0.001$), but not between agent and effort (X^2= 1.06, $p = 0.30$). An interaction was also found between reward and effort (X^2= 6.59, $p = 0.01$).

Perceived similarity predicted task performance and model parameters.

Participants had been asked to self-report the extent to which they felt similar to each of these other agents, and so in exploratory analyses I sought to use perceived similarity to predict relevant task behaviour and model parameters (Figure 9).

I ran a linear mixed-effects regression which predicted choice probability (the percentage of trials on which participants chose the high effort option for a particular agent) based on fixed effects of agent (self, in-group, out-group) and the difference in perceived similarity between the other agents, as well as a subject-level random intercept (Figure 9). This analysis showed a significant similarity-by-agent interaction ($b = 0.007$ [0.004 – 0.010], $z = 4.299$, $p < 0.001$; Figure 9A). I also ran a regression (with the same fixed effects) to predict the model discounting parameters. This also showed a significant similarity-by-agent interaction ($b = 0.012$ [0.007 – 0.017], $z = 4.820$, $p < 0.001$; Figure 9B). In line with these regressions, I also found that the difference in intergroup similarity was a predictor of the difference in intergroup choice probabilities ($r = 0.53$, $p < 0.001$; Figure 9C) and the difference intergroup discounting parameters ($r = 0.62$, $p < 0.001$; Figure 9D). In other words, participants who perceived the out-group as more similar to themselves showed less of a difference in their willingness to exert effort for out-groups compared to the other conditions.

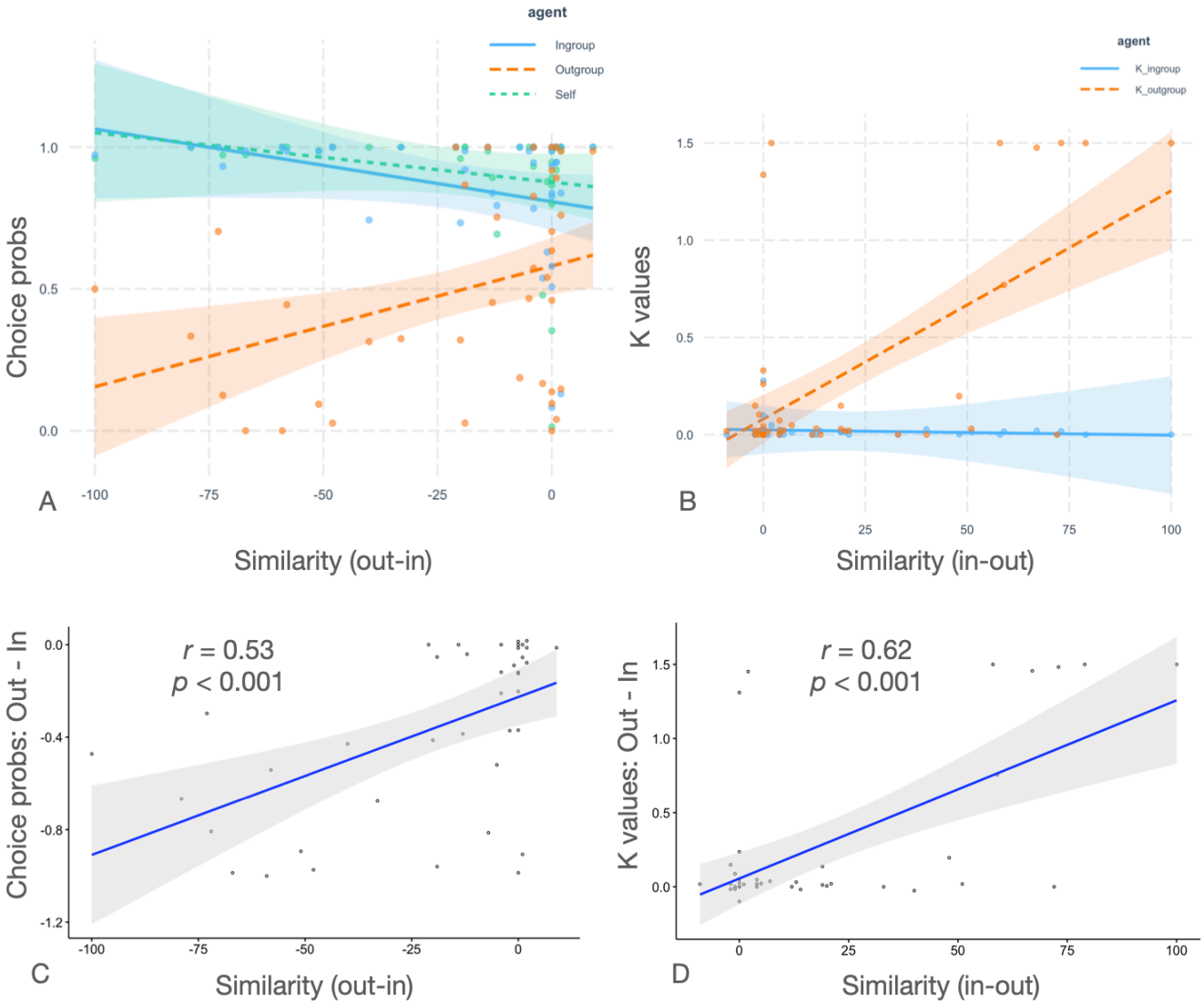


Figure 9 — Perceived similarity predicted task performance (left panels) and model parameters (right panels). (**A**) Difference in perceived similarity for out-group and in-group members predicted participants choice probabilities for out-group members in a linear mixed model regression, but not for in-group members or self. (**B**) Likewise, the difference in perceived similarity predicted out-group discounting parameter values, but not in-group values. (**C**) The difference in similarity judgments between out-group and in-group members predicted the difference in choice probabilities. (**D**) Likewise, the difference in similarity judgments between out-group and in-group members also predicted the difference in discounting parameters values.

Discussion.

These data are consistent with an intergroup bias in prosocial motivation. Participants' willingness to exert effort for in-group members strongly resembled their willingness to exert effort for themselves, while they showed significantly more discounting for out-group members. In Study 2.2 (which featured increased overall effort demands from Study 2.1), there was no significant difference in participants' probability of choosing the higher effort option (versus the baseline) in the self condition versus the in-group condition. Choice probability for both self and in-group conditions differed significantly from the out-group condition. This was aligned with the results of the model comparison, which showed that the winning model featured two separate discounting parameters for in-group (including both choices for both self and in-group) and out-group.

In pre-registration I predicted there would be main effects of agent, effort, and reward, all of which were present in the data. I also hypothesised interactions between each of these variables. Interactions were found between agent and reward, and between effort and reward, but not between agent and effort. As noted in the pre-registration, the agent-by-reward interaction suggests that participants are forgoing certain higher reward options simply out of the principle of not wanting to benefit another agent (in this case, out-group members), and not necessarily because of any increased effort cost. The lack of an agent-by-effort interaction suggests that while participants were sensitive to effort costs in general (they were significantly less likely to incur them for out-group agents, as shown by the main effect of effort), they were not sensitive to increases in effort in a way that differed between agents. In other words, the slopes indicating changes in choice probability across effort levels in the generalised linear model (Figure 8) are not significantly different between overall choice behaviour for in-group and out-group targets.

However, the slopes of the effort levels differ significantly for the varying reward levels, which represents the effort-by-reward interaction. For example, participants are much more sensitive to increases in effort in the lowest reward level, while they are less sensitive to changes in effort at the highest reward level.

Interactions among agent, effort, and reward have been found in previous versions of the Prosocial Effort Task (Lockwood et al., 2017; Lockwood et al., 2021), whereas no agent-by-effort interaction was found in this one. One possible interpretation is that in an intergroup scenario, participants are more sensitive to changes in reward levels across agents than changes in effort levels. In this account, participants are still sensitive to effort overall (being less inclined to exert effort, particularly for out-group members), but that their motivation is primarily determined by whether they want earn a given level of reward for in-group or out-group agents.

An alternative explanation is that participants *would* show different levels of sensitivity to changes in effort levels across agents (i.e., an effort-by-agent interaction), except that the effort costs in this paradigm are not sufficiently high to elicit them. The agent-by-effort interaction effects found by Lockwood and colleagues were found in an in-lab—rather than online—experiment, using a handheld dynamometer instead of button clicks under time pressure. If one imagines an even more extreme effort costs (say, one in which participants have to perform a weightlifting exercise to earn reward on each trial), it is possible to hypothesise that participants would become increasingly nuanced in their sensitivity to increases in effort. One reason to be concerned about this possible alternative explanation is the slight ceiling effect, observed primarily for decisions on self trials. Previous versions of the prosocial effort task (with grip strength as the effort manipulation) showed greater aversion to the highest effort conditions, even when earning rewards for oneself (Lockwood et al., 2017). In this version, many participants

chose to work for themselves and in-group members on nearly every trial, regardless of effort or reward.

While this alternative explanation cannot be entirely ruled out based on these data, I still maintain that the effort manipulation was robust. This is primarily because all other expected effects associated with effort aversion in previous versions of the task were present. As a secondary consideration, participants also were approximately normally distributed across the different effort calibration levels (with most being placed in the medium calibration, and only several being placed in each of the easy and hard calibrations). Nonetheless, this is a trade-off within the paradigm that I have used: having participants perform many trials allows for a comparison of minute differences in reward and effort level sensitivity, but also means that the effort manipulation cannot be overly taxing as this would elicit fatigue effects after only a few trials (Müller & Apps, 2019). One way for future research to disentangle these competing explanations would be to design a task specifically to employ a more taxing effort manipulation across fewer trials.

Despite the change in effort calibration (with an overall increase in effort required for Study 2.2 than Study 2.1), the winning model remained the same across both samples. This model was consistent with models of previous versions of the Prosocial Effort Task (Lockwood et al., 2017; Lockwood et al., 2021), primarily in that it favoured a parabolic discounting function over a hyperbolic or linear one. The in-group and out-group discounting parameters fitted by the winning model showed more discounting for out-group than in-group choices, and also correlated with self-reported judgments of similarity. However, it is important not to read too deeply into these model parameters, as the non-linear interactions between different parameters make precise interpretations difficult. For example, while discounting was higher for

out-group choices, the model mostly alighted on the most extreme values of 0 or 1.5 (Figure 7B), with few distributed values somewhere in the middle.

Another source of complexity in interpreting these results is the large range of individual differences in choices for out-group targets (Figure 6A). Some participants chose almost always to work for the out-group target, while some chose almost never to work for the out-group—with a fairly uniform distribution of the rest between these extremes. This suggests that whatever the overall trend of the effect might be, there is potential for variation in individual decision strategies.

General Discussion

To my knowledge, this is the first set of studies to directly investigate the effect of social group membership on prosocial motivation. Previous studies on prosocial motivation have tended to focus on willingness to put in effort to help an anonymous individual (Lockwood et al., 2017; Lockwood et al., 2021) or in social contexts which do not necessarily elicit group-based social identity (such as with close family and friends in Guassi Moreira et al., 2020; Guassi Moreira et al., 2021; or with friends and charities in Depow, Lin, & Inzlicht, 2022). This is aligned with Batson's (2022) prescription of a greater need to focus on the role "collectivism" plays in prosocial motivation—that is, an individual's willingness to exert effort on behalf of members of one's social group.

One consistent finding from this prosocial motivation literature (e.g., Lockwood et al., 2017; Depow, Lin, & Inzlicht, 2022) is a self-bias: people are more willing to put in effort for rewards that go to themselves than those that to other agents. However, this tendency for self-bias is to a great extent ameliorated in the present studies. While Study 2.1 showed a significant

(though very slight) difference between choices for oneself and the in-group target, the sample performing the same task with recalibrated effort levels in Study 2.2 showed no statistically significant difference. In both studies, the model comparison favoured a model where the same discounting parameter accounted for choices in both self and in-group trials. This suggests that under certain circumstances people can value rewards for others—at least in-group members—very closely to how they value rewards for themselves. As noted in the discussion section above, this task prioritises relatively minor effort exertions over many trials (rather than large effort exertions over few trials), and so may be more representative of, say, one's willingness to hold a door open for a passing individual than one's willingness to donate a kidney.

In-group "bump" vs out-group "dip."

I presented three hypothetical patterns of behaviour which could be borne out in these data (Figure 1), though I did not pre-register strong a priori hypotheses about which of these accounts would best explain the data. The results were consistent with the hypotheses of Figure 1A. Reward valuation for in-group members did not significantly differ from valuation of reward for oneself (i.e., an in-group bump). Reward valuation for out-group members showed significant increase in discounting (i.e., an out-group dip). This finding is consistent with the general trend of research on prosocial and altruistic behaviour that the desire to achieve both positive outcomes for in-group members and outcomes for out-group members play a role in prosocial decisions (Mummendey and Otten, 1998; Hewstone, Rubin, & Willis, 2002; Halevy et al., 2012).

Previous work suggests that people prefer to cooperate to maximise absolute group gains rather than compete to maximise relative gains over an out-group (Halevy et al., 2012), and indeed in the present study participants do show exhibit a willingness to exert effort on behalf of

out-group members in approximately 50% of trials (Study 2.2; Figure 6A). There is even an effect of reward level in out-group decisions (Figure 7, third panel), which shows that participants chose to work for out-group members on trials with higher rewards at stake. An alternative pattern of behaviour that could have arisen is that participants could have chosen not to work for out-group members at all. Rather, they did appear to be maximising their choices for the out-group's gain (as opposed to never working for them) conditional on the relevant levels of effort-based discounting observed in the data. Certainly, it is important that this task does not feature a zero-sum game (such as an ultimatum game or dictator game), and the rewards for out-group members do not come at the expense of rewards for oneself or one's in-group. Additionally, it is important to note that there is quite a bit of individual variation in participants' decision strategies on out-group trials. Further research will be required to tease apart individual motivations for choosing whether or not to exert effort on behalf of one's out-group.

One other potential concern about these data is the possibility of comparing the three present agent conditions (self, in-group, and out-group) with a fourth: a non-group-affiliated target agent. On the one hand, the presence of this non-group agent would allow for a comparison with decisions for the out-group. Would participants choose less for the out-group agents than they would for unaffiliated agents? This version would be another means of using this paradigm to investigate another version of this question (i.e., whether in-group and out-group valuations are above or below a baseline rewards value for non-group-affiliated other agents). However, there are downsides to this variation as well. One is that the experiment at present is already fifty minutes long, and adding another 75 more trials would increase it to over an hour (while reducing the number of trials in each condition would have produced insufficient data to ensure the robustness of the models). This increase may result in fatigue effects (Müller

& Apps, 2019). It would also be the case that this altered version cannot provide data that would falsify the in-group bump effect (i.e., in-group decisions would still likely resemble self decisions). Finally, while there is precedent for the use of a non-group-affiliated agent in a minimal groups paradigm (e.g., Cikara et al., 2014), it was my intention to gear this study toward investigating differences in decisions based on a standard two-group scenario.

Internal validity vs external validity.

Another trade-off worth noting in this paradigm is that between internal and external validity. I contend this paradigm is highly internally valid, as it successfully controls for confounding effects such as temporal discounting, fatigue, individual variability in the amount of effect exerted (e.g., in the Introduction [Chapter 1], I discussed how the empathy selection task cannot control exactly what the participants are doing when they select the empathy deck), and related concerns. However, achieving this level of experimental precision often comes at the cost of resemblance to situations humans are likely to encounter in real-world scenarios (McGrath, 1981; Finkel, Eastwick, & Reis, 2017). Indeed, even the usage of minimal groups over real-world groups prioritises laboratory control over ecological generalisability. Prosocial decisions in an intergroup context as especially susceptible to ecological variation, for example with differing cultural norms (Kok et al., 2020) or social norms (Pech, Gishoma, & Caspar, 2022) prescribing increased prosociality toward out-group members. Ultimately, this reflects both a strength and a limitation of this approach: a high level of experimental precision is achieved, but its generalisability may be limited in individual cases of ecologically contextualised circumstances.

Additionally, many factors (besides the ones tested or controlled for here) are known to influence an individual's prosocial tendencies. For example, the time of day an individual performs a task results in systematically different levels of prosociality (Francis et al., 2020); inclination toward prosocial behaviour also varies throughout the lifespan, with older adults being more likely to engage in prosocial acts (Lockwood et al., 2021); there is evidence of gender variation in prosociality (Eagly, 2009); and even within a national culture, regional stereotypes can prime individuals for different levels of prosociality (Gavreliuc, Gavreliuc, & Semenescu, 2020). So while inclusion of any of these factors could conceivably impact the results of these experiments or their generalisability in a given social circumstance, there is no evidence that they affect an individual's levels of prosocial motivation as a function of group membership. The difference between the key conditions is still solely about group identity.

Exploring the role of perceived similarity.

Participants' reports of perceived similarity strongly correlated with key measures from participants' behaviour and fitted model parameters. This suggests a potential account of where these intergroup effects are coming from: if the minimal groups paradigm induces participants artificially to feel more similar to one agent than another, does that explain the extent to which they choose to earn points for that agent? In the following chapter, I will explore this question by investigating correlations between perceived similarity and prosocial motivation in an independent data set, and by experimentally manipulating perceived similarity itself in a version of the Prosocial Effort Task.

In summary.

Across these two studies, I have shown how current work in prosocial motivation can be extended to an intergroup context. These data suggest an account consistent with both an in-group bump and out-group dip, where participants treated decisions for their in-group members similarly to decisions for themselves while heavily discounting rewards for out-group members. I have also shown evidence of a potential connection between this intergroup bias and participants' perceived similarity with the other agents, a possibility which I will explore in greater depth in the following chapter.

CHAPTER 3:

PROSOCIAL MOTIVATION AS A FUNCTION OF PERCEIVED SIMILARITY

Abstract. *In this chapter, I investigate the effect of perceived similarity with a target agent on an individual's willingness to exert effort in a prosocial capacity. In the final study of the previous chapter (Study 2.2), I presented data which showed that a participant's self-reported perceived similarity with the target agent predicted both task behaviour and model-based parameter estimates. In Study 3.1 of this chapter, I replicate this finding in an independent data set—extending the finding to show that the effect holds for younger individuals, but not older ones. In Study 3.2 and 3.3, I present the results of experiments in which perceived similarity is manipulated directly. In Study 3.2, participants were given the opportunity to make decisions on behalf of a high- or low-similarity target agent—for which there were no significant effects. In Study 3.3, I replicated the same similarity-based paradigm while increasing the discrepancy between the high- and low-similarity agents. This modification yielded significant effects of perceived similarity. I conclude by arguing that these studies represent a novel test of the influence of perceived similarity on prosocial decision-making—without the confound of information or cues about latent group membership. I argue that these data suggest that while perceived similarity may have a minor effect on prosocial decisions, the explicit structure of group boundaries is a much stronger driver of prosocial behaviour.*

Introduction

In the studies reported in Chapter 2, I found that participants showed different patterns of prosocial motivation for in-group and out-group members. What drives those intergroup effects? One hypothesis, which I explore in this chapter, is that people perceive themselves to be more similar to in-group members (even in the absence of information about an individual's traits) and that this perceived similarity drives increased prosociality. To what extent are people inclined to help in-group members simply because they perceive them to be more similar to themselves than out-group members?

Homophily, perceived similarity, & prosocial behaviour.

Lazarsfeld & Merton (1954) formalised the idea that interpersonal similarity plays a key role in human social relations; they called the phenomenon "homophily." In particular, they believed this was a key determinant in the formation of friendships: people tend to become friends with other people who are like themselves (for the most influential modern account of homophily, see McPherson, Smith-Lovin, & Cook, 2001). Researchers in the second half of the twentieth century picked up on this idea and applied it not only to friendship, but to mating—"homogamy" (e.g., Secord & Backman, 1964; Novak & Lerner, 1968). A meta-analysis by Montoya, Horton, & Kirschner (2008) showed that within this literature (460 effect sizes from 313 laboratory and field studies) both actual similarity ($r = 0.47$) and perceived similarity ($r = 0.39$) were significant predictors of interpersonal attraction (though the authors do not state whether it is possible to dissociate overlapping effects, e.g., where couples are both actually similar and perceive themselves to be). However, the effect of actual similarity was most important in no-interaction and short-interaction studies, and became sharply attenuated in

studies looking at longer interactions. The effects of perceived similarity, meanwhile, were maintained regardless of length of interaction.

Homophily has also been investigated as a predictor of helping behaviour. In one early study, researchers approached people in a telephone booth, asking to borrow a dime in order to make a call (Emswiller, Deaux, & Willits, 1971). They categorised people into one of two types—"hippie" or "straight"—as well as by their gender. A researcher would approach a subject of either matched or mismatched gender and type. The comparison of the matched and mismatched conditions supported their hypothesis that people would be more likely to give a dime to people who resembled themselves. More recent studies have shown that perceived similarity correlates with compassion-motivated behaviour, while pride-motivated behaviour correlates with perceived similarity only with those higher in a social hierarchy (Oveis, Horberg, & Keltner, 2010); that people are more willing to work with a robot when they perceive that robot to be similar to themselves (You & Robert, 2008); that human infants showed preferential looking responses toward agents who either treated similar others well or dissimilar others poorly (Hamlin et al., 2013); that perceived similarity mediated willingness to help immigrants, where similarity was defined on an ethnic basis (López-Rodríguez, Cuadrado, & Navas, 2017); that perceived similarity predicts defection rate in a prisoner's dilemma task (Fischer, 2009); and that perceived similarity predicts response in the ventral striatum—associated with positively reward for oneself—when watching another individual receive rewards (Mobbs et al., 2009).

However, I contend that the above studies are representative of two general issues found in this literature: The first is that perceived similarity is typically assessed *indirectly*, as a self-report measure. In the vast majority of studies, participants are asked whether they regard themselves as similar to a target agent, rather than having that perception of similarity

experimentally manipulated by the researchers. This is also true of the correlational similarity results from Study 2.2 in the previous chapter: participants performed an experiment designed to manipulate and measure one variable—their willingness to engage in prosocial behaviour—and then that variable was shown to be correlated with self-reports of perceived similarity.

The second problem is that the dimensions of assessed "similarity" are typically confounded, most often with group membership or social identity. For example, the classic study by Emswiller and colleagues (1971) separated participants into two social categories ("hippie" and "straight"). Did the participants behave in a certain way because they felt "similar" to the other agents, or because they believed themselves to be members of the same social group? By a similar token, López-Rodríguez and colleagues (2017) defined similarity on an ethnic basis. Again, it is unclear whether people are making decisions because they perceive themselves as similar, or because they believe themselves to be members of the same group. As long as a measure of similarity implies information about membership in social groups, it is not possible to distinguish between these two separate factors. A key focus of my work in this chapter will be to experimentally manipulate perceived similarity without introducing information about social groups—thereby allowing me to disentangle the distinct roles of perceived similarity and group membership in prosocial motivation.

Manipulating perceived similarity directly.

Several other studies have attempted to manipulate perceived similarity directly. Krebs (1975) split participants into two groups: one was led to believe they were paired with a person who was interpersonally similar, and the other with a person who was dissimilar. The participants were told that the initial determination was done based on the results of a personality

survey, while also being given other cues, such as the confederate hinting at a mutually-enjoyed activity or casually professing belief in a shared value. Krebs showed that participants in the high-similarity condition were more likely to exhibit an empathic response (based on self-report, as well as psychophysiological measures, such as heart rate and skin conductance) than participants in the low-similarity condition.

However, Batson and colleagues (2005) tested a similar hypothesis, designing an experiment to manipulate similarity directly while measuring the resulting variation in empathic response. In contrast to Krebs, they found no evidence that one's perceived similarity with a stranger affects an individual's empathic response. But they did find a significant effect of nurturance—whether the target individual elicited an impulse to care for and protect offspring. Batson and colleagues argued that empathy is brought about by such a nurturance response, rather than a similarity-based one, implying that this would in turn influence an individual's willingness to engage in prosocial behaviour.

Crucially, neither Krebs nor Batson and colleagues were careful to manipulate similarity without implying group membership. In the study run by Krebs, the participants had face-to-face interactions with the confederate to whom they were supposed to feel similar or different. Krebs reported that the "performer was a 22-year-old graduate student who had recently arrived in the area. He was visible to the subjects from the chest up." Given that the 60 subjects were males with an average age of 20, it seems likely group-based identities such as gender and age were important factors. Likewise, Batson and colleagues manipulate similarity by describing the target agent as having the same or different gender, age, and occupation as the participant. These factors are indeed a basis for similarity judgments; they are also standard bases on which to form social groups.

Few attempts have been made to satisfactorily disentangle the effects of similarity from those of group membership. Mussweiler and Ockenfels (2013) demonstrated that effects of similarity elicited systematically different patterns of altruistic punishment in an economic game than did group membership. However, their similarity manipulation was one of priming: participants either completed a task in which they were instructed to find the similarities between the pictures or one in which they were instructed to find the differences. Thus, the participants were under the "influence" of similarity-based thinking; but the extent to which they identified as similar with the other agent was not explicitly manipulated. Meanwhile, Lau and colleagues (2018) attempted to show how perceived similarity disambiguates group membership when the true delineations of social structure are not known. But they do not show whether perceived similarity, independent of implied group membership, drives willingness to engage in prosocial behaviour.

Thus despite the long history of similarity-based explanations of altruism, empathy, and related phenomena, to my knowledge no studies have yet investigated the effects of perceived similarity on willingness to engage in prosocial behaviour by manipulating similarity directly, without implying confounding information about group membership. In this chapter, I will present a series of studies which are intended to satisfy those conditions.

Interaction between age and prosocial motivation.

While many studies of prosocial behaviour have tended to look at young adults, some studies have suggested that there are significant differences in prosociality and empathy across the lifespan (Eisenberg et al., 2005; Imuta et al., 2016; Hepach & Warneken, 2018). A recently published version of the Prosocial Effort Task by Lockwood et al. (2021) looked at differences in

prosocial decision-making for younger adults versus older adults—finding that older adults were generally more prosocial. While Lockwood and colleagues asked participants to self-report about their perceived similarity to the other agent in the task, they neither analysed nor reported these results. I therefore sought to replicate the results of Study 2.2 in the previous chapter—which showed correlations between perceived similarity and participant performance in the Prosocial Effort Task—in an independent data set, while also exploring the extent to which differences in perceived similarity drive the age-based differences in prosociality observed by Lockwood and colleagues.

Hypotheses and summary of studies.

In this chapter, I will present three studies. The first one replicates the findings from Study 2.2 in the previous chapter—showing a correlation between self-reported perceived similarity and prosocial motivation, as measured by key behavioural measures and model-based parameter estimates. I will also to use these data to explore whether there is an age-based difference in prosociality on the basis of perceived similarity. In the subsequent two studies, I will manipulate similarity directly (by reporting to participants whether they scored similar results on a personality test with a target agent, about whom they knew nothing besides these experimentally manipulated "scores"). Given that the intergroup version of the Prosocial Effort Task showed strong group differences, using a manipulation of similarity without confounds of group membership will show the extent to which similarity, independent of group membership, affects prosocial motivation. Studies 3.2 and 3.3 differ in how interpersonally similar to the participant they present the target agents to be.

Study 3.1: Perceived Similarity Correlations in an Independent Dataset

Participants. One-hundred ninety-six participants were recruited for the present study. Seven were excluded due to disbelief in the deception, one due to technical error, and one due to incomplete data. The final sample included 187 participants: 95 younger adults (age 18-36, m = 24; 56 female) and 92 older adults (aged 55-84, m = 69; 43 female). A power analysis based on a previous study indicated there was 91% power to detect a medium effect size (d = 0.5) with at least 92 participants in each group. Age ranges were 18-36 in the younger cohort and 55-84 in the older cohort.

Participants were recruited through university databases, social media, and the local Oxford community. All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants with previous or current neurological or psychiatric disorder were excluded, as were participants with non-normal or non-corrected-to-normal vision. The older cohort was given Addenbrooke's Cognitive Examination (Hsieh et al., 2013) to exclude participants with severe cognitive impairment that are more prevalent in older populations. Anyone scoring over 82 was excluded. Participants were paid at a rate of £10 per hour. They were told an addition bonus payment of £5 was possible for both themselves and the other participant at the end of the experiment depending on their performance in the task.

In the analyses that follow, I will not replicate the analyses from the original paper. Any analyses reported are my own, unless indicated otherwise.

Method.

Design. The task structure was identical to the one described in Lockwood and colleagues (2017; Figure 1). Participants completed 150 trials. This included 75 decisions for themselves and 75 decisions for the other person. Each trial involved a choice between a baseline option of 1 credit for no effort and an option that varied in the level of effort and reward. Effort levels corresponded to different percentages of the participant's maximum voluntary contraction (at 30%, 40%, 50%, 60%, and 70%). This was measured by a handheld dynamometer: a device designed to measure the force exerted by a person's grip.

The experiment was designed to minimise potential effects of fatigue: participants had to squeeze for 1 second out of a 3 second window to earn the reward. Each trial lasted 10 seconds. Participants were only required to exert grip-force effort when they chose to higher-effort option. It should be noted that many of the "higher-effort" trials were not especially demanding; anything below 50% maximum voluntary contraction requires a negligible amount of contraction

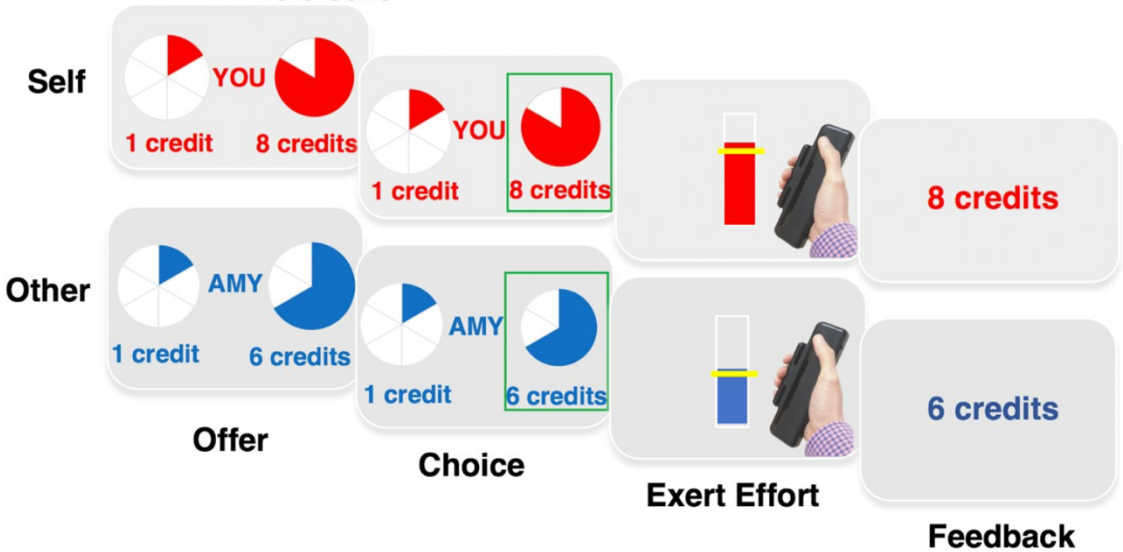


Figure 1 — Schematic of the Prosocial Effort Task (from Lockwood et al., 2021). Participants completed trials on which points other either with to themselves or another agent. On each trial, participants chose between a lower-effort baseline option and a higher-effort option which varied in effort and reward level. Effort was operationalised by force exerted via grip. Feedback was given on each trial.

to hold for a single second. Each participant also had three breaks of indefinite length. Trials were counterbalanced such that the same number of self and other trials with equivalent effort and reward levels were presented in blocks of 50. Thus, any potential effects of fatigue would occur equally in both the self and other conditions.

Apparatus. Stimuli were presented on a PC using MATLAB (The Math-Works Inc., USA) and Psychtoolbox. The dynamometer was model TSD121B-MRI from BIOPAC Systems Inc., USA.

Role assignment procedure. This was identical to the procedure followed by Lockwood et al. (2017) and reported in full by Lockwood et al. (2021). Participants were told their choices resulted in monetary benefits for another individual. They were told their counterpart (actually a confederate of the experimenter) was in an adjacent room. Both "participants" selected a ball from a box, which would "assign" them to one of two roles. Each wore a glove and was asked not to speak, to avoid learning any identifying information about one another. The confederate was led by another "experimenter" to the other side of the doorway in which the participant completed the experiment. Both individuals were asked to wave to one another, with gloved hands, to ensure each other's presence. The experiment tossed a coin to determine who would pick first. Each participant then selected a coloured ball, which "determined" their role. The participant's role was always the "decider." The confederate was the "receiver."

Dynamometer calibration. Participants began by squeezing a handheld dynamometer with as much force as possible to determine their maximum voluntary contraction. This accounted for any baseline differences in participants' grip strength. This measurement was used to determine individual participant's effort levels.

Task procedure. In each trial, participants made decisions between a baseline low effort offer that earned one point, and an offer that varied in points (2, 4, 6, 8, 10) and required effort (30%, 40%, 50%, 60%, 70% of maximum voluntary contraction). Effort and reward levels varied independently over trials. Each effort-reward combination was done three times by each participant. This resulted in 150 trials in total: 75 for self and 75 for other. To earn points in the higher effort option, the participant had to achieve the specified level of effort for one second within a three second window. Failure to do so resulted in zero points earned, as did non-choices. The baseline option was set at one point to incentive participants to choose this option if the higher effort option was not worth it, rather than choosing the higher effort and not exerting the force. All trials lasted the same duration to ensure there were no temporal discounting effects. It should also be noted that participants exerted the required amount of force on >95% of trials. This suggests participants were confident in being able to achieve the reward once they had decided it was worth earning, thus ruling out potential effects of risk aversion.

Pre-task self-report judgments. Before starting the main experiment, participants were asked to make a self-report judgment on a scale of 0-9: "How similar do you feel to the other participant?"

Results.

The primary results from these data were reported in Lockwood and colleagues (2021). This included an analysis of participants' choice behaviour and a computational model of effort discounting fit to their task performance. Lockwood and colleagues found that older adults

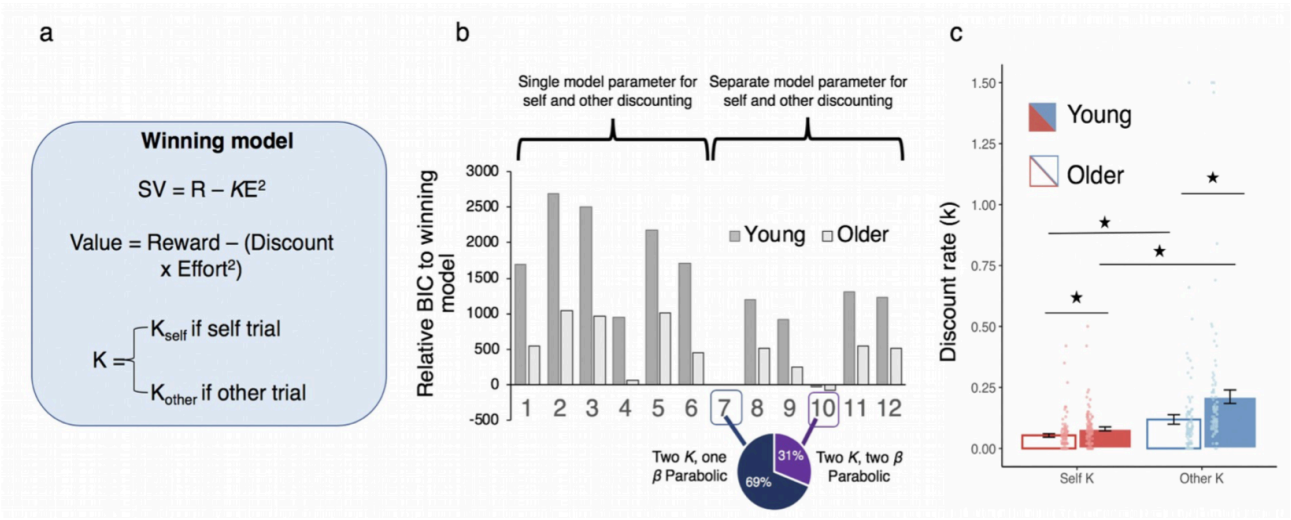


Figure 2 — Older adults discount rewards by effort less than younger adults, particularly for others (from Lockwood et al., 2021). (**a**) The winning model based on participants' task performance was comparable to the winning intergroup model from Chapter 2: it featured a parabolic discounting function, with separate discounting parameters values for self and other. (**b**) Model comparison, based on group-level BIC. (**c**) Younger adults showed higher discounting rates than older adults. These model parameters were consistent with the behavioral findings—that older adults tended to earn more rewards for the other agent than did younger adults.

showed less overall discounting of rewards for other agents (Figure 2). However, Lockwood and colleagues left the results of perceived similarity self-report assessment unanalysed. Using these data, I investigate the extent to which perceived similarity predicted key measures of prosocial motivation from the main experiment.

Perceived similarity predicted task behaviour and model parameters.

I ran a similar suite of statistical tests to those presented in Study 2 of Chapter 2 to assess the extent to which perceived similarity predicted key measures of prosocial motivation in these data (Figure 3). To test whether perceived similarity predicted participants' task behaviour, I ran a linear mixed-effects regression, which predicted the number of credits won for self or other

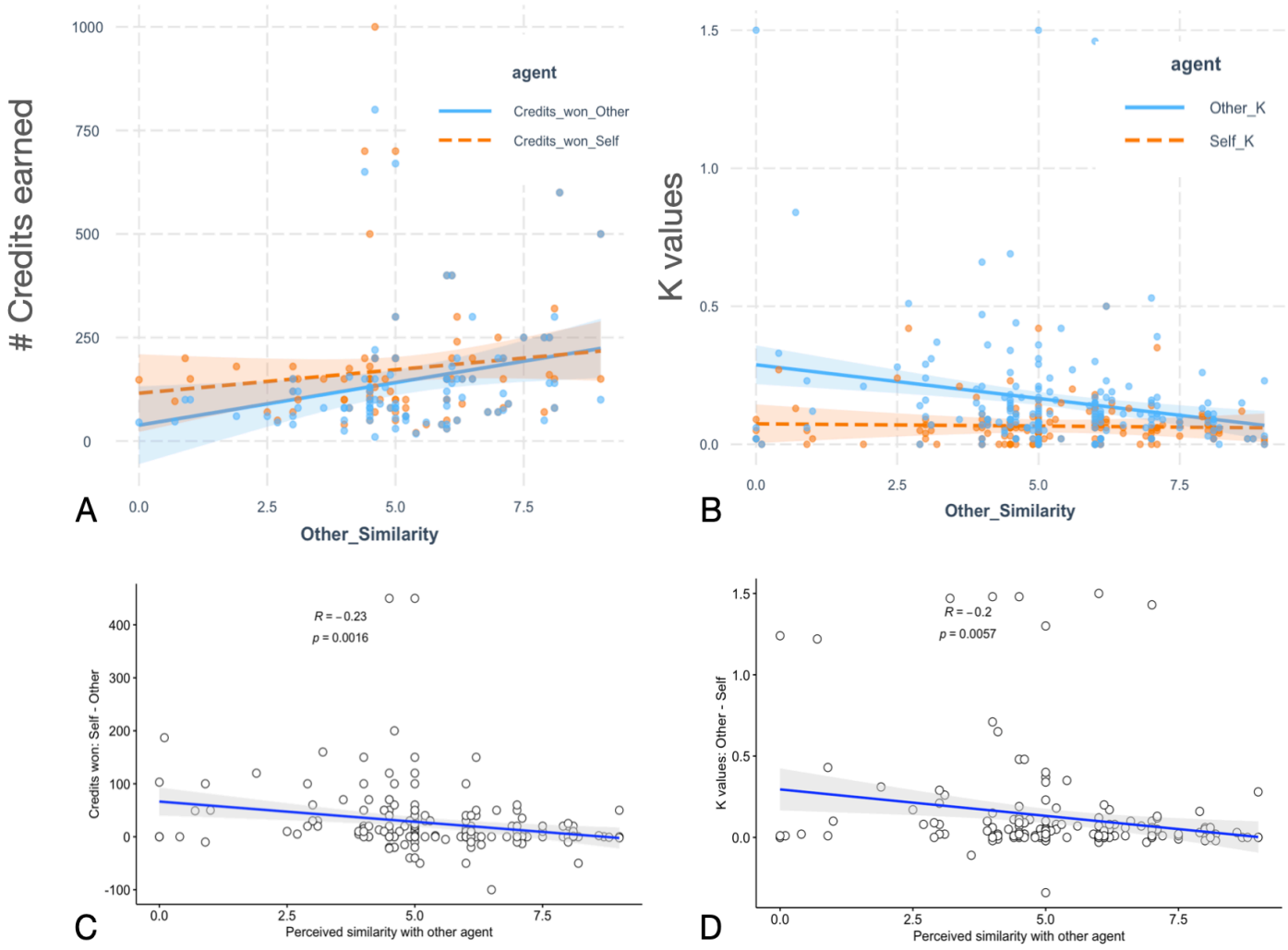


Figure 3 — Perceived similarity of other agent predicted task behaviour (left panels) & model parameters (right panels). (**A**) A linear mixed-effects regression predicting number of credits earned for self and other showed a significant interaction between fixed effects of agent and perceived similarity. (**B**) Likewise, a regression predicting model effort discounting parameters also showed a significant interaction between agent and perceived similarity. I also found that perceived similarity predicted (**C**) the difference in credits won for self versus other and (**D**) the difference in model discounting values.

based on fixed effects of agent (self, other) and perceived similarity with the other agent, as well as a subject-level random intercept. This analysis showed a significant similarity-by-agent interaction (b = -7.640 [-12.302 – -2.977], z = -3.212, p = 0.001; Figure 3A). I ran a regression with the same fixed effects to predict the model discounting parameters, extracted from the

winning computational model with separate discounting parameters and noise parameters (see Lockwood et al., 2021). This regression also showed a significant similarity-by-agent interaction (b = 0.023 [0.007 – 0.039], z = 2.749, p = 0.006; Figure 3B). In line with these regressions, I also found that similarity was a predictor of the difference in credits won for self versus other (r = -0.23, p = 0.002; Figure 3C), as well as the difference between self and other discounting parameters (r = -0.20, p = 0.006; Figure 3D). The larger the difference between self and other discounting parameters, the less similar participants reported themselves to be; this was due to decreased prosocial decision-making for the other agent, as predicted by perceived similarity. In a conceptual replication of my correlational finding in Study 2.2, these results suggest that participants demonstrated a more prosocial pattern of decision-making when they reported feeling greater similarity to the other agent—according to both their behavioural data and the computational model.

This effect was driven by younger participants.

I fit an additional linear mixed-effects regression incorporating the effect of cohort (young versus old). The analysis yielded a three-way interaction of similarity, agent, and age cohort (b = -0.037 [-0.069 – 0.005], z = -2.260, p = 0.024). Perceived similarity predicted key task behaviour and model parameters in younger, but not older, adults (Figure 4). Younger adults showed the same pattern of effects as the analysis of the whole dataset reported above: a linear mixed-effects regression showed a significant interaction between agent and similarity for predicting model discounting parameters (b = 0.042 [0.016 – 0.068], z = 3.220, p = 0.001; Figure 4A) and number of credits earned (b = -9.277 [-15.998 – -2.556], z = -2.705, p = 0.007); and a correlation showed that similarity predicted between-agent difference in model discounting

parameters (r = -0.34, p < 0.001; Figure 4C) and between-agent difference in number of credits earned (r = -0.27, p = 0.008).

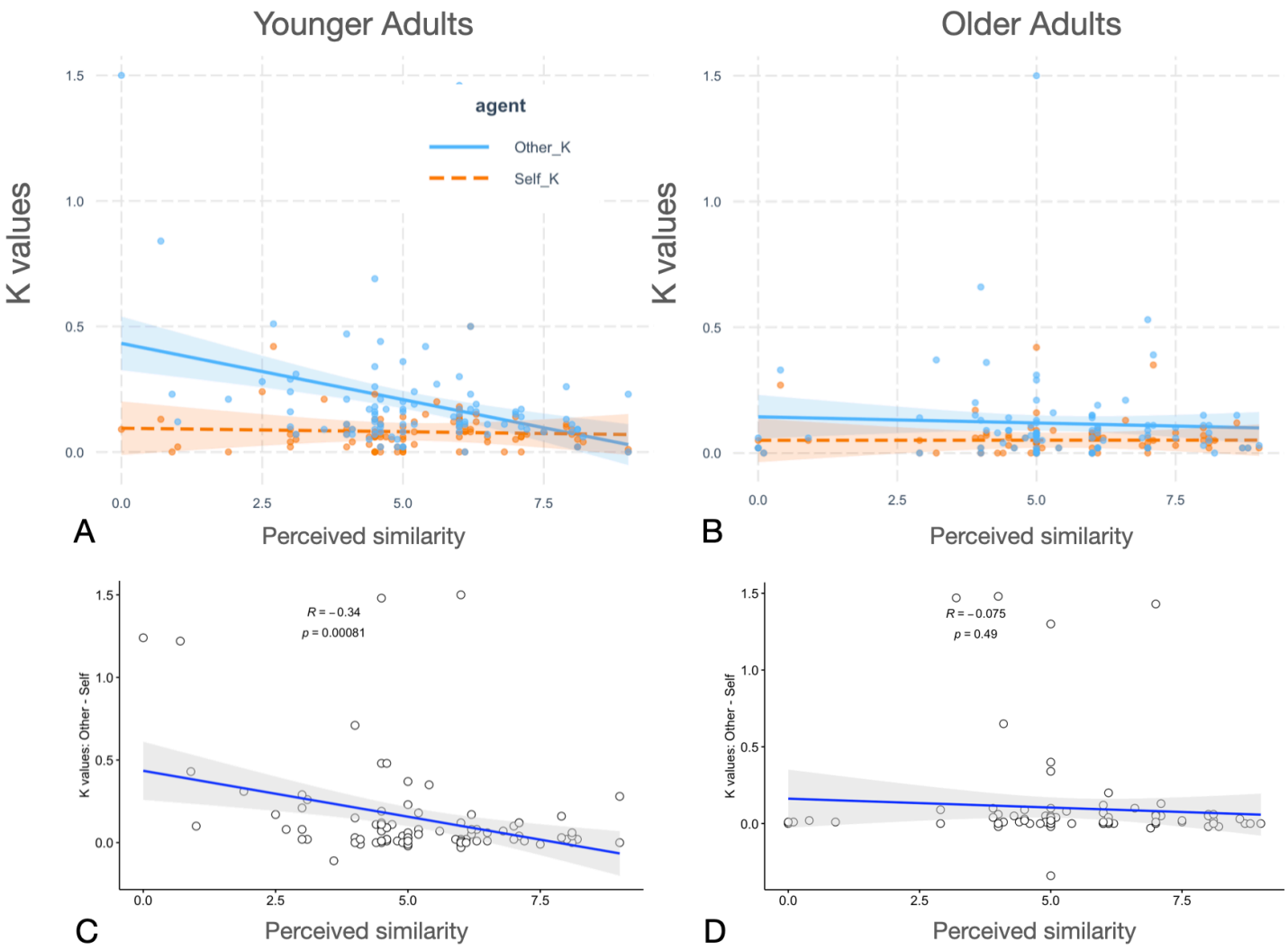


Figure 4 — The similarity effect is driven by younger participants. Three-way interaction between perceived similarity, target agent, and age. (**A**) Model effort discounting parameter (k; the rate at which effort becomes "not worth it") as a function of perceived similarity, for both self and other. Whether perceived similarity predicts the model parameters depends on the condition. (**B**) The interaction is ameliorated in older participants. In neither condition does perceived similarity predict model parameters. (**C**) In the younger cohort of participants, perceived similarity with other agent predicts difference in k values between Other and Self conditions. Larger differences in k values indicate a less prosocial strategy (e.g., minimal self discounting and maximal other discounting would yield value of 1.5—per model parameter definitions). (**D**) In the older cohort of participants, perceived similarity does not predict model parameters.

Older adults showed none of these effects: a linear mixed-effects regression showed no significant interaction between agent and similarity for predicting model discounting parameters (b = 0.005 [-0.015 – 0.025], z = 0.496, p = 0.620; Figure 4B) or number of credits earned (b = -6.114 [-12.658 – 0.429], z = -1.831, p = 0.067); and a correlation showed that similarity was not a significant predictor of between-agent difference in model parameters (r = -0.08, p = 0.49; Figure 4C) or between-agent difference in number of credits earned (r = -0.19, p = 0.07).

Discussion.

These results give further support—albeit correlational—to the hypothesis that perceived similarity might be an important driver of prosocial decisions. They also suggest that this effect might vary with age: the prosocial decisions of younger participants depended much more heavily on perceived similarity than did the prosocial decisions of older participants.

Study 3.2: Manipulation of Perceived Similarity in Prosocial Effort Task

As reported in Study 2.2 and Study 3.1, two independent datasets produced by similar experimental paradigms provide correlational evidence suggesting that perceived similarity may be a crucial driver of prosocial motivation. In the following two experiments, perceived similarity is directly manipulated in the experimental design. The key question tested here is whether perceived similarity can drive differences in prosocial decision-making of its own accord; I designed the experiment with the specific objective of distilling the effects of perceived similarity on its own, rather than as the by-product of other manipulations (e.g., in a minimal groups paradigm) or the participant's baseline assumptions (e.g., when participants are given no identifying information about the other agent).

Preregistration. I preregistered methods, recruitment, analyses, and hypotheses for this experiment at https://osf.io/3gfxm. The hypotheses in this preregistration were based on other versions of the Prosocial Effort Task (e.g., Lockwood et al. 2021) and my a priori hypotheses.

Participants. Forty participants were recruited on the Prolific subject recruitment platform (19 female; mean age = 24.0). All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately forty minutes to complete.

Method.

Design. This implementation of the Prosocial Effort Task was the same as the one described in Study 2. 2 with one main difference: instead of the intergroup manipulation, I used a similarity manipulation (Figure 5), described below.

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Prolific subject recruitment platform.

Procedure. The procedure was the same as described in the "Intergroup Prosocial Effort Task" experiments, with the substitution of the similarity manipulation instead of the group induction.

Similarity manipulation procedure. The aim of this procedure was to manipulate participants' perceived similarity with the other agents—while avoiding introducing any information about potential social groups. I based the procedure loosely on the similarity learning paradigm described in Gershman & Cikara (2020), who sought to employ a related

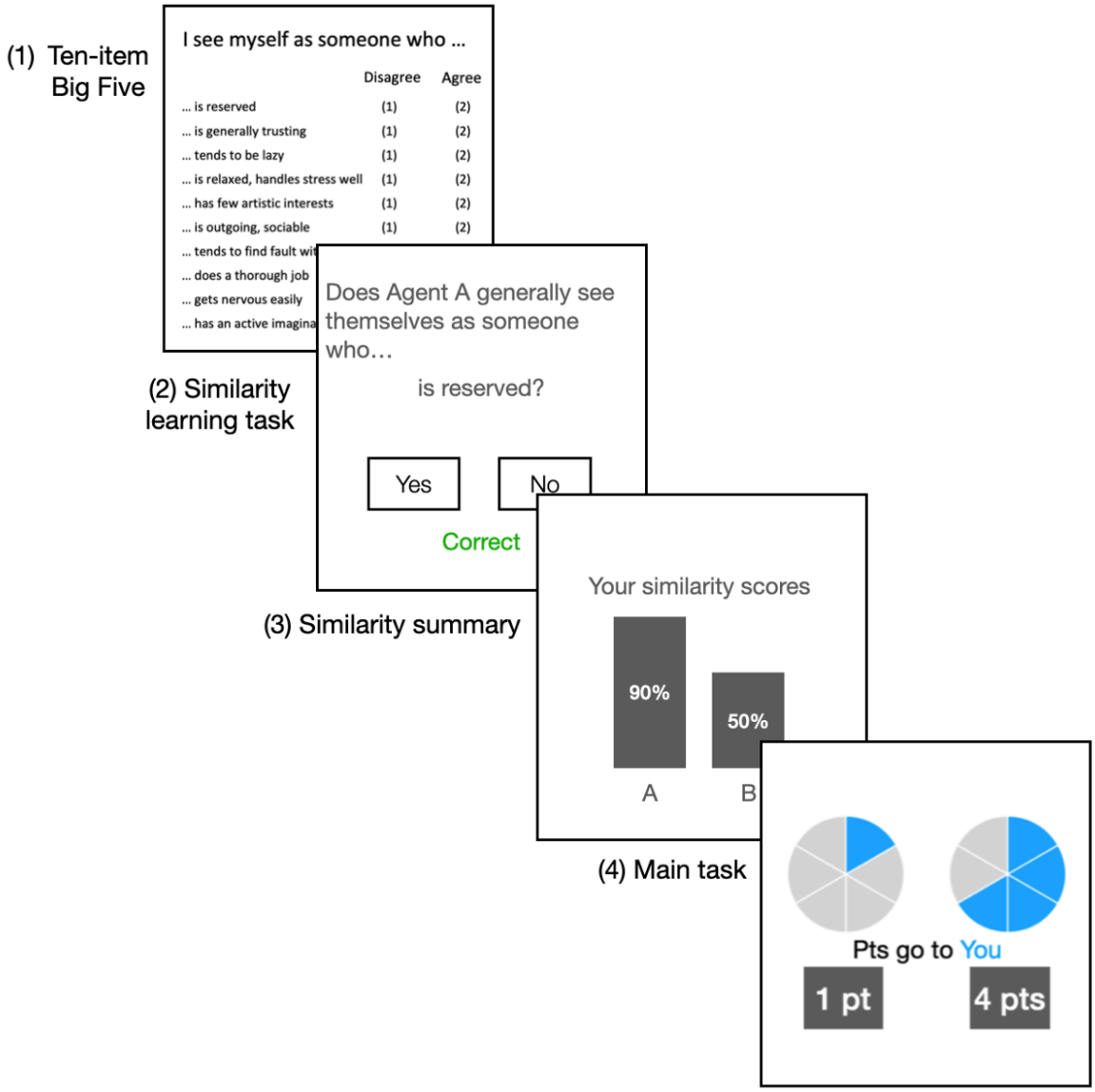


Figure 5 — Experiment schematic for manipulation of perceived similarity in the Prosocial Effort Task. (**1**) The procedure begins with the participant filling out a 10-item version of the Big Five inventory. (**2**) This is followed by a learning task, in which the participants learns the answers of two other "agents." (**3**) The participant is told that one of these agents scored 90% similarity with the participant, the other 50%. (**4**) This is followed by the main task, which is identical to the structure of the Intergroup Prosocial Effort Task used in Study 2.2.

experimental manipulation (in their experiment, they imply similarity information about agents without directly specifying any group structure, then ask the participants make inferences about what the possible group structure might be). In particular, Gershman and Cikara employ a

learning paradigm wherein participants learned about the preferences of other agents for individual movies across several trials—rather than simply being summarily informed which agent had more similar preferences in the final tally. In their paradigm, Gershman and Cikara used movie preferences as the dimension of similarity; I opted to use personality scores, instead of movies, as cinematic preferences can give strong clues about a viewer's demographic makeup.

To begin, participants were asked to fill out a 10-item version of the Big Five personality assessment (Rammstedt & John, 2007). I presented the assessment with binary options (i.e., "Agree" or "Disagree") rather than the standard gradation of options (e.g., "Strongly agree"). This was in order to give the participant binary feedback (correct or incorrect) in the similarity learning trials. Participants were told that this quiz was a strong indicator of how similar they were likely to feel to another person if they were to meet them in real life.

Participants were then presented with a list of nine "possible names" of "other participants" drawn from a "previous version of this experiment." These names were "hidden" behind individual tiles labelled by a question mark. No matter which tile the participant picked, they were told they had chosen a participant named "Kadvvvz"—a name which has the benefit of conveying no identifying information. The name "Kadvvvz" and all their data (e.g., their similarity score) were presented throughout the experiment in the colour blue. Participants were then presented with another group nine of question-marked "participant" tiles. Having selected one of the tiles, the participant was told they had chosen "Cpdnam", which was presented throughout the experiment in the colour red.

Then the participant underwent two blocks of "similarity learning" trials. In each trial, they were asked to guess what the other agents had answered on the Big Five questions. For example: "Does Kadvvvz see themselves as some who generally… is reserved?" Where "is

reserved" is one of the ten items in the assessment. After guessing at the binary answer ("Yes" or "No"), they received feedback on whether they had responded correctly. The participant was presented with each of the items for each participant, with the items presented the same order as the initial assessment (e.g., Question 1 always came first). Participant order was not randomised, to avoid a working memory demand while learning each agents' answers.

In the second block of similarity learning trials, the prompts about the other participants' answers were preceded by a question about the participant's own answers: "What did you answer on the initial questionnaire? Do you generally see yourself as someone who… is reserved?" To which the participant could respond as they originally did: "Agree" or "Disagree". This functioned both as an attention check (if they had filled out the questionnaire randomly, without trying to respond honestly, they may not be able to recall their initial answers), as well as an attentional cue to compare one's own responses with those of the others. Participants who scored less than 70% correct on the attention check were dismissed from the study (n = 1).

The responses of the other "agents" were experimentally manipulated to achieve a specific level of agreement with the participant. The high-similarity agent "responded" to nine of ten questions the same as the participant. Which question elicited disagreement was randomly selected for each participant. The low-similarity agent "responded" to only half of the question the same as the participant (again, randomly selected). These results were summarised for the participant in a screen which showed them their "similarity scores" with the other agent. In a bar graph, the high-similarity agent was presented, in blue, as having 90% "overall similarity" whereas the low-similarity agent was presented, in red, as having 50% "overall similarity."

In summary, participants learned about the other agents' similarity of "preferences" via two rounds of a short learning task and a graph summarising the similarity scores presented after completing the learning trials.

Post-task self-report judgments. After completing the main experiment, each participant was asked to report how "similar" they perceived themselves to be to each of the two others players for whom they were ostensibly earning points. The intent of the question was to assess whether the test had really induced participants to feel more or less similar to the other agents. In full, it read: "How similar do you feel to Kadvvvz? (This is not an attention check; it's okay if the answer is different than the initial test.)" Participants were then asked to report the scores from the initial assessment: "How similar did the test say you were to Cpdnam? (This IS an attention check; please try to report the results as best you remember them.)" Participants were also asked: "Outside of this task, how similar do you feel to people in general? (For example, if you were to meet an average person on the street, how similar would you likely feel to that person?)" to measure people's baseline assumptions about interpersonal similarity. Responses from each of these questions were collected using a slider that indicated one side as minimal similarity and the other side as maximal similarity.

Results.

To examine whether directly manipulated similarity influenced people's willingness to be prosocial, I analysed participants' choice behaviour in the experiment and fit a computational model of effort discounting to their effort-based choices, to the self, high-similarity and low-similarity conditions.

Participants showed no difference between high- and low-similarity agents.

This first analysis tested whether participants chose the high-effort option more or less often for high- versus low-similarity agents. Figure 6 shows participants' mean probabilities of choosing the high-effort option for self, high-similarity, and low-similarity agents. I ran a linear mixed-effects regression with choice probability as an outcome variable (the percentage of high-effort choices made by each participant) and agent as a predictor variable (self, high-similarity, or low-similarity), including a subject-level random intercept. While there was a significant difference for self (b = 0.623 [0.543 – 0.704], z = 16.175, p < 0.001), there no significant difference between high- and low-similarity choice probabilities (b = -0.487 [-0.616 – -0.358], z = -7.413, p < 0.001).

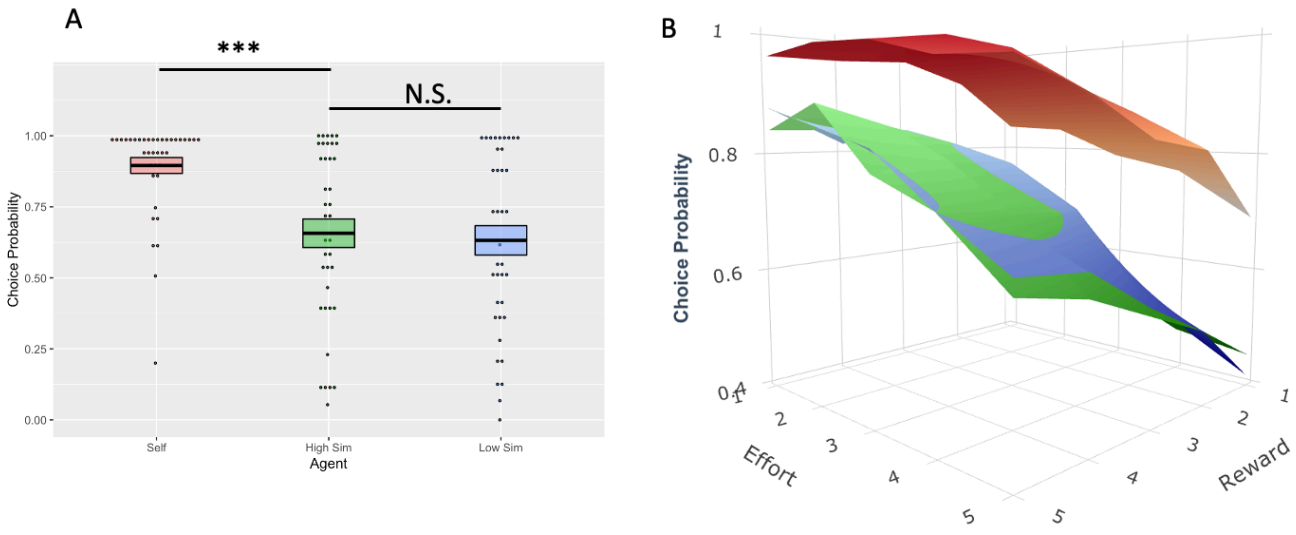


Figure 6 — Choice probabilities of participants' decisions to earn rewards for themselves, the high-similarity agent, or the low-similarity agent. (**A**) Participants almost always chose to earn for themselves. The choice probabilities between the high and low similarity agents were much closer than between self and either of the other agents. (**B**) Choice probability as a function of effort and reward levels.

In line with analysis presented in Chapter 2, I ran general linear mixed-effects model to look for possible interactions (Figure 7). Whereas the linear mixed-effects model presented in the previous paragraph looks at overall choice probabilities across the entire set of trials, this general linear model looks at trial-by-trial behaviour. For this test, I specifically compared only trials on which the participant was choosing for high- or low-similarity agents, and not for self. This would uncover any possible interactions—for instance, if there was a significant difference between high- and low-similarity but only for the highest effort levels. However, the model revealed no such interactions. There were main effects of effort ($X^2 = 407.66$, $p < 0.001$) and reward ($X^2 = 362.13$, $p < 0.001$), but not of agent ($X^2 = 1.49$, $p = 0.222$). There were no significant interactions between effort and reward ($X^2 = 3.67$, $p = 0.4529$), agent and reward ($X^2 = 0.29$, $p = 0.593$), or agent and effort ($X^2 = 1.57$, $p = 0.815$).

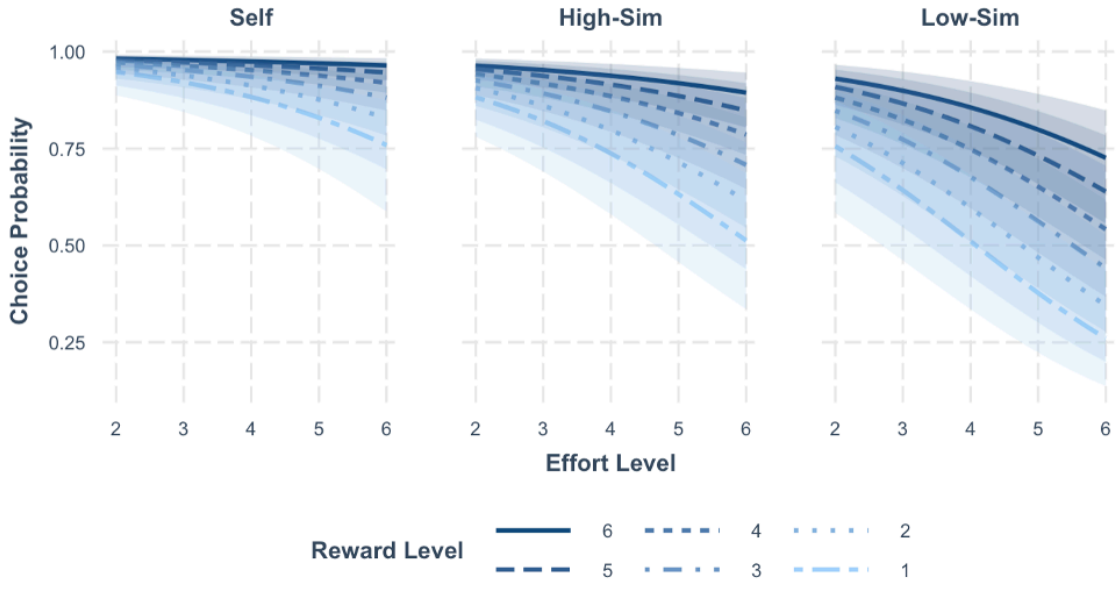


Figure 7 — Generalized linear model. I used a GLM predicting choice probability on individual trials with effort and reward level as fixed effects and participant as a random effect.

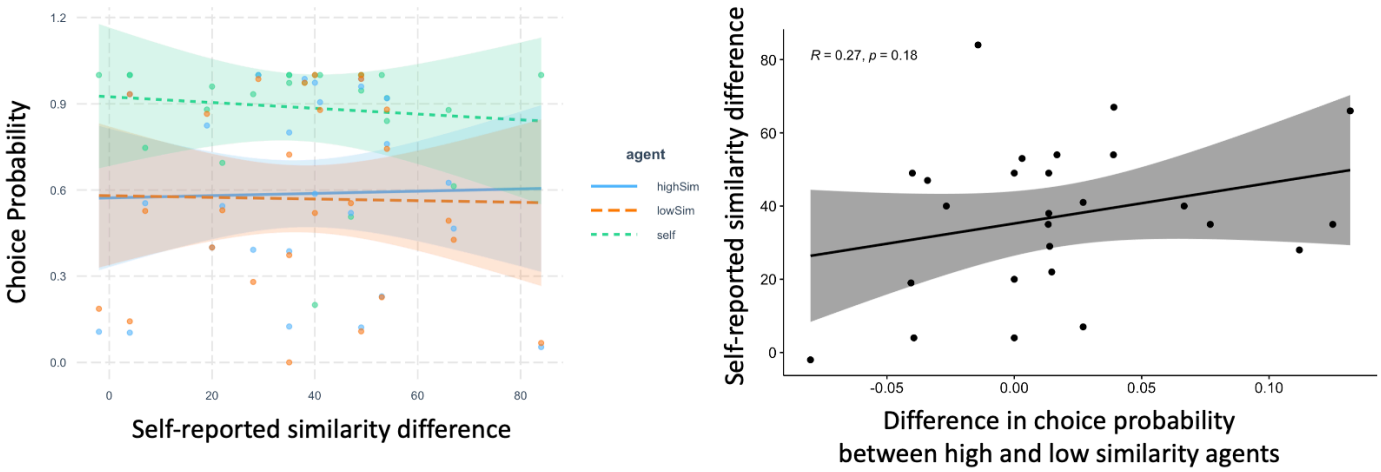


Figure 8 — No significant relationship between the participants' self-reported perceived similarity and their "revealed" perceived similarity, as measured by the difference in choice probability between the high- and low-similarity agents. (**A**) Linear mixed model predicting agent as a function of choice probability and self-reported similarity difference. (**B**) Correlation between the difference in self-reported similarity with the other agents with the difference in choice probability when choosing for those agents.

I also compared the difference between participants' self-reported perceived similarity with the two agents with their behaviour in the task (Figure 8). A linear mixed-effects model predicting choice probability based on agent and difference in self-reported similarity showed no effect of similarity difference (b = -0.002 [-0.005 – 0.002], z = -0.952, p = 0.341). The correlation between the difference in self-reported similarity and the difference in choice probabilities between the agents was not significant (r = 0.27, p = 0.18).

In line with the analysis from Studies 2.1 and 2.2, I also fit a series of computational models to each participant's choice behaviour in the task (Figure 9). I compared discounting parameters in four conditions: baseline; self, other; high-similarity, low-similarity; and self, high-similarity, low-similarity. The model of best fit had only two discounting parameters: self and other.

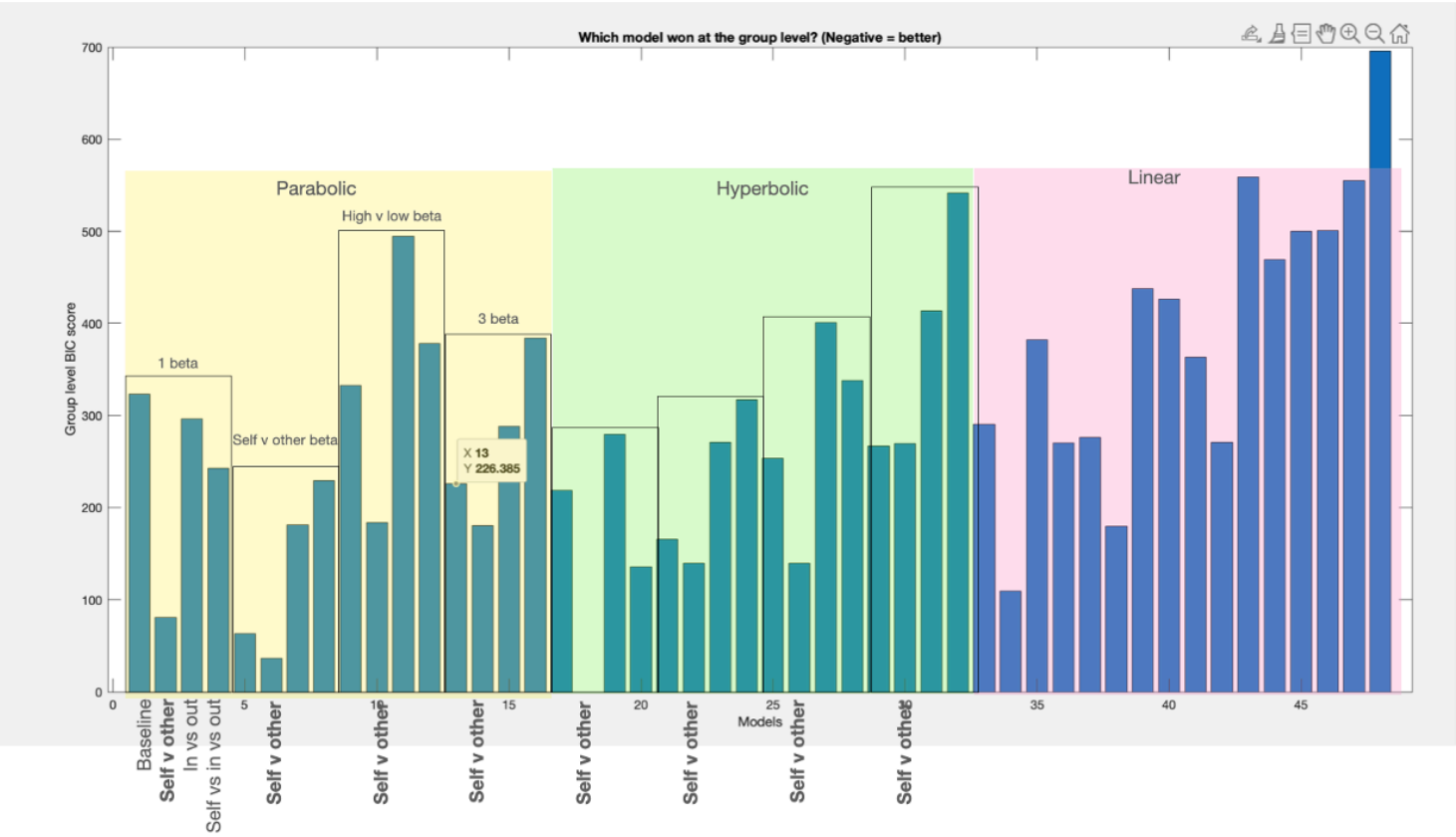


Figure 9 — Modelling showed a consistent preference for self and other discounting parameters rather than self, high-sim, and low-sim discounting parameters.

I also checked for a possible effect of "general" similarity—whereby people might differ in how similar they perceive themselves to be to people in general. If participants were to operate under the assumption that they are very similar to others in general, it might nullify any effects from the manipulation of differences in perceived similarity. I found no correlation between self-reported general similarity and participants' difference in choice probabilities between high- and low-similarity agents ($r = 0.11$, $p = 0.26$; Figure 10). Additionally, participants' mean difference in self-reported similarity between the high- and low-similarity agent was 33.1% (where the similarity difference indicated by the personality learning task manipulation was 40%). Participants also, on average, passed the attention check, reporting a mean difference between the high- and low-similarity agent of 39.1%.

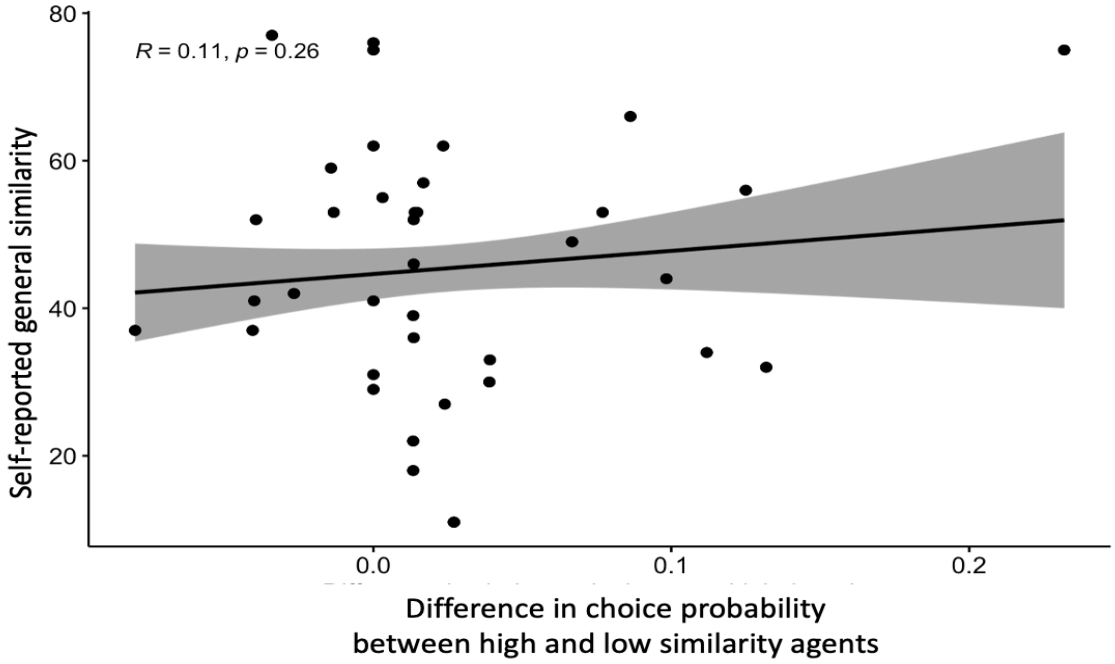


Figure 10 — No correlation between how similar participants report being to "people in general" and their difference in choice probabilities for the high- and low-similarity agents.

Discussion.

These data suggest that participants did not display different patterns of prosocial choices for high-similarity versus low-similarity agents. With the exception of the GLM analysis, which showed a slight variation between choices for high- and low-similarity agents, participants did not significantly differ in their choices for other agents as a function of similarity. There are two possible explanations for this null result: (1) perceived similarity was not a significant factor in participants' prosocial decisions, or (2) the perceived similarity manipulation was not sufficiently effective in convincing participants of their relative similarity with the other agents.

One reason to believe that the manipulation succeeded in making participants feel more or less similar to the other agents is that the majority of participants indicated so in the two relevant post-hoc questions. Most participants indicated that they perceived themselves to be

more similar to the high-similarity agent, and also were able to give a reasonable estimate of their test-revealed similarity differences given at the beginning of the task.

However, one possibility is that our manipulation *was* effective, but that the difference in similarity scores (90:50) was not severe enough to elicit effects. I sought to address that possibility in a follow-up study.

Study 3.3: Manipulation of Perceived Similarity (Increased Difference)

In order to test whether a significant effect could be obtained using this perceived similarity manipulation, I used the same paradigm as reported in Study 3.2 but increased the difference in similarity—from 90:50 percent similarity scores between the high- and low-similarity agent to 90:10.

Participants. Sixty-four participants were recruited on the Prolific subject recruitment platform (25 female; mean age = 26.46). All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately forty minutes to complete.

Method.

The methods in this study were identical to those in Study 3.2—with the exception that participants were shown a larger difference in similarity scores between the two other agents. The participant always scored 90% similarity with the high-similarity agent and 10% with the low-similarity agent.

Results.

Participants showed a preference for the high similarity agent.

Figure 11 shows participants' mean probabilities of choosing the high-effort option for self, high-similarity, and low-similarity agents. I ran a linear mixed-effects regression with choice probability as an outcome variable (the percentage of high-effort choices made by each participant) and agent as a predictor variable (self, high-similarity, or low-similarity), including a subject-level random intercept. In addition to a significant effect of the self condition (b = 0.893 [0.817 – 0.968], z = 23.234, p < 0.001), there was also a significant difference between high- and low-similarity choice probabilities (b = -0.062 [-0.106 – -0.017], z = -2.690, p = 0.007). This stands in contrast to the findings from Study 3.2, in which there was no significant difference between choices for the high- and low-similarity agents.

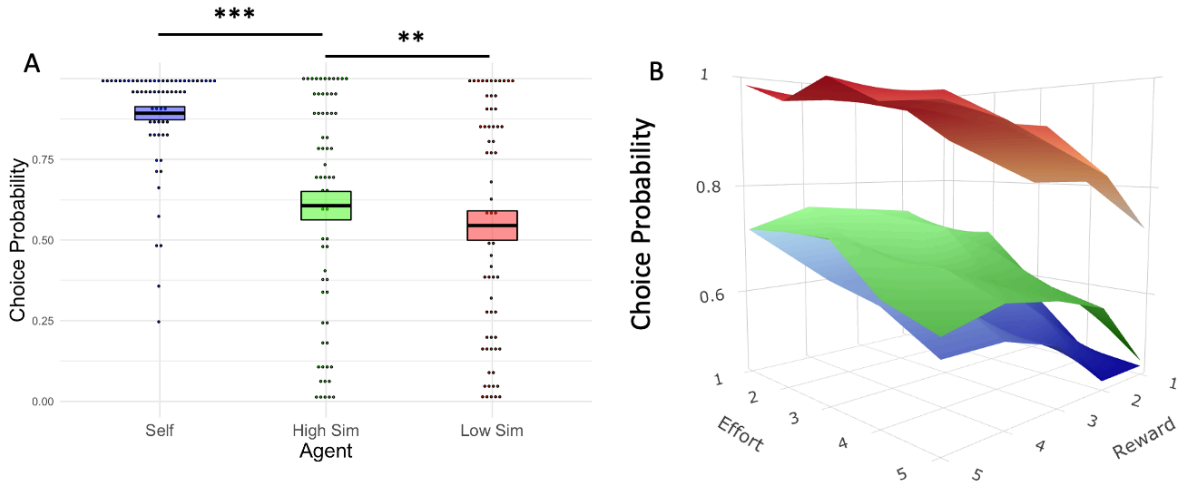


Figure 11 — Choice probabilities of participants' decisions to earn rewards for themselves, the high-similarity agent, or the low-similarity agent. (**A**) Participants almost always chose to earn for themselves. The choice probabilities between the high and low similarity agent were much closer than between self and either of the other agents, though still significant. (**B**) Choice probability as a function of effort and reward levels.

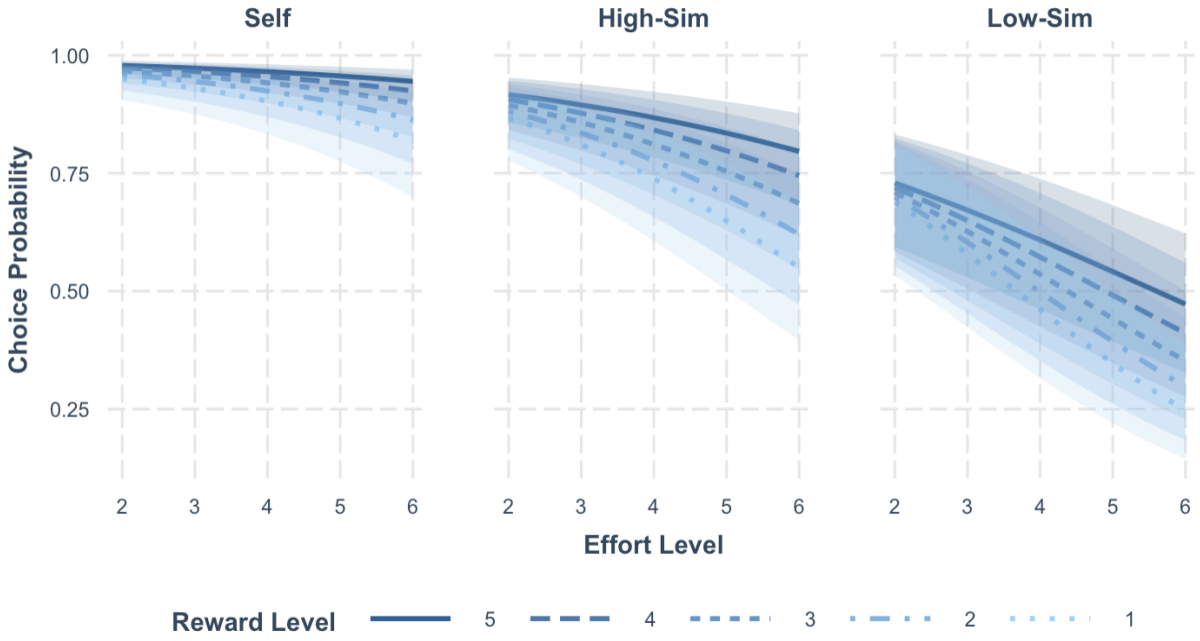


Figure 12 — Generalized linear model. I used a GLM predicting choices on individual trials with effort and reward level as fixed effects and participant as a random effect.

I ran general linear mixed-effects model to look for possible interactions (Figure 12). The model revealed main effects of agent ($X^2 = 1602.05$, $p < 0.001$), effort ($X^2 = 293.18$, $p < 0.001$), and reward ($X^2 = 138.54$ $p < 0.001$). There as also a significant interaction between effort and reward ($X^2 = 11.10$, $p < 0.001$) and between agent and reward ($X^2 = 7.59$, $p = 0.001$), but not between agent and effort ($X^2 = 1.59$, $p = 0.207$). This was the same the pattern of interactions found in the Intergroup Prosocial Effort Task (Study 2.2). The same logic applies: participants were sensitive to effort levels across all conditions (hence the main effect of effort), but they were not more sensitive to changes in effort levels in the different agent conditions (i.e., the slopes of choice probabilities in each panel of Figure 12 are not statistically different).

Discussion.

This final study presented results with the highest possible discrepancy (in units of 10 basis points, without resorting the implausible manipulation of 100% and 0% similarity) between high- and low-similarity agents. In contrast to the previous study, which featured a more subtle similarity manipulation, significant effects of perceived similarity were found: participants were indeed more likely to make prosocial decisions on behalf of high-similarity agents than low-similarity ones. Overall, this indicates a pattern of results which resembles the intergroup findings in Study 2.2, but with a large drop-off in prosocial decisions toward high-similarity agents than in-group agents. I will discuss the implications of these findings in the following section.

General Discussion

The studies presented in this chapter provide evidence for the effect of perceived similarity on prosocial decisions. As I argued in the Introduction to this chapter, the paradigm employed here addressed two overarching issues in the literature on similarity: similarity was manipulated directly, and it was not confounded with cues about potential social group membership. In Study 3.3, it was shown that participants were willing to put in significantly more effort to help high-similarity agents over low-similarity ones. This is in line with correlational evidence from Study 2.2 and Study 3.1, which show that self-reported perceived similarity predicted prosocial behaviour in two separate versions of the Prosocial Effort Task. The overall evidence suggests that perceived similarity does play a role in prosocial motivation—however, a far more subtle one than group membership.

However, in Study 3.2, no significant difference was found between decisions made for high- versus low-similarity agents. One possible explanation for this null finding was that participants did not interpret 50% similarity to be sufficiently "low-similarity" to warrant a different decision strategy. A key distinction for evaluating this potential explanation is between absolute and relative similarity. Study 3.2 used manipulated similarity values of 90% and 50% for the high- and low-similarity agents. If participants attend to absolute similarity values, then one would expect to see them make similar decisions for 50% similar agents no matter what the similarity values of other agents might be. But if participants attend to relative similarity, these perceptions would change depending on the other agent. For example, if manipulated similarity values were 50% for the high-similarity agent and 10% for the low-similarity agent (where the difference in values is the same as the 90:50 similarity values), then participants might treat the 50% agent like the 90% similarity in the current version. With the current data, it is not possible to rule out the explanation that the null result is due to participants not feeling sufficiently dissimilarity to the 50% agent. If the present 90:50 version of Study 3.2 had yielded a positive result, this direct comparison of absolute versus relative similarity (with values of 50:10) would be worthwhile. However, given the present null result any difference between the two versions might be difficult to detect.

All told, I contend that these data suggest that perceived similarity is not the primary driver of differences in prosocial decisions for in-group versus out-group members. While significant effects of similarity were found, they were not nearly as dramatic as the intergroup effects in Study 2.2. Instead, I contend that findings which have been attributed to similarity are actually due to implication of social groups—which do elicit large effects. A crucial aspect of this argument is that the similarity manipulation used in the present studies offered no

information or cues about latent social groups—e.g., political affiliation, ethnicity, etc. These findings provide evidence that when no such group-based cues are given, the effect of similarity is severely attenuated.

Perceived similarity and prosociality in younger vs older adults.

One potentially fruitful avenue of future inquiry would be to study the differential effects of perceived similarity in younger versus older adults. As shown in Study 3.1, correlational effects of perceived similarity on prosocial decision-making were observed in younger, but not older, adults. One possible explanation is that younger participants pay more attention to, or place a higher value on, information about similarity than do older adults; they are therefore more likely to take it into consideration when making prosocial decisions. Future studies could examine whether older adults fail to encode information about similarity, or successfully encode the information but do not factor it in to decisions about helping others. A second explanation would be that younger adults are less certain (or more flexible) when it comes to social groupings, while older adults are more certain and less flexible about existing social groupings. Put another way, perhaps younger adults are more flexible how they map information about similarity onto group memberships. However, it is more likely that older adults are mostly fixed in whether they place a new individual in one social category or another; therefore information about similarity independent of social groupings does not provide additional decision-making utility. The implication in this case would be that younger adults might show a willingness to place new individuals in a social groups on a wider range of factors (including interpersonal similarity) than older adults. This claim is supported by Cutler and colleagues (2021), who show

that older adults are more prosocial overall but less prosocial to out-groups than in-groups—consistent with the claim that older adults make decisions more strongly based on identity.

A social correlate of Hamilton's Rule.

Explanations of prosocial behaviour based on similarity can be traced back to a classic finding from evolution biology, called Hamilton's Rule (also known as the theory of kin selection; Hamilton, 1964; Dawkins, 1976). Hamilton's Rule is a highly consistent empirically-based finding about the circumstances under which ostensibly altruistic behaviour is likely to occur. On a conceptual level, it says that an individual actor is likely to help a conspecific target when the benefit to the target is greater than the cost to the actor. This balance is weighted by the genetic relatedness of the actor and the target. If they are highly related, then the balance need only tip slightly in favour of the target's benefit. If they are not highly related, the benefit to the actor would need to be large. Relatedness is defined as genetic overlap that is unique with respect to the greater population (e.g., a shared genetic code for having two eyes or four limbs is not a sign of relatedness in an average population of humans; relatedness is a relative concept). More formally it can be written as:

$$r * B > C$$

where *B* is the benefit to the target, *C* is the cost to the actor, and *r* is the coefficient of relatedness (defined between 0-1). That is, altruistic behaviour is predicted to occur when the relatedness-weighted benefit of an action is greater than the cost.

A similar rule could be hypothesised for social psychology: instead of defining relatedness genetically, it could be defined in terms of perceived social similarity. For instance, Batson et al. (2005) note exactly this point in connection with Hamilton's Rule: "if those with

whom we share more genes are more similar to us in appearance and attributes, then it is not necessary to have a highly sensitive detector of degree of genetic relatedness for kin selection to operate."

Though Hamilton's Rule is typically applied only in ethology, some research has sought to identify Hamilton's Rule using a network analysis of who people tend to turn to for support in times of need (Rushton, 1989). Much of this research is based on an investigation of which observable cues can be a proxy for genetic relatedness (e.g., similar face shape or skin colour). For example, Park and Schaller (2005) used a version of the implicit association task to show that people were more likely to associate others with kinship terms when they shared similar attitudes. More recently, Burton-Chellew and Dunbar (2015) found that people were more likely to turn to relatives than to non-relatives in a time of need, suggesting that they expected help from other individuals according to their degree of genetic relatedness. Globally, people also tend to give more to intra-national than inter-national charities for the same kind of causes (Cutler et al., 2021). But no study, to my knowledge, has ever investigated a social correlate of Hamilton's Rule—investigating willingness to exert prosocial effort by directly manipulated the degree of perceived social similarity experimentally. Moreover, the Prosocial Effort Task (Lockwood et al., 2017) is uniquely well-suited to probe a social correlate of Hamilton's Rule, as it provides explicitly defined B's and C's: the target receives a specific number of points based on the participants willingness to exert effort at various levels of difficulty. The more a participant perceives herself as socially related to the target, the more likely she would be to exert effort to earn the target greater reward.

The work presented in this chapter provides a tangential test of a possible social correlate of Hamilton's rule. As discussed, there is a general trend toward higher prosocial motivation

when there is a great difference in similarity. The intent of the paradigm used here was to study the effects of perceived similarity in the absence of group-identifying information, as well as to compare these effects with the same paradigm using a group-based manipulation. To test Hamiltonian hypotheses more directly, the experimental paradigm would need to explore finer grained distinctions in perceived similarity, as well as different kinds of similarity manipulations (other than personality-based ones). An additional concern is that the effects of similarity were not especially strong; while they are generally in line with Hamiltonian predictions, it is not obvious from these data that interpersonal, non-group-identifying similarity is a primary driver of human prosocial decision-making. Therefore, while a version of this paradigm could provide a more directly Hamiltonian test, it will remain the purview of future research to explore the hypotheses generated by this model in greater depth.

In summary.

The present studies show that how similar a participant is made to feel to another person influences their willingness to exert effort to obtain rewards for that individual. This claim is supported both by correlational evidence showing perceived similarity predicts other prosocial decisions (Study 2.2 & 3.1), as well as in a study that manipulates perceived similarity directly (Study 3.3). Crucially, perceived similarity was manipulated in a way that did not introduce any confounding cues of about social group membership. While significant effects were found, they were not nearly as large as the effects found in the intergroup studies of Chapter 2. This suggests that while perceived similarity does play a role in prosocial decision-making, it is a much subtler one than the role played by group membership.

CHAPTER 4:

ANTISOCIAL MOTIVATION AS A FUNCTION OF GROUP MEMBERSHIP

Abstract. *In this chapter, I investigate whether people are willing to incur an effort cost to harm another agent, depending on whether that agent is an in- or out-group member. In previous chapters, my experiments have investigated prosocial motivation (whether people are willing to incur an effort cost for the benefit others) and this chapter looks at its inverse: antisocial motivation. I begin by outlining two possible patterns of behaviour, which are explored in existing literature, and lead to two competing hypotheses that the present studies are intended to address. The first hypothesis is that participants might consistently choose not to harm an out-group member due to effort aversion—why actively go out of one's way to cause harm? The second hypothesis is that, conversely, participants might consistently choose to harm out-group members, because intergroup bias is such a compelling behavioural force. In the present studies, participants are given the opportunity to choose between taking points (monetary reward) away from another individual by exerting a specific amount of effort versus an option with no effort cost in which no points are taken away. On different trials, the target individual is a member of the participant's in- or out-group. In Study 4.1, I present the results of this inverted reward scheme version of the Prosocial Effort Task. I find that participants almost always chose to take points away from out-group members, with minimal discrimination between high- and low-effort costs. There is significant variation in the amount of intergroup bias shown by individual participants. In Study 4.2, I replicate these findings. In the latter version, I modify the instructions to account for a possible alternative explanation of inequity aversion. In these data, participants showed a consistent lack of discrimination between high- and low-effort trials—showing that they were motivated to harm out-group members regardless of the cost to themselves. Combining these two samples, I show that a measure of subclinical psychopathy predicts one's willingness to harm both out-group and in-group members—but not the willingness to harm out-group members more than in-group members.*

Introduction

In the studies presented in Chapters 2 and 3, participants are given the opportunity to earn points in exchange for effort. The extent to which participants are willing to exert effort for the monetary benefit of themselves or others is the core question which experiments involving the Prosocial Effort Task have sought to answer. In this chapter, I will explore the inverse of this question: Under what conditions are people willing to exert effort to the *detriment* of other individuals? I call this "antisocial motivation." The previous literature on this kind of willingness to harm points to two potential hypotheses: (1) that people tend to harm others only when there is an underlying group- or individual-level benefit to doing so, and therefore would be unlikely to engage in harm that patently serves no benefit (e.g., Henrich et al., 2006); and (2) that people tend to violate this general inclination toward justifiable harm when placed in an intergroup scenario, even when that harm toward an out-group member serves no identifiable purpose (e.g., Aaldering et al., 2018). However, to my knowledge, no single study has pitted these two competing hypotheses against one another directly, particularly to investigate willingness to harm an out-group member in the absence of standard justifications for actions against out-group members such as commitment of moral violations or a competitive intergroup setting. As demonstrated in the previous two chapters and consistent with previous literature (Hull, 1943; Zipf, 1949), people are typically effort-averse; they are willing to overcome this aversion only when the benefit to doing so is sufficiently high (Lockwood et al., 2017). Will people exert effort to harm out-group members even when there is no direct benefit to doing so?

Harm as an instrumental social tool.

One dominant hypothesis in behavioural studies of willingness to harm is that harm serves a constructive social purpose when it comes in the form of punishment. Several influential accounts have argued that punishment is in fact a key ingredient in the evolutionary foundations of human cooperation (Fehr & Gächter, 2002; Henrich et al., 2006). The core of these accounts often poses costly punishment as a puzzle: why would an individual incur a cost to oneself to bring some form of harm or retribution to another? This form of punishment is "costly" because it comes at one's own expense; it is "altruistic" because the group benefits from the deterrence of free-riders (Barclay, 2006). In an experimental setting, Fehr & Gächter (2002) showed that without the ability to punish defectors—those who go against the established social order—cooperation collapses and society devolves into anarchic chaos, a Hobbesian war of all against all. Henrich and colleagues (2006) used a similar paradigm to show across fifteen diverse cultures that humans are willing to administer costly punishment in proportion to the amount of unequal behaviour observed, that the severity of this punishment differs across societies, and that the amount of costly punishment predicts the presence of altruistic behaviour. These experiments suggest that the inclination to harm others as punishment played an important and beneficial role in the foundations of human group dynamics across societies. This position is further supported by developmental evidence that human infants are sensitive to nuanced mechanisms of harm and helping, such as being more likely to help an individual they have inadvertently harmed (Hepach, Vaish, & Tomasello, 2016; Hepach, 2017).

While such costly punishment is beneficial from the perspective of an overall group, it still must be acted upon by an individual. As with willingness to engage in prosocial action (e.g., Lockwood et al., 2017), people appear to be sensitive to the cost-benefit dynamics of enacting

punishment and are willing to punish when the cost to themselves is sufficiently low in conjunction with the severity of the punishment to the other agent (Egas & Riedl, 2008). Beyond group-level benefits, punishment can also benefit the individual who enacts it: people are likely to perceive altruistic punishers as more trustworthy, group focused, and worthy of respect than non-punishers (Barclay, 2006). Another line of work has extended questions about willingness to punish by looking at the individual circumstances in which people choose to punish. Cacault and colleagues (2015) designed a paradigm which gave participants the option to choose between a purely prosocial, a purely antisocial, and a mixed option. They found that some people chose to engage in "strong aggression"—choices that were detrimental to both victims and perpetrators. While this suggests there is large variance in individual choices to punish, there are also consistencies. For instance, people are willing to administer punishments which have no deterrent effect (i.e., the perpetrator does not know they are being punished) but only a retributive effect (Crockett, Özdemir, & Fehr, 2014). Even though punishment typically both inflicts damage (the retributive effect) and communicates a norm violation (the deterrent effect), the study by Crockett and colleagues found that retribution alone was sufficient for people to choose costly punishment. Together, these studies suggest that people are sensitive to many different kinds of benefits (e.g., retributive vs deterrent; reputational gain) associated with enacting punishment.

What about the associated costs of enacting punishment? Crockett and colleagues (2015) provided evidence that people value harm to others more than harm to themselves—that is, they perceived there to be a higher cost associated with harm that comes to another person versus harm that comes to oneself. In this study, participants could trade off their own profits for the pain experienced by themselves or others. Participants were consistently willing to pay more to

reduce the number of electric shocks experienced by others than experienced by themselves. People also required more compensation to increase another's sufferings relative to their own. Crockett and colleagues also found that aversion to pain for both self and others correlated negatively with a measure of subclinical psychopathy. Indeed, one hallmark of psychopathy is diminished sensitivity to the affective costs usually associated with enacting antisocial harm toward others (Anderson & Kiehl, 2012), and in an analysis of the data presented across both studies in this chapter will provide evidence of a correlation between antisocial motivation, as measured by my behavioural task, and a measure of subclinical psychopathy.

Another prominent line of research has focused on harm in the context of decisions made on the basis of Utilitarian moral principles (Kahane et al., 2018). Prior literature on Utilitarian judgments tended to centre on the principled avoidance of harm (e.g., choosing an emotionally unintuitive option to save a greater number of people in a trolley problem). This line of research provides evidence of a dissociation between the "negative" (one's attitude toward instrumental harm) and "positive" (one's inclination toward impartial beneficence) aspects of prosocial decisions from a Utilitarian perspective. One use of this dissociation has been to tease apart people's judgments who rely on a consequentialist vs non-consequentialist decision strategy (Everett et al., 2018): under what conditions do people judge an agent as acting more morally? Everett and colleagues show that while non-consequentialist agents are preferred overall when making decisions about instrumental harm, there is no distinction in preference between the strategies when socially distal agents (e.g., a politician vs a family member) make decisions about impartial beneficence. Other evidence suggests that priming intuitive judgment decreases Utilitarian judgments about instrumental harm but not impartial beneficence (Capraro, Everett, & Earp, 2019). In short: people treat the costs of harm differently than the benefits of help, and also

make assessments of others morality based on how an actor takes these various costs and benefits into account.

Together, the above lines of research suggests that willingness to harm another agent is typically dependent on whether that harm can be justified—in a way that either benefits a group or an individual. Just as with most behaviour, people are sensitive to the relative costs and benefits of engaging in it, and therefore tend only to engage in harm when there is a benefit to offset the associated cost. Deviation from this pattern of behaviour is usually synonymous with some form of maladaptive syndrome, such as psychopathy (Drayton, Santos, & Baskin-Sommers, 2018). However, another potential violation of this standard pattern exists in intergroup contexts: How does the standard cost-benefit of harm shift when that harm is directed at an out-group member?

Increased willingness to harm in an intergroup context.

One running theme of the literature on costly punishment is that people tend to use group membership as a heuristic—that without doing the full cost-benefit analysis, they assume that what is good for one's group is good for oneself and that what is bad for the opposing group is good for one's own group. An archetypal example of this is Baron (2001), who showed that people conflate their own self-interest with the interest of their group: if you require people to calculate actual gains and losses, this parochial cooperation is reduced. A recent study presented participants with a version of an economic game in which they could display universal cooperation (benefitting both in-group and out-group members) or parochial altruism (benefitting in- but not out-group members; Aaldering et al., 2018). Aaldering and colleagues found that people consistently chose parochial altruism over universal cooperation, consistent with Baron's

account that losses for one's out-group entail a gain for their in-group. Similarly, Apps and colleagues (2018) demonstrated that participants were often willing to reject an even fifty-fifty split in an economic game when playing with an out-group member. The evidence from these studies suggests that group membership can often override standard considerations in a cost-benefit analysis, such as whether a monetary reward is associated with a positive or negative valence, and even moral considerations such as fairness.

Another important line of research by Cikara and colleagues shows the effects of intergroup "schadenfreude"—pleasure derived from the pain of out-group members. In one experiment, Cikara, Bruneau, and Saxe (2011) showed avid baseball fans depictions of failures from their team or a rival team while undergoing magnetic resonance imaging. Whether failures (or successes) belonged to one's in- or out-group changed participants' hemodynamic response to the stimuli. Successes of the in-group and failures of the out-group both elicited activation in the ventral striatum (associated with reward), while failures of the in-group and successes of the out-group elicited activation in the anterior cingulate cortex and insula. Similar neural responses can also be modulated with stereotypes (Cikara & Fiske, 2011), in which participants showed different responses to positive or negative outcomes for an out-group member depending on whether that out-group was stereotyped as enviable or pitiable. Participants reported more favourable responses to the negative outcomes of enviable out-groups and showed greater willingness to harm enviable out-groups than pitiable ones. This pattern of results was extended to show that participants were least averse to negative outcomes for high-status targets, were more likely to endorse these negative outcomes, and showed a physiological response of facial muscles consistent with smiling, while manipulating the status information significantly changed participants' responses (Cikara & Fiske, 2012; Cikara & Fiske, 2013). This schadenfreudic

response can be modulated by self-reference, with spontaneous response in the medial prefrontal cortex (an area associated with self-reference) predicting a decrease in willingness to harm out-groups (Cikara et al., 2014). It can also be modulated by empathy: a greater empathic response for in-groups elicits more out-group schadenfreude, while a greater empathic response for out-groups elicits less (Bruneau, Cikara, & Saxe, 2017).

Together, these studies show that people are perfectly willing to harm out-group members—mostly as a heuristic for the benefit of their in-group, which in turn is a proxy for their own personal gain. However, this willingness depends on many contextual factors: the exact in-group/out-group payoff matrix, the amount of empathy currently felt toward in- or out-group members, the stereotypical content associated with the out-group, whether the punishment promotes adherence to social norms and supports cooperation, and the amount of pure enjoyment one gets from inflicting harm on others. But what has not been addressed in this literature is the extent people are willing to harm out-groups in the absence of this kind of information. Without potential benefit to oneself or stereotypic content of the out-group or any empathy-inspiring personal information about the target individual, do people still show an increased willingness to harm members of an out-group?

Furthermore, in none of the experiments discussed above do participants have to put in a higher effort cost to inflict harm on out-groups: harm is either passively observed, or comes as a minimally effortful allocation of money or shocks. Perhaps people are willing to harm others—but only when doing so does not require the executing of a physical action (e.g., the "footbridge" dilemma in Thomson, 1985) The experiments presented in this chapter offer participants an opportunity to exert effort to harm another individual, while offering no compensating benefit in return.

Hypotheses and summary of studies.

In the experiments that follow, I present the results from two studies which use a minimal groups paradigm (Tajfel et al., 1971) to present participants with a choice between working to harm another individual (from the in- or out-group) or doing no work with no harm to the other individual. This inverse reward scheme of the Prosocial Effort Task (as presented in Chapters 2 and 3 with the standard reward scheme) allows me to test whether people are willing to harm others—who either in-group or out-group members—when that harm is associated with an effort cost, and there is no direct incentive (moral, monetary, or otherwise) to desire negative outcomes for members of either group. I will also provide correlational evidence showing increases in this kind of antisocially motivated behaviour are associated with a self-report measure of subclinical psychopathy.

Study 4.1: Inverse-Rewards Prosocial Effort Task

Participants. Twenty-seven participants were recruited on the Prolific subject recruitment platform (9 female; mean age = 28.2). All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately 30 minutes to complete.

Method.

Design. This implementation of the Prosocial Effort Task was the same as the one described in the primary intergroup Prosocial Effort Task (Study 2.2) with two main differences:

instead of exerting effort to *earn* points for in- or out-group agents, participants worked to take points *away* from them (Figure 1; the self condition was also removed, leaving only in- and out-group conditions). In each trial, participants were presented with a choice between high- and low-effort options. In the high-effort option, participants had the opportunity to exert a given level of effort in order to take a given number of points away from a specified target agent (either an in-group or out-group member). In the low-effort option, participants were required to exert no effort, and no points were taken from the target agent. If participants failed to choose between the high- and low-effort options, the target agent lost a single point (as opposed to zero in the low-effort option, and at least two in the high-effort option). Participants performed three blocks of fifty trials each (consisting of one trial for each permutation of the five effort levels and five reward levels for each of the two agents). The task was designed to mimic the dynamics of

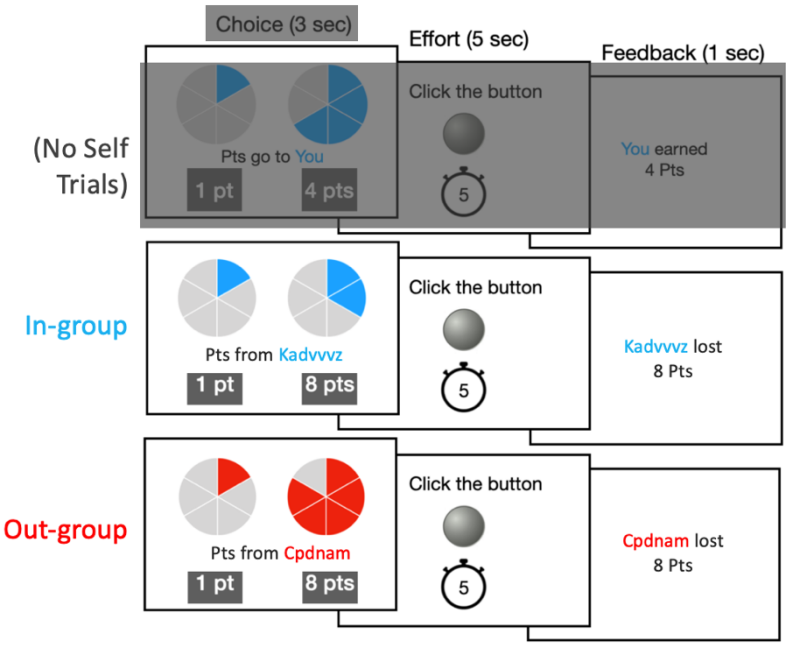


Figure 1 — Inverse Rewards Scheme for The Intergroup Prosocial Effort Task. The task structure was based closely on the intergroup and similarity-manipulated versions of the Prosocial Effort Task presented in previous chapters, except in each trial participants were offered an exchange of a given level of effort in order to take points *away* from a target agent. Participants were presented with the opportunity to take points away from in- and out-group targets, but not themselves.

previous versions of the Prosocial Effort Task as closely as possible, while inverting the logic of the rewards scheme.

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Prolific subject recruitment platform.

Procedure. Participants followed an identical procedure to the one used in Study 2.2, with the exception of the inverted rewards scheme and the absence of self trials. In each trial, participants chose whether to exert effort to take points away from in-group or out-group target agents. This entailed several modifications. The self condition was removed from the inverted rewards scheme version, because it seemed unlikely this would elicit any variance in participants' responses. This left only trials in which they chose whether to work against other agents: members of their in- or out-group. At the onset of each trial, participants were prompted with "Points come from [Target Agent]" where the in- and out-group target names were the same as in the original version. The variation between the reward levels remained the same as in the original version, with the same number of points being *subtracted* from the other participants as could be earned in the original. The effort levels remained identical to the previous version, requiring different numbers of button presses.

Inverted reward scheme. There was also a modification to the paradigm's reward structure to maintain overall coherence. As described in the previous paragraph, when the participant chose the high-effort condition (to trade button-clicks for points), the stated number of points were deducted from the relevant agent. When the participant chose the low-effort option, no points were deducted from the target agent. This gave participants the opportunity to employ a strategy in which the other target never lost any points. If the participant chose the high-effort option but failed to complete the button pressing as prescribed in the five-second

window, one point was deducted from the target agent. One point was deducted (as opposed to no points) in order to disincentivise participants from choosing indiscriminately between the high- and low-effort options, since otherwise both cases would require a single button click (the initial choice) and the same outcome (0 points lost).

The three-second time limit was also removed for the initial decision between the high- and low-effort options. Keeping this limit in place would mean that a participant who wanted not to take any points away from the target agents could simply leave the experiment running, and it would eventually conclude without requiring their continued attention.

Instructions. Before the main experiment, participants were given the following instructions: "In this task you'll have the opportunity to take points away from other Prolific workers by clicking a button on screen. Each of these Prolific workers starts off with a £1 bonus. Each point you take away from them will decrease that bonus. For example, if you take away all their points, they will be left with a bonus of £0. How much you want to take away is up to you." Otherwise, the instructions were identical to the original version of the experiment, with the exception of the explanations of the reward.

Questionnaires. Before beginning the main experiment, participants were given the same questionnaires as in the original task, related to empathy, apathy, and alexithymia. Additionally, participants were given a standard questionnaire of psychopathy (Levenson, Kiehl, & Fitzpatrick, 1995).

Post-task self-report judgments. After completing the main experiment, each participant was ask to report how "similar" they perceive themselves to be to each of the two others players for whom they were ostensibly earning points ("How similar do you feel to [name]?" rated on a slider scale from "completely dissimilar" to "completely similar"). Participants were also asked

about their beliefs of whether an agent would earn points for them ("Given the chance, how many points do you think [name] would have earned for you?" on a slider scale "as few as possible" to "as many as possible"). Participants were also asked to rate on the same scale: "In general, how many points do you think people earn for others in this task?" These were the same post-hoc similarity judgments included in the Intergroup Prosocial Effort Task from Chapter 2.

Results.

Participants almost always chose to take points away from out-group members.

Figure 2 shows participants' mean probabilities for choosing the high-effort option for in-group vs out-group agents. In line with the analyses presented in Chapters 2 and 3, I ran a linear mixed-effects regression to determine participants' difference in choices for different group members. The regression was run with choice probability as an outcome variable and agent as a predictor variable, including a subject-level random intercept. There was a significant difference between in- and out-group choice probabilities ($b = 0.536$ [0.386 – 0.687], $z = 6.985$, $p < 0.001$).

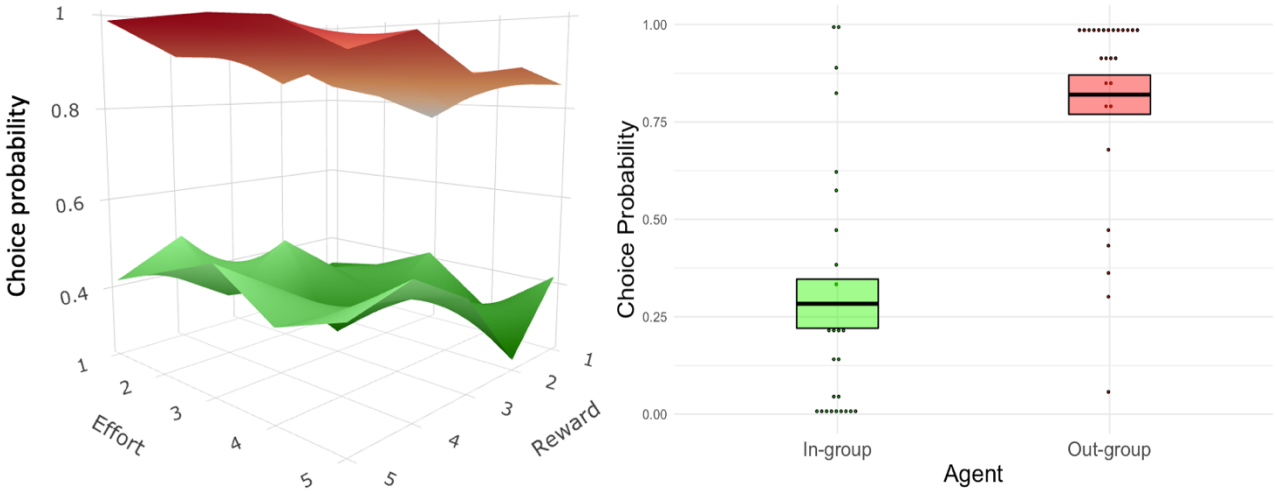


Figure 2 — Participants almost always chose to take points away from out-group members. (**A**) Choice probability as a function of effort and reward, with green as in-group agents and red as out-group agents. (**B**) Participants chose on more than 80% of trials to take points from out-group members.

Strikingly, participants on average chose to exert effort to take points away from out-group agents on 82.0% of trials, with many participants choosing to take points away from out-group members on every possible trial.

Participants chose to take points regardless of effort costs.

I ran a general linear mixed-effects model to look for possible interactions (Figure 3). The model revealed main effects of agent ($X^2 = 982.89$, $p < 0.001$) and reward ($X^2 = 70.72$, $p < 0.001$), but not of effort ($X^2 = 1.27$, $p = 0.261$). This result is fairly surprising, as it is at odds with the usual findings in this paradigm, in which participants show strong effects of effort aversion as the difficulty levels increase. This implies that participants exhibited a similar pattern of decisions on higher- and lower-effort trials. There was also a significant interaction between

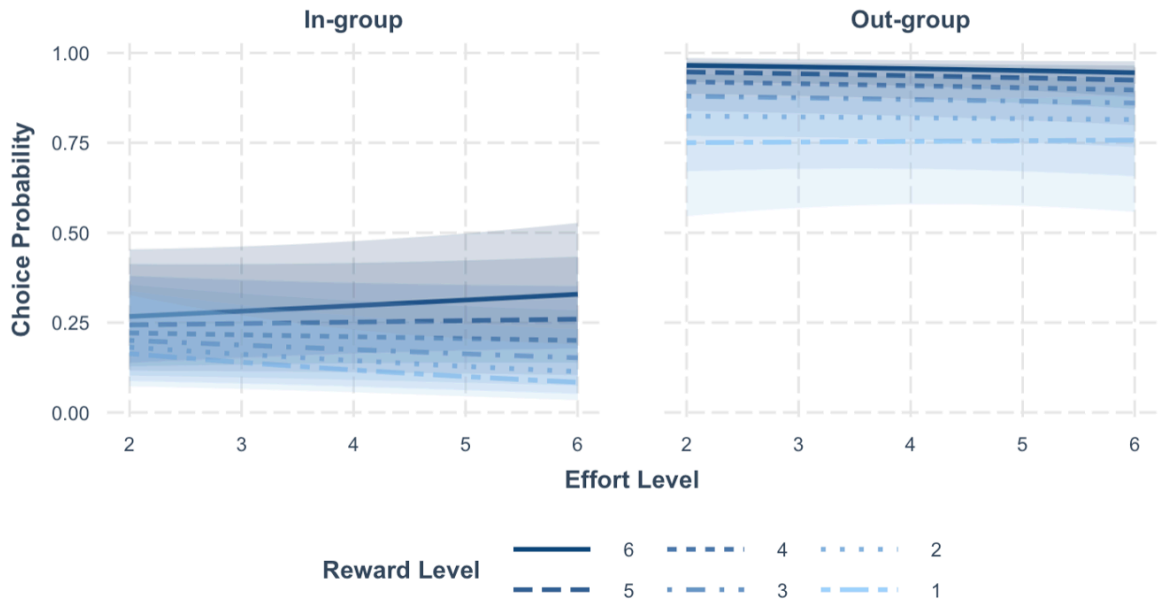


Figure 3 — Generalized linear model. This GLM analyzed trial-by-trial behaviour, predicting choices on individual trials for the two target agents with effort and reward level as fixed effects and participant as a random effect. The slopes of the lines are relatively flat, suggesting that participants tended not to make decisions on the basis of effort.

agent and reward ($X^2 = 6.11$, $p = 0.024$) but not between agent and effort ($X^2 = 0.15$, $p = 0.697$) or effort and reward ($X^2 = 0.53$, $p = 0.468$).

Participants varied in the strength of their group bias.

Figure 4 shows variation in participants' group-based strategies. The majority of participants showed strong group effects. Approximately two-thirds of participants showed a steep slope in the difference between out-group and in-group choice probabilities (Figure 4A). A steeper slope indicates greater intergroup effect, with greater discrepancies associated with decisions made for in- and out-group members. Approximately one-third of participants showed a slight intergroup effect, while three showed a negligible group effect and only one showed the opposite effect.

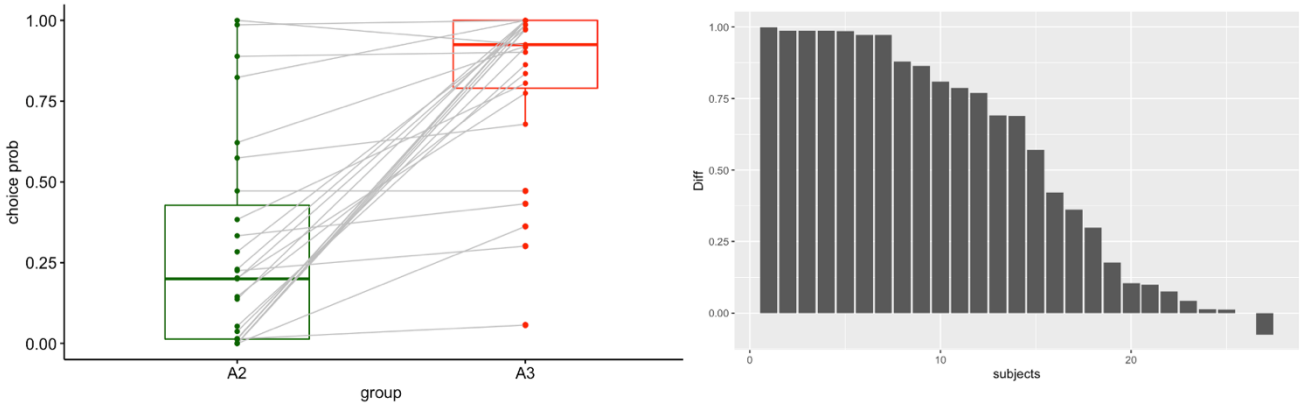


Figure 4 — Individual decision strategies. (**A**) This plot shows the difference between individual participants' mean choice probabilities for in- and out-group agents. Approximately 1/3 of participants showed little or no group bias (flat slope) while 2/3 showed a rather severe group bias (steep slope). (**B**) This plot shows the magnitude of participants' intergroup differences in choice probabilities.

Discussion.

This experiment was designed to create a direct comparison of the strength of two countervailing behavioural forces: that people exert effort only when there is a justifiable or instrumental reason to do so (especially in order to cause harm to another person), and that people are willing to invest additional effort when it comes to working toward the detriment of an out-group member. The results from these data suggest that the latter force won out decisively in the present study. Participants consistently worked to take points away from out-group members, despite no benefit to themselves, their in-group, or in response to any kind of moral infraction or norm violation. This is particularly striking when juxtaposed to the standard effort aversion effects found in this task (both in the online button-clicking operationalisation of effort employed the experiments for this thesis, as well as in the grip-force operationalisation of Lockwood et al., 2017). Whereas participants on average chose the high-effort option to *earn* points for out-group target agents approximately half the time in the Intergroup Prosocial Effort Task (Study 2.2), participants in the present study chose more than eighty percent of the time to exert an equivalent amount of effort to take points *away* from comparable out-group target agents.

One relevant consideration is the well-known effect of framing, in particular how one problem with the same logical structure can yield two different behavioural responses when presented in two different ways (Tversky & Kahneman, 1986). A possible relevant reframing here is that in the initial version of the task (Study 2.2), effort is dedicated toward earning points for others; therefore participants infer that earning points for others must be valuable or desirable. By contrast, in the present version, effort is dedicated toward taking points away from others; therefore participants infer that there must be at least one relatively high-value decision

pathway and the only plausible outlet is taking points away from out-group members. This is consistent with the idea of the effort paradox, whereby people infer the value of something based on how much effort they put in to obtain it (Inzlicht, Shenhav, & Olivola, 2018; Campbell, Chung, & Inzlicht, 2022). Another possible reframing is that what is perceived as "work" when pressing a button to earn points for oneself or another agent is perceived as a kind of "fun" when it comes to pressing a button to take points away from an out-group agent, and therefore the same action is not aversive in the same way. After all, this is the premise of many popular video games: repeatedly pressing a button to effect some detrimental plight on an out-group agent.

Similarly, recent evidence suggests that people prefer to exert a modest amount of effort over doing nothing (Wu, Ferguson, & Inzlicht, 2021). Perhaps participants would find it aversive to perform a thirty minute experiment while only ever choosing the default, low-effort option. Again, the path of least resistance in alleviating this aversion is to work against the out-group.

Yet another relevant consideration might be the commonplace intergroup dynamics associated with games. It is well-known that intergroup bias is exacerbated in competitive settings (Sherif et al., 1961), and as anyone who has ever been to an athletic event knows, positive affect responses in the face of negative outcomes for an opposing team are a staple of intergroup competition (Cikara & Fiske, 2013). Perhaps people are willing to adopt the conventional dynamics of intergroup competition—i.e., the other team's loses take on a positive valence—even in the absence of the standard incentive structure of intergroup competition. Although possible, this does seem to contradict the exquisite sensitivity to cost-benefit trade-offs typically demonstrated by people when doling out punishment.

Yet while any of these explanations might potentially be relevant, none of them fully explains why participants chose to work so consistently against out-group target agents (as

opposed to forgoing to the highest effort levels, for instance). One potential avenue for future research to explore might be in the underlying motivations between individual differences in intergroup bias.

Study 4.2: Inverse-Rewards Prosocial Effort Task (Replication & Extension)

One possible explanation for the effects in Study 4.1 is inequity aversion, the well-documented inclination of people to prefer equitable outcomes (e.g., Fehr & Schmidt, 1999; Fehr & Schmidt, 2010; Shaw, Choshen-Hillel, & Caruso, 2016). Perhaps participants believed that these other participants—who were in fact fictitious—were receiving a "bonus" on top of what they had already earned, leading some of the real participants perceiving the in-group and out-group participants might earn more without exerting any additional effort. For example, this would be the case if (as was suggested in the instructions) the other agents went through a similar paradigm, earning the same amount of money that the participant does in the current one. Then any money which the participant does not take away from them would be an additional, unearned add-on over what the participant would receive themselves. Perhaps participants were choosing to take money away from others because they perceived this to be an inequitable outcome. In Study 4.2, I sought to replicate the findings in Study 4.1, while presenting the participants with instructions that would allay this potential confound.

Preregistration. I preregistered methods, recruitment, analyses, and hypotheses for this experiment at https://osf.io/632dh. The hypotheses in this preregistration were based on replicating the findings from Study 4.1.

Participants. Twenty-seven participants were recruited on the Prolific subject recruitment platform (15 female; mean age = 28.7). All participants provided written informed consent. The

study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately 30 minutes to complete.

Method.

Procedure. The procedure and methods were identical to Study 4.1, with the exception of a slight modification to the instructions, described below, to address a possible confound of inequity aversion.

Instructions. Before the main experiment, participants were given the following instructions: "Each of the others Prolific workers completed a similar version of this experiment earlier today. They did a similar button-pressing task with the same length and difficulty. For their participation, they earned the same baseline amount as you. And based on their button-clicking rate, they qualify for a £1 bonus. Based on your own button-clicking rate (which was measured at the beginning), you also qualify for a £1 bonus. In this task you will have the opportunity to take away from their bonus. For example, if you take away all their points, they will be left with a bonus of £0. How much you want to take away is up to you. Both you and the other participants will be paid the final amount within 48 hours." All other aspects of the instructions were identical to Study 4.1.

Results.

I ran the same suite of analyses as presented in Study 4.1, described above, to determine whether the same pattern of effects successfully obtained in a new sample, presented with a

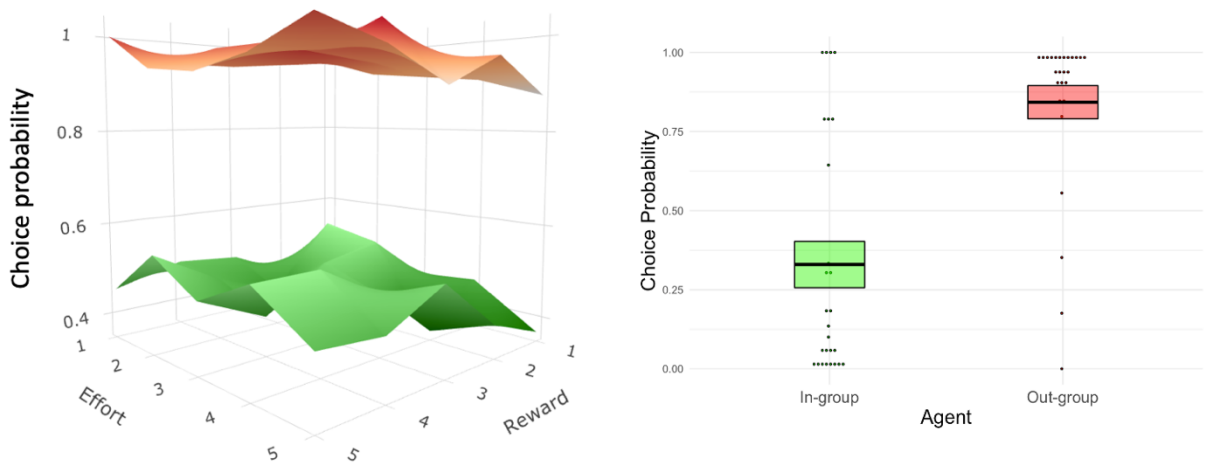


Figure 5 — Replication: participants almost always chose to take points away from out-group members. (**A**) Choice probability as a function of effort and reward, with green as in-group agents and red as out-group agents. (**B**) Participants, once again, chose on more than 80% of trials to take points from out-group members.

slight modification to the instructions. In the full dataset combining participants from both studies (N = 54) I also sought to find correlations between the self-report assessment of psychopathy and the participants' task behaviour.

The findings from Study 4.1 were successfully replicated.

To replicate the analysis from Study, I ran a linear mixed-effects regression to determine participants' difference in choices for different group members (Figure 5). The regression was run with choice probability as an outcome variable and agent as a predictor variable, including a subject-level random intercept. As in Study 1, there was a significant difference between in- and out-group choice probabilities of a comparable magnitude ($b = 0.521$ [0.369 – 0.672], $z = 6.727$, $p < 0.001$). The effect of choosing to take points away from out-group members was not diminished.

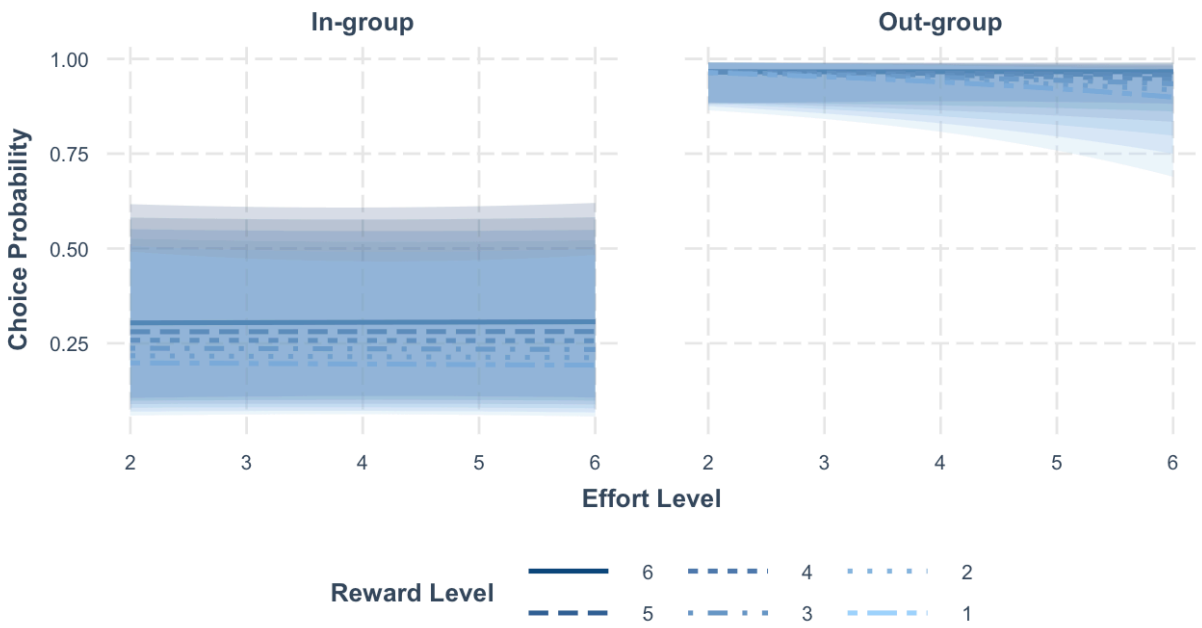


Figure 6 — Generalized linear model. I used a GLM predicting choices on individual trials with effort and reward level as fixed effects and participant as a random effect.

I ran a general linear mixed-effects model to replicate the null finding on the effect of effort (Figure 6). In line with Study 4.1, the model showed main effects of agent ($X^2 = 890.16$, $p < 0.001$) and reward ($X^2 = 9.80$, $p = 0.002$), but not of effort ($X^2 = 1.49$, $p = 0.222$). However there were no significant interactions between agent and reward ($X^2 = 0.01$, $p = 0.925$), agent and effort ($X^2 = 1.86$, $p = 0.172$), or effort and reward ($X^2 = 0.76$, $p = 0.385$).

Psychopathy predicted choice probabilities—but not intergroup bias.

With the combined data from Studies 4.1 and 4.2, I ran correlations between the outcome measures of interest and the results of self-report questionnaires given to the participants before completing the task. Figure 7 shows the correlation between participants' scores for Primary Psychopathy and their mean choice probabilities for in- and out-group agents. Primary

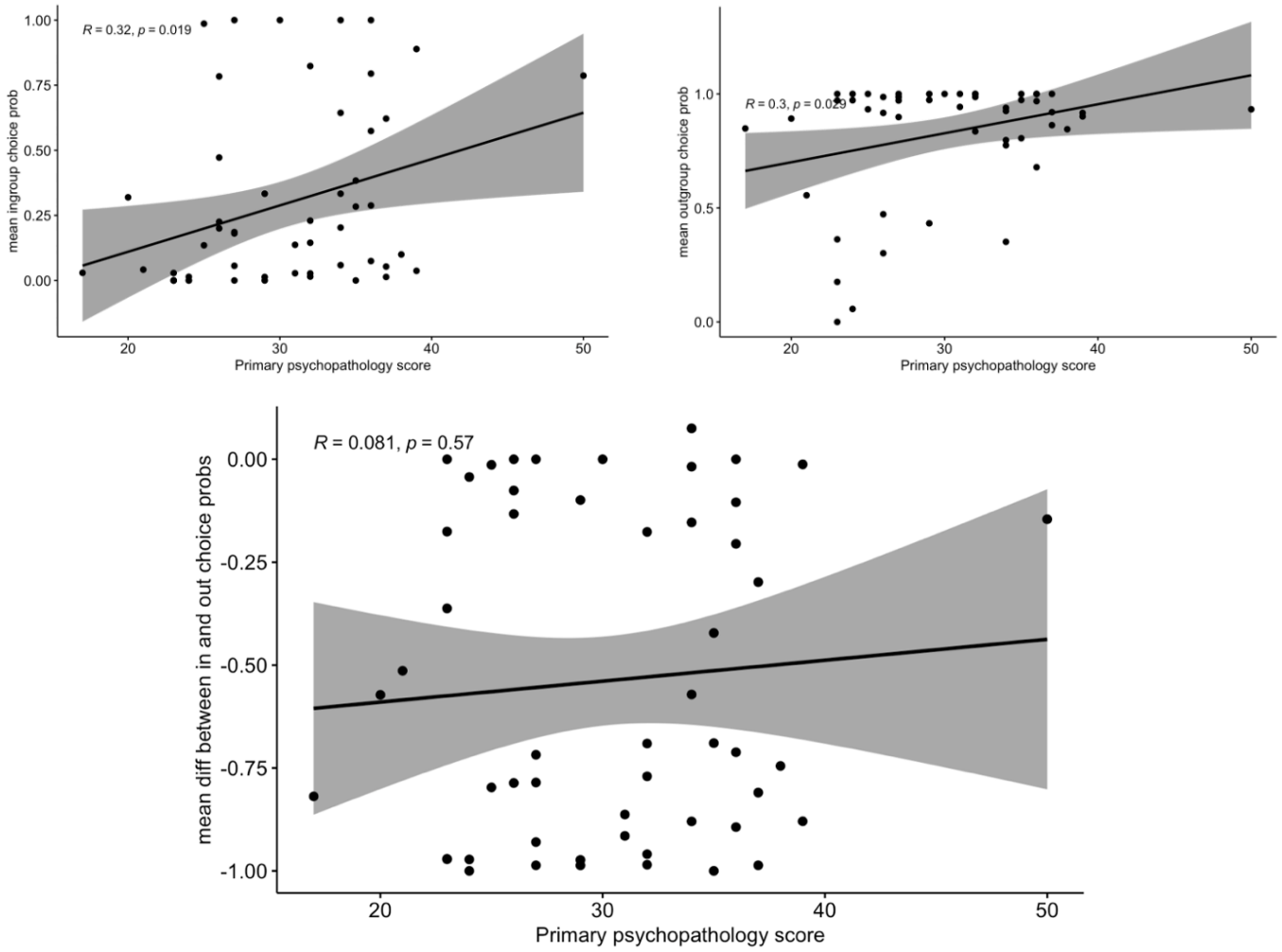


Figure 7 — Psychopathy predicted participants' mean choice probabilities for both in- and out-group members, but not the difference between them. (**A**) Mean in-group choice probability as a function of primary psychopathy. (**B**) Mean out-group choice probability as a function of primary psychopathy. (**C**) There was no correlation between primary psychopathy scores and intergroup bias (the difference between choice probabilities for out-grout and in-group agents).

psychopathy was predicted by both in-group choice probability ($r = 0.32$, $p = 0.019$; Figure 7A) and out-group choice probability ($r = 0.30$, $p = 0.029$; Figure 7B). However, psychopathy did not predict greater intergroup bias: there was no significant correlation between primary psychopathy and the difference in choice probabilities between out-group and in-group targets ($r = 0.08$, $p = 0.57$; Figure 7C). This suggests that subclinical psychopathy predicted a participant's

overall willingness to take points away from other individuals, but not on the basis of group affiliation.

Discussion.

The data from Study 4.2 successfully replicated the findings from Study 4.1—if anything, with a stronger effect of out-group and with an even smaller effect of effort. Based on the correlational data, I found that the higher a participant scored in primary psychopathy, the more likely they were to be willing to takes points away from others. However, psychopathy did not predict intergroup bias: participants scoring higher on primary psychopathy were just as likely to take points from out-group vs in-group targets as other participants.

General Discussion

The studies in this chapter provide evidence that participants were willing to exert effort to harm an out-group member, despite no apparent individual- or group-level benefit to offset the effort cost. While this is consistent with the effect of intergroup schadenfreude (Cikara, Bruneau, and Saxe, 2011; Cikara & Fiske, 2013), none of the paradigms employed by Cikara and colleagues were specifically designed to pit a participant's willingness to harm an out-group member against an associated effort cost. The extant literature on this kind of out-group harm to date has only shown that people are willing to engage in it or that it results in a positively valanced affective state, but it has not quantified the costs that people are willing to incur to engage in this behaviour. The studies presented in this chapter provide a first step toward outlining what kinds of costs people are willing to incur in order to bring harm to out-group members.

However, it is more difficult to reconcile these results with studies surveyed in this chapter's introductory section on harm as an instrumental social tool. The upshot of that literature is that people are highly attuned to the reward structure associated with punishment: people tend to harm only when there's something to be gained (e.g., Crockett, Özdemir, & Fehr, 2015). As discussed, this gain does not have to be specifically an individual gain but could contribute toward a larger evolutionary or social benefit. In the present studies, there is no explicit benefit for the participant for engaging in out-group harm; is there an implicit benefit for some broad social goal baked into the logic of the task? I considered possible explanations in the Discussion of the Study 5.1. However, further research is required to disentangle these possibilities.

Another possible explanation might come from the distinction between instrumental harm and impartial beneficence (Kahane et al., 2018). Evidence suggests that priming intuitive judgment decreases Utilitarian judgments about instrumental harm but not impartial beneficence (Capraro, Everett, & Earp, 2019). If participants in the present task are relying on intuition—say, because the task requires them to do many trials, and subsequently may rely more on intuitive reactions to the cost-benefit trade-offs of any given trial—then perhaps their actions are less Utilitarian when it comes to causing harm. In the prosocial version of the task (e.g., Study 2.2), this intuitive basis for action does not affect their inclination toward impartial beneficence. Thus, participants show dramatically anti-Utilitarian behaviour in the harming version of the task, and more Utilitarian-consistent behaviour in the helping version of the task.

As for the existing research on psychopathy and intergroup bias, the present study appears to be at odds with existing evidence—though few studies have looked at the effects of psychopathy on intergroup bias specifically. One study found in an economic game, those who

scored highly for secondary psychopathic traits showed an elevated intergroup bias, making more generous offers toward in- than out-group members (Gillespie et al., 2013). In a similar vein, another study found that psychopathy predicted negative attitudes toward asylum seekers (Anderson & Cheers, 2018). Meanwhile, Crockett and colleagues (2015) found that aversion to pain for both self and others correlated negatively with psychopathy (though without directly testing an effect of group membership). Perhaps the explanation lies in the link between effort and harm: in the present experiment, enacting harm requires both an endorsement of the negative outcome and a willingness to execute the effort necessary to attain it. However, most studies meant to elicit out-group harm (such as the studies mentioned above, employing an economic game and self-report measures) require only endorsement of a negative outcome but have no associated effort cost. Further research is required to understand a potential relation between psychopathy and any potential costs of perpetrating intergroup harm.

In summary.

I have shown that people are willing to incur an effort cost to cause harm to out-group members, seemingly regardless of the effort level required to take a reward from the other agent. This stands in contrast with the standard effort aversion effects demonstrated in the previous chapters. This willingness to cause harm was replicated and shown to be robust about a possible confound of inequity aversion. It also appears that this antisocial motivation was predicted by psychopathic tendencies—though psychopathy did not affect the severity of the intergroup bias.

CHAPTER 5:

A NOVEL ASSESSMENT OF EFFORT AND ACCURACY IN THEORY OF MIND

Abstract. *While motivation has been shown to play an important role in many social-cognitive processes (Contreras-Huerta et al., 2020), previous work has not yet quantified the role of effort costs in theory of mind. In this chapter, I present a novel set of stimuli designed to assess variation in effort and accuracy in mental state inferences performed by neurotypical adults. While social-cognitive capacities such as theory of mind are of great interest in psychological research, there is a surprising lack of paradigms for assessing the extent to which individuals are more or less capable of inferring the mental states of other agents. Existing theory of mind paradigms tend to suffer from at least one of two problems: lack of ground-truth (the mental states in question are defined by the researchers and not necessarily causally related to an agent's behaviour), or ceiling effects (most neurotypical adults are able to attain 100% accuracy). The stimuli set I present is designed to overcome these limitations. The stimuli set consists of 104 five-second video clips of individuals playing the "Assassin Game," a novel game designed specifically in service of creating these stimuli. Across two studies, I demonstrate that these clips show a wide range of variation, both in terms of the average probability of participants identifying the correct assassin and target, as well as their perceived difficulty in doing so. In the first study, I show evidence for this basic effect. In the second study, I show the effect is reliable within individuals across a span of several months. I argue that this novel paradigm overcomes the two limitations of current theory of mind paradigms noted above, because (a) the mental states to be inferred are causally related to the behaviour of the (real) individuals controlling the shapes, and (b) because of the empirical variation observed in the distribution of individual scores on the task. I argue that these stimuli represent an advancement in our ability to present participants with stimuli requiring mental state inference that elicit a range of accuracies and perceived difficulties in neurotypical adults.*

Introduction

Many social-cognitive processes are known to be effortful, and engaging in them requires that an individual be sufficiently motivated to do so (Contreras-Huerta, Pisauro, & Apps, 2020). In the preceding chapters, I show how motivation to help an agent changes as a function of what information the actor knows about the target agent. To what extent does the same kind of social information influence prosocial motivation in other processes? As I will argue below, one social-cognitive process for which only indirect evidence of an effortful component has been shown is theory of mind—the ability to perform inferences about the mental states of other agents. Across the next two chapters, I will present a novel paradigm for assessing both perceived effort and accuracy while engaging in mental state inference. I will build toward using this paradigm to measure how motivation to engage in mental state inference (which has been shown to facilitate prosocial action; Gaesser, Keeler, & Young, 2018) changes depending on social information, such as group membership. To what extent are people less willing to put in effort to understand out-group minds? Are people worse at making inferences about out-group, than in-group, minds? Is any observed deficiency in understanding out-group minds due to lack of motivation or lack of capacity? In this chapter, I will focus on demonstrating variation of perceived effort and accuracy within the stimuli (Study 5.1), as well as their reliability in measuring individual differences in theory of mind ability across a span of several months (Study 5.2).

Attribution of goal-directedness as mental state inference.

It has long been known that simple geometric shapes can be animated such that humans attribute mental states to them (e.g., Heider & Simmel, 1944). Heider & Simmel's famous demonstrations showed that people are capable of attributing rich narrative content about the

mental lives of agents, even when the agent is a triangle moving across a two-dimensional plane. In the past two decades, this basic observation has been formalised in several experimental paradigms. One version, which draws inspiration directly from Heider & Simmel, comes from Abell, Happé, and Frith (2000). In this task, two triangles are shown moving in one of three conditions—sequences of random motion, goal-directed motion, or interactive motion. Abell and colleagues used this task to delineate social-cognitive differences in children with Autism Spectrum Disorder versus neurotypical children. The children with Autism tended to select "inappropriate" mentalising terms—that is, a mental state other than the one intended by the experimenters—more often, when compared to neurotypical children, in sequences requiring the observer to attribute a theory of mind to the triangles on the screen.

Other paradigms, while relaying on the same basic insight, have deviated more significantly from the original Heider and Simmel setup. Gao, Newman, and Scholl (2009) developed a paradigm in which a number of dots (e.g., 50) are moving at random across a screen. An additional dot is deployed which does not move at random, but instead follows a target dot. Gao and colleagues created different demonstrations, in which the non-random targeting dot either goes directly after the target ("heat-seeking") or introduces noise into their path of pursuit. They showed that this quantitative change in trajectory led to qualitative differences in participants' perception of the scene. For example, the non-noisy, heat-seeking agent was endorsed by participants as "chasing" behaviour whereas the noisy agent was perceived as "stalking" behaviour (Gao & Scholl, 2011).

One key unifying theme of these shape-motion theory of mind paradigms is that they require observers to assume the intentional stance (Dennett, 1987). The core claim is that these stimuli are only interpretable if the observer imputes goals (a mental state) to the observed agent.

The motion of these shapes is not explainable by the motion of non-agentive bodies—an alternative Dennett called assuming the physical stance—and therefore the most effective means of predicting future actions depends on an inference of the agents' underlying desires and beliefs assuming it is behaving approximately rationally to achieve its desires based on its beliefs (Baker, Saxe, & Tenenbaum, 2009). While goal-directedness does not cover the entire expanse of mental phenomenon, it is undoubtedly an important feature of intentional agency (Berke & Jara-Ettinger, 2022)—and therefore to the extent that motion-shape paradigms require inferences about goal-directedness, they are relevant to assessing an individual's social-cognitive processing as it relates to theory of mind.

This intentional stance framework was formalised by Baker, Saxe, and Tenenbaum (2009), who showed that people interpret the behaviour of goal-directed agents through a process of "inverse planning." This idea leverages the computational formalisations of rational planning provided by reinforcement learning (Sutton & Barto, 1998; Gershman, Zhou, & Kommers, 2017), which provides a normative definition of how a rational agent ought to behave in order to achieve a specified goal based on their beliefs about the surrounding environment. Inverse planning models, by contrast, infer the agent's unobservable goals and beliefs conditional on their observable actions (Baker et al., 2017). Baker and colleagues tested this formalisation of Dennett's intentional stance in an experimental paradigm showing a simple shape-based agent moving across an environment toward one of several possible goals. These experiments showed that people consistently attributed goals to the shape-based agent in line with the predictions made by the inverse planning modes (Baker, Saxe, & Tenenbaum, 2009; Baker et al., 2017). In this chapter, I use the precedent from this lineage of shape-motion theory of mind tasks to develop my stimuli.

Effort in theory of mind and empathy.

I provide an extensive review of the evidence for the potential cognitive effort costs associated with theory of mind and empathy in Question II of Chapter 1, the most directly applicable aspects of which I will summarise here. Many mental processes are known to be perceived as effortful (Kool & Botvinick, 2018; Westbrook & Braver, 2015; Shenhav et al., 2017). Though the exact relationship between cognitive and physical effort is difficult to disentangle, cognitive effort has been shown to elicit similar discounting effects as in physical effort tasks (Chong et al., 2018). Cognitive effort is often associated with the exertion of cognitive control (Shenhav, Botvinick, & Cohen, 2013), but can also be a property of specific cognitive tasks (e.g., Piolat, Olive, & Kellogg, 2004). Shenhav and colleagues (2017) define cognitive effort as the quantity describing what must be overcome for an individual's information processing capacity to be utilised its application in a target task, which largely depends on the extent to which the full information processing capacity is achieved. Another distinction, which I have followed throughout this thesis, is the dissociation between motivation and capacity to perform a task (Contreras-Huerta, Pisauro, & Apps, 2020). As Contreras-Huerta and colleagues have noted, social behaviours are often goal-directed and therefore require expenditure of effort. Their framework for motivated social cognition proposes that much variation in social behaviour is due not to capacity limits, but to variations in motivation to engage in a specific social task. In other words, social cognition may itself be intrinsically effortful. Given that people find other cognitive and social tasks mentally demanding, one might predict that, under certain conditions, mental state attribution might be similarly effortful.

Though theory of mind is often considered to be an automatic process (Coricelli, 2005), several accounts have proposed that it may also include an effortful component. Experimental

evidence has suggested that theory of mind may be effortful because it requires intentionality tracking (Lin, Keysar, & Epley, 2010), as well as an understanding of an individual's privately constructed perception of reality (Hodges, Denning, & Lieber, 2018). There is also evidence that increased motivation to achieve a certain outcome can elicit greater engagement in theory of mind in order to persuade another agent to work toward a shared goal (Altiero, Kang, & Peecher, 2022). One survey provided evidence that people self-report consistent levels of "mind-reading motivation"—willingness or desire to consider others' mental states—across time (Carpenter, Green, and Vacharkulksemsuk, 2016). A further distinction can be drawn by studying theory of mind as an implicit vs explicit process (Low & Perner, 2012). It is possible that some elements of theory of mind are intuitive, automatic, and heuristic, while other involve deliberate, reason-based processes—as is the case for other elements of social cognition (Satpute & Lieberman, 2006; Happé, Cook, & Bird, 2017). It has also been proposed that Autism Spectrum Disorder can be understood as a deficit not in capacity to engage in theory of mind, as it commonly thought (Baron-Cohen, 1997), but in the motivation to do so (Chevallier et al., 2012). However, none of these accounts was able to provide evidence based on a task which can reliably dissociate between capacity and motivation to engage in mental state inferences. Specifically, none have provided stimuli that can test whether people find attributing a mental state more difficult, even though they are able to do so accurately—a cognitive effort-based cognate to varying levels of grip force below an individual's max grip strength in physical effort studies (e.g., Lockwood et al., 2021).

It has also been proposed that there is a cognitive effort cost associated with the related process of empathy (Keysers & Gazzola, 2014; Zaki, 2014; Cameron et al., 2019). While empathy is often difficult to disentangle from theory of mind (Batson, 2009), empathy is

generally associated with an assessment of a target agent's affective states whereas theory of mind is associated with assessment of the target's cognitive states (Schurz et al., 2020). The empathy selection task (Cameron et al., 2019; Ferguson & Inzlicht, 2022) has been used to provide evidence that participants show standard signals of effort aversion in choosing between engaging in empathy versus a control task, and that empathy is therefore cognitive costly. As I argue at length in the Introduction (Chapter 1), given the close relationship between the mental processes of empathy and theory of mind, it is reasonable to suspect that theory of mind, if measured appropriately, could elicit some of the same cognitive effort costs which have been associate with empathy.

Thus, my goal in the present chapter is the development of a task where participants reliably perceive the attribution of correct mental states as more difficult in one situation than another. Until such a task has been developed, it will be difficult to provide empirical evidence of the role motivation plays in theory of mind. An important consideration of the novel stimuli described below is that they elicit different levels of perceived difficulty.

Current paradigms in theory of mind.

The existing literature on theory of mind has been criticised for the lack of alignment between its concepts and its experimental paradigms (Quesque & Rossetti, 2020). Quesque and Rossetti argue that in order to truly measure theory of mind, a paradigm must meet two separate criteria. The first is the "non-merging" criterion (which requires that participants make a clear distinction between their own mental states and the other agent's, as opposed to projecting their own thoughts and feelings onto the other). The second is the "mentalising" criterion (which requires that the relevant inferences cannot be explained by lower-level processing; see also

Heyes, 2014). Quesque and Rossetti argue that a large proportion of existing paradigms which purport to measure theory of mind in fact do not meet these two criteria.

In the stimuli set presented in this chapter, I will seek to address three problems in current theory of mind tasks: (1) lack of ground-truth with respect to mental state inferences, (2) ceiling effects in neurotypical adults, and (3) lack of distinction between capacity and motivation to perform mental state inference (i.e., that no existing stimuli vary in how difficult people perceive the mental state attribution to be even though they are able to accurately perform the task). The first problem is that the so-called "mental states" relevant in theory of mind tasks are often fictional or hypothetical, not based on actual people carrying out actual behaviour as motivated by their actual intentions. They are, in other words, a researcher's attempt to create a plausible facsimile of mental states. In this case, whether or not a participant gets a theory of mind problem "correct" depends less on their ability to assess the true causes of a particular instance of behaviour and more on whether participants can arrive at the same conclusion as the researchers. The second problem is that most theory of mind assessments require a binary inference: an observer either identifies the "correct" mental state or not. This would not be a problem in itself, except that in many assessments (detailed below) most neurotypical adults are almost always able to identify the relevant mental state. Failure to identify the correct mental state, then, does not signify less adeptness at mental state inference so much as a potential lack of theory of mind altogether. These paradigms are useful for studying conditions such as Autism Spectrum Disorder or the development of social-cognitive processes in children (Senju, 2012). But they are severely limited for studying individual variation in neurotypical adults.

The following comprise a selection of the most commonly used assessments of theory of mind—all of which suffer from one or both of the limitations described above:

The Sally-Anne Task (Wimmer & Perner, 1983). This task assesses an individual's ability to attribute a false belief to another agent. In this task, a participant witnesses a protagonist (Sally) hiding an item in a given location; once the protagonist leaves, another agent (Anne) moves the item to a different hidden location. The observer knows the true location of the object, but Sally, presumably, does not. Does the observer impute the (false) belief to Sally that the item remains in the first location? While this task undoubtedly taps into a key mind theoretic resource—the ability to distinguish between one's own beliefs and those of another agent—the majority of neurotypically developing children above the age of five are able successfully to solve this problem (Happé, 1995). The task is not capable of assessing variation in mental state inference among neurotypical adults. To the task's further detriment, the two characters are usually rendered as a series of drawings, and therefore neither Sally nor Anne, in fact, have mental states.

The Director Task (Keysar, Lin, & Barr, 2003). In this paradigm, a director gives instructions to a participant to move objects around in a grid. Before receiving instructions, participants hide an object in a bag, which is unknown to the director. When the director gives instructions to move an object that resembles the hidden one in the bag, the participant may still attempt to move the object—even though the director would, presumably, not know to instruct them to do so. The experimenters use reaction time as a proxy for the participant's engagement in taking the other agent's perspective, with longer reaction time suggesting greater "depth" of perspective-taking. However, follow-up studies have suggested that this task does not rely primarily on mentalising, as participants show the same behavioural patterns when the "director" is replaced by an inanimate object, such as a camera (Santiesteban et al., 2015).

The Triangles Task (Abell, Happé, & Frith, 2000). As described above, this task requires participants to assign mental states descriptions to triangles moving across a screen. In this task, there is a lack of ground truth about what the correct mental state is. It was also designed to test theory of mind capacity in non-normally developing populations (such as children with Autism Spectrum Disorder), and therefore most neurotypical adults perform the task adequately.

The Mind in the Eyes Task (Baron-Cohen et al., 2001). In this task, participants view a series of pictures of eyes and adjacent facial regions. Participants choose between choices of words which best describe the emotion or mental state conveyed by the eyes. The test was originally designed to study Autism Spectrum Disorder and therefore is of limited use in studying neurotypical adults. Like the Triangles task above, there is no ground truth between the stimuli and the relevant mental states. Furthermore, as Oakley and colleagues (2016) note, affective states are not the same as cognitive states, and given that theory of mind pertains only to inferences about cognitive states, the task has been argued to be invalid.

Strange Stories (Happé, 1994). This task uses naturalistic stories to assess whether individuals are able to attribute appropriate mental state explanations to the characters. It allows for greater variation and nuance than standard binary theory of mind assessments. However, the "correct" mental state is not veridical; it is defined by the experimenter. It therefore represents the experimenter's interpretations of the mental states in question, not the causal connection between mental states and any given behaviour.

Movie for the Assessment of Social Cognition (Dziobek et al., 2006). Similarly, this task presents participants with visual vignettes about which they are asked questions about the actor's mental states. Its drawbacks are the same as Strange Stories: it requires the participant not to determine the actor's true mental states, but the researcher's interpretation of them.

Surveying these prominent theory of mind tasks, I contend there is currently no paradigm which adequately assesses variation in theory of mind in neurotypical adults. Below, I describe the development of a stimulus set designed to address this deficiency. Across two tests, I show how these stimuli capture variation in accuracy and perceived effort in mental state inference.

Hypotheses and summary of studies.

Across the two studies that follow, I present a novel set of stimuli designed to assess differences in accuracy and perceived effort in mental state inference. In the first study, I describe the stimuli and provide an initial assessment of their use. In the second study, I replicate those initial findings, showing that these stimuli are reliable assessments of individual performance across the span of several months. The stimuli are based on the "Assassin Game," which I created for the purpose of making the stimuli. As described below, in each round participants are required to identify the mental states underlying various goal-directed behaviours of real people controlling digital avatars. Recordings of individuals playing the game were then segmented into five-second clips. My aim was to create short clips that could be used to assess theory of mind in neurotypical adults. In order for it to be an effective theory of mind task, there should be significant variation in both accuracy and perceived difficulty associated with the mental state inferences required by different clips. I will argue that these clips rectify the three problems associated with most current theory of mind paradigms, as described above. These stimuli: feature a causal relationship between mental states and behaviour, demonstrate empirical variation among performance by neurotypical adults, and are capable of distinguishing between capacity and motivation to perform mental state inference.

Study 5.1: Theory of Mind Stimuli Assessment

I developed a set of stimuli to act as a novel assessment of perceived effort and accuracy in theory of mind. These stimuli consist of 104 five-second video clips showing segments of what I call the "Assassin Game." Each video clip shows the movements of five players, each controlling an individually identifiable shape. Each clip features the players engaged in the

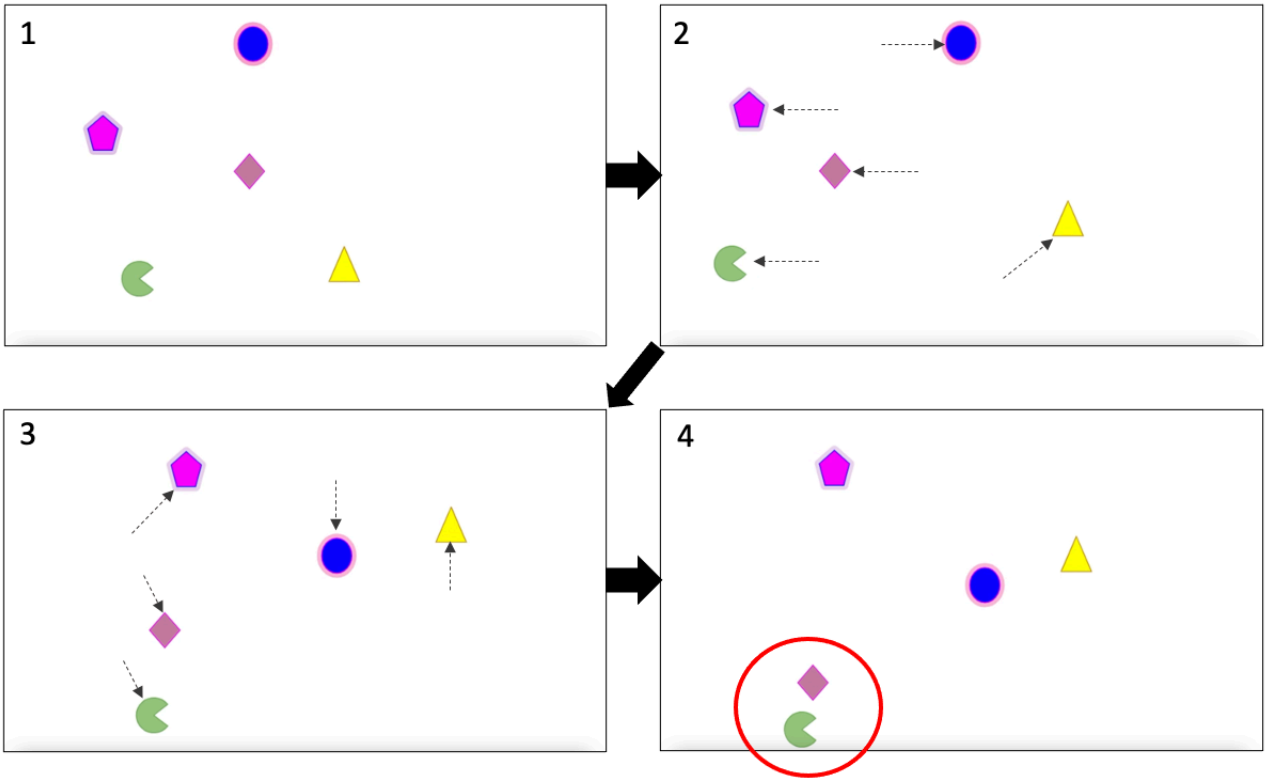


Figure 1 — Four frames from a single round of the Assassin Game. One of the five players has been assigned as the "assassin" and given a "target" which their goal is to make contact with. All players not assigned to be the assassin are tasked with avoiding a potential assassin, on the possibility they are the target. The frames shown here are not necessarily consecutive (in that players have only made one move) but show four progressive points within the round. Motion vectors are approximate. (**1**) Initially, each agent starts off in a pseudo-randomised position. (**2**) As each player begins to move from their starting position, there is significant ambiguity about which player is the assassin. (**3**) Eventually, the assassin will go directly toward the target, while other players do not maintain consistent goal-directed actions. It should be noted that the assassin can attempt to trick targets by not immediately going toward them, and that other players can pretend to be assassins by temporarily going toward another player. (**4**) The round is over when the assassin (in this case, the pink diamond) touches the target (the green Pacman).

game, and the participant must infer the mental states of the players controlling the shapes—at least enough to infer which player is the assassin.

The Assassin Game. The Assassin Game consists of five players simultaneously controlling the movement of individually identifiable shapes on their computer screen (Figure 1). The goal of the game is for the assassin player to "kill" the target player. The assassin does this by touching the target shape with their own shape. On each round, one player is randomly assigned to be the assassin. Only the assassin is informed of their role and which shape is the target. All other players must operate under the assumption that they *may* be the target. A round ends once the assassin has successfully touched the target.

Video capture. The game was implemented using the Google Slides web app, which allows multiple people to simultaneously modify the elements of a slideshow frame. The stimuli were created across two separate sessions. In each session, I convened five colleagues to play approximately thirty rounds of the game over the course of an hour. Each person was seated in front of a laptop computer in the same room, without the ability to see one another's screens. As the session "host," I also was seated at a computer. Each player was sent a link to a shared frame of a slideshow in Google Slides. This slide contained five shapes—a triangle, a circle, a "Pacman," a square, and a pentagon—with each player assigned to control one of the shapes. The rules, as described above, were explained to each player. The host's computer also displayed the same shared frame, but instead of controlling a shape the host's computer was used to screen-record each round of the game. A rectangular boundary was set as the game "board" in which players had to remain during each round (the size of the boundary was enlarged in a subset of the videos from the first session, to provide multiple scales of the shape-to-screen ratio). Rounds featuring a breach of protocol were discarded.

The five players controlled their shapes using their computer's keypad: the left, right, up, and down arrows. During each round an audible metronome was set at 60 beats per minute. Each player was required to make exactly one key-stroke on each beat (remaining stationary was not allowed). This severed two functions. The first was to accommodate the slow latency of simultaneous updates in Google Slides. That is, it takes a split-second for one person's move to appear on another person's screen. With the metronome set at this tempo, each player could see the other players' movements before making their next key-stroke. The second function was to keep each shape moving at the same velocity. With everyone moving at the same tempo, no one could "outrun" their opponents by simply clicking their keys faster.

Each player had a unique slide deck designating their role in each round. The slide deck was positioned on their screen so that only the current round was visible. Before each round, they would click to the next slide. On each round one player was informed that they were the assassin and which shape was their target; other players received no information. Assassin and target assignments were approximately equally distributed among all players. Between the two sessions, sixty-two rounds were completed.

Video editing. As each round continued until the assassin completed their mission, the raw video clips were of variable lengths. The iMovie video editor was used to segment each video into clips of five-seconds. I used my own discretion to select these five-second segments. My intention was to obtain a wide range of player behaviour in the clips. I also omitted segments in which no relevant behaviour took place (e.g., protracted periods in which the real assassin maintained the subterfuge of not overtly going after their target), as these would be incomprehensible to third-party observers. I attempted to include a range of clips—from the obvious, in which the assassin is "heat-seeking" for the entire five seconds (usually right before

the termination of a round) to more subtle movements. An example of a more subtle movement would be when the assassin follows a faux-target for most of the clip, only to deviate in the last moment toward their actual target. I allowed multiple clips from the same round to overlap, as long as they depicted qualitatively different behaviour (for instance, if the assassin deviated from their subterfuge at the last moment of one five-second clip, only to spend the majority of the next five-second clip heat-seeking). In total, this resulted in 104 five-second video clips.

Participants. Eighty-five participants were recruited on the Prolific subject recruitment platform (47 female; mean age = 27.27; mean Prolific score: 99.1; distinct birth countries represented = 24). All participants provided written informed consent. The study was approved by the local Ethics Committee. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately fifty minutes to complete.

Method.

Design. Participants were presented with each of the 104 video clips in randomised order. On each trial, they were shown a video clip. The clip started when the participant clicked "play." After the conclusion of each clip the participant was asked to select which shape was the assassin from pictures of each of the five candidate shapes. Once the participant made their selection, they were asked which of the five shapes was the target. (The list of shapes were pseudo-randomised, so that the shapes were not presented in the same order each time.) On the following screen, participants were asked to use five sliding scales (from 0 to 100) to rate: the "difficulty"

in selecting the assassin; the "difficulty" of the target; their "uncertainty" in selecting the assassin; their "uncertainty" about the target; and the "difficulty" of the video overall.

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Prolific subject recruitment platform.

Questionnaire. Immediately after providing consent, each participant completed the Questionnaire of Cognitive and Affective Empathy (QCAE; Reniers et al., 2011).

Post-task self-report judgments. After completing the main experiment, each participant was asked to use a sliding scale to report how "interesting" they found the task, how "difficult" they found it, and how often they felt they forgot their answers before responding to the questions.

Attention checks. Participants were asked at the end of the experiment to report the extent to which they paid attention during the entirety of the experiment—both in a slider and a verbal description. They were ensured this report would not affect their payment. There were also three "catch-trials" during the main task, in which text appeared in the video instructing the participant to disregard the movement of the shapes and instead select a specific shape as the assassin in the subsequent prompt.

Results.

The clips showed a wide range of variation in participants' ability to accurately infer the assassin and the target, as well as the perceived difficulty of doing so.

Figure 2 shows the variation in both accuracy (in terms of the percentage of participants who correctly identified the assassin and, separately, the target) and perceived difficulty for individual video clips. The intention was to develop stimuli that covered a range of accuracy and

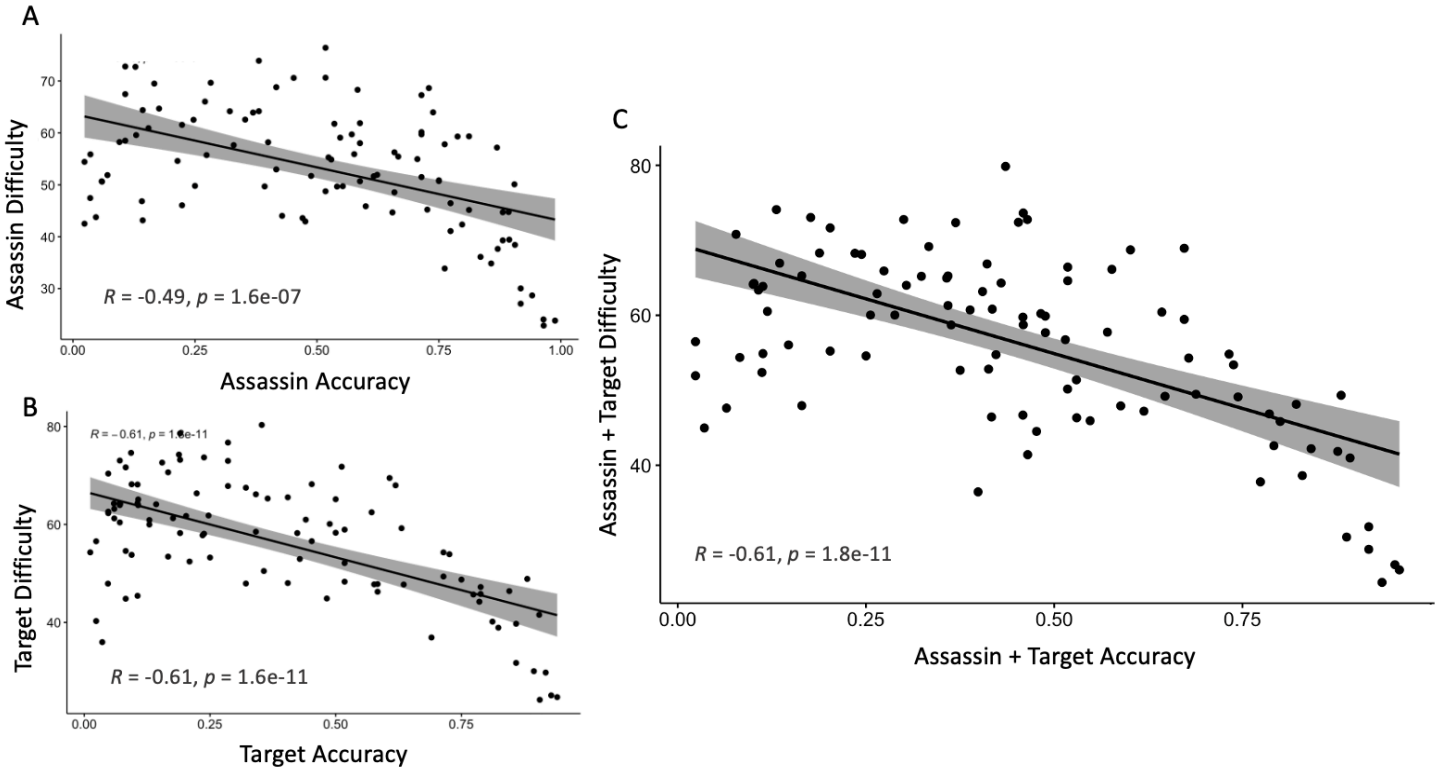


Figure 2 — Video clips varied in their average accuracy and perceived difficulty. (**A**) This plot shows each video clip's average accuracy for participants selecting the correct assassin as a function of how difficult the participant reported the assassin selection to be. (**B**) Each video clip's average target accuracy as a function of target difficulty. (**C**) Each video clip's combined accuracy (assassin + target) as a function combined difficulty.

difficulty levels, and indeed the stimuli are approximately uniformly distributed between 0 and 100 for accuracy (from almost no one to almost everyone getting the clip correct). The vast majority of clips were above chance level of accuracy (20%), and thus participants were clearly able to identify the target and assassin, attributing the correct mental states to the appropriate agents. There is also little indication of a ceiling effect, with only three clips judged correctly for both assassin and target near 100% of the time. While there is a correlation between accuracy and difficulty (r = -0.61, $p < 0.001$; Figure 2C), there is still a large amount of variation in perceived difficulty across all ranges of accuracy—as well as for assassin judgments (r = -0.49, $p < 0.001$; Figure 2A) and target judgments (r = -0.61, $p < 0.001$; Figure 2B) on their own. For

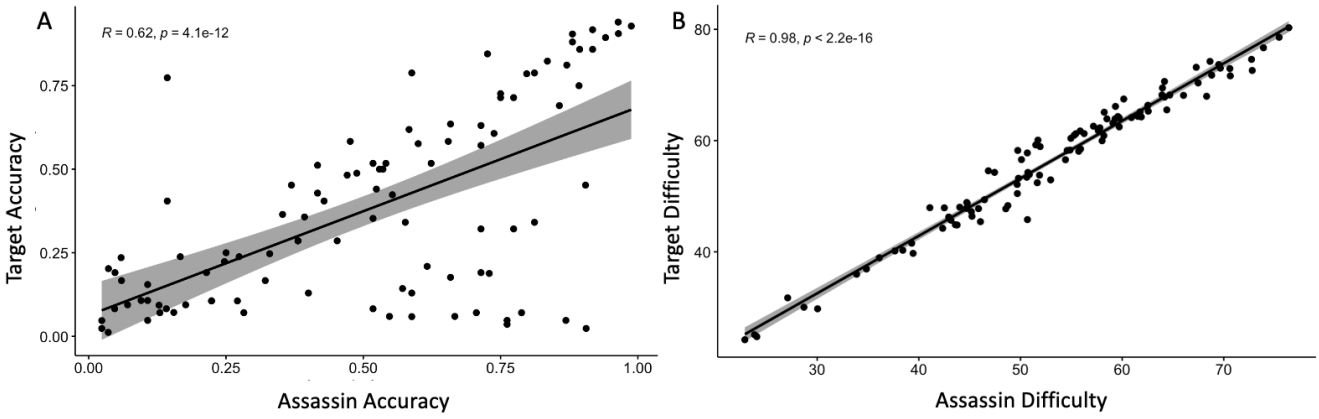


Figure 3 — There was a dissociation between assassin and target accuracy; but the perceived difficulty between targets and assassins was almost identical. (**A**) For any given clip, assassin accuracy only imprecisely predicted target accuracy. This suggests that selecting the correct assassin and target represent two substantively different tasks. (**B**) Participants did not appear to distinguish between the difficulty in assessing the assassin and the target agents.

clips with less than 50% accuracy, there is no correlation with perceived difficulty (r = -0.046, p = 0.72).

There was also a similarly-sized correlation between assassin accuracy and target accuracy across all clips (r = 0.62, p < 0.001; Figure 3A). While one would expect these to be correlated (e.g., greater certainty about the identity of assassin should lead to greater certainty about who their target is), it is important that they are not identical. This remaining variance shows that selecting the correct assassin and target represent two substantively different tasks. While most of this variation consisted in videos for which the assassin accuracy was higher than target accuracy (i.e., participants could tell which agent was the assassin, but not which agent they were targeting), there remained a few videos for which participants were able to accurately identify the target but not the assassin. Meanwhile the perceived difficulty of assessing the assassin and target were so highly correlated as to be functionally identical (r = 0.98, p < 0.001; Figure 3B). It should also be noted that participants' reports of "difficulty" and "certainty" were

nearly identical (though with reversed sign) for both assassins (r = -0.98, p < 0.001) and targets (r = -0.99, p < 0.001).

Participants varied in their accuracy and perceived difficulty.

In addition to assessing variability of accuracy and difficulty on average for each clip, it is also possible to examine variation between participants (Figure 4). Participants' overall accuracy did predict their reports of perceived difficulty, but the correlation was fairly weak compared to the effects at the level of individual clips in terms of combined assassin and target judgments (r = 0.26, p = 0.02; Figure 4C), as well as individually for assassins (r = 0.23, p =

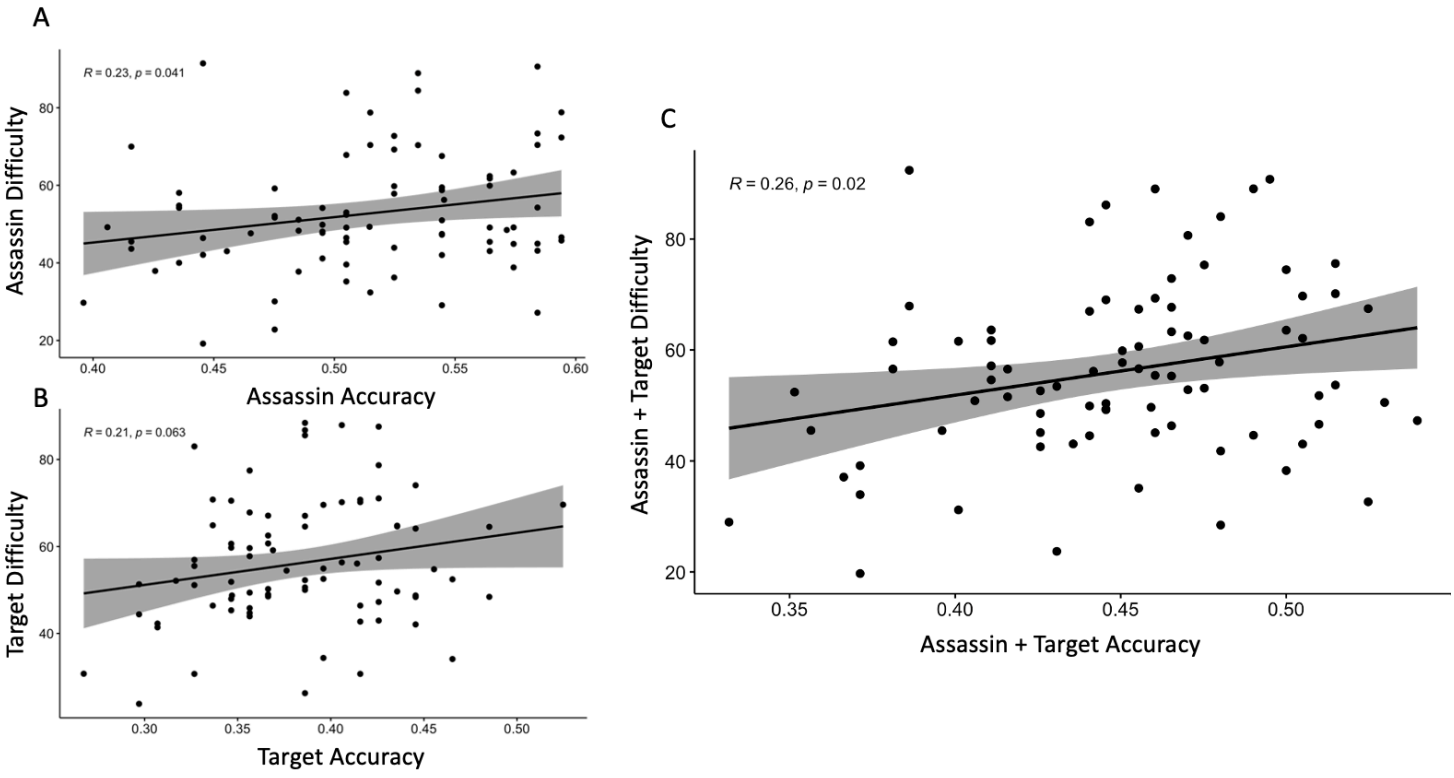


Figure 4 — Participants varied in their average accuracy and perceived difficulty. (**A**) This plot shows each participant's average accuracy for selecting the correct assassin as a function of perceived difficulty. (**B**) Each participant's average target accuracy as a function of target difficulty. (**C**) Each participant's combined accuracy (assassin + target) as a function combined difficulty.

0.041; Figure 4A) and targets (r = 0.21, p = 0.063; Figure 4B). The range of participant accuracy scores was less than the accuracy scores of the clips. To some degree one would expect these ranges to have an inverse relationship: if many clips are almost always correctly assessed by participants, then that raises the floor of the participants' overall scores; if many clips almost always incorrected assessed, that lowers the ceiling. Given the uniformity of the range of the clips, participants scores—for both accuracy and difficulty—fell within a fairly centralised range.

A measure of empathy predicted participant's perceived difficulty—but not accuracy; this effect is driven by cognitive (rather than emotional) empathy.

Participants' scores on the QCAE self-report assessment of empathy did not predict their average accuracy scores combined for both assassins and targets (r = -0.15, p = 0.17; Figure 5A). However, the empathy scores were a significant predictor of the average perceived difficulty

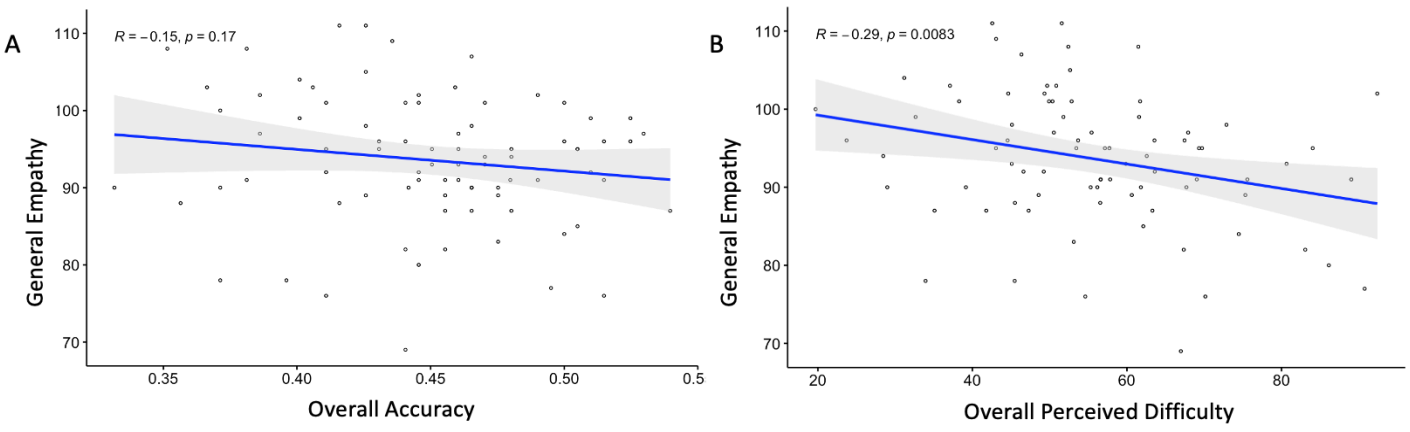


Figure 5 — A self-report measure of empathy predicted participants' average ratings of perceived difficulty, but not their accuracy. (**A**) There was no significant correlation between participants' scores on the QCAE assessment of empathy and their accuracy in assessing assassins and targets. (**B**) However, there was a significant correlation between empathy and average perceived difficulty.

across all trials (r = -0.29, p = 0.008; Figure 5B). This correlation shows that the higher participants self-reported empathy scores, the less difficult they rated the task to be.

There is a further dissociation when these correlations are broken down into the components of cognitive and affective empathy (Figure 6). As with general empathy, a participant's accuracy was predicted by neither cognitive empathy (r = -0.2, p = 0.07; Figure 6A) nor emotional empathy (r = -0.01, p = 0.92; Figure 6B). Emotional empathy scores contributed little to the correlation between empathy and difficulty (r = -0.10, p = 0.38; Figure 6D). The

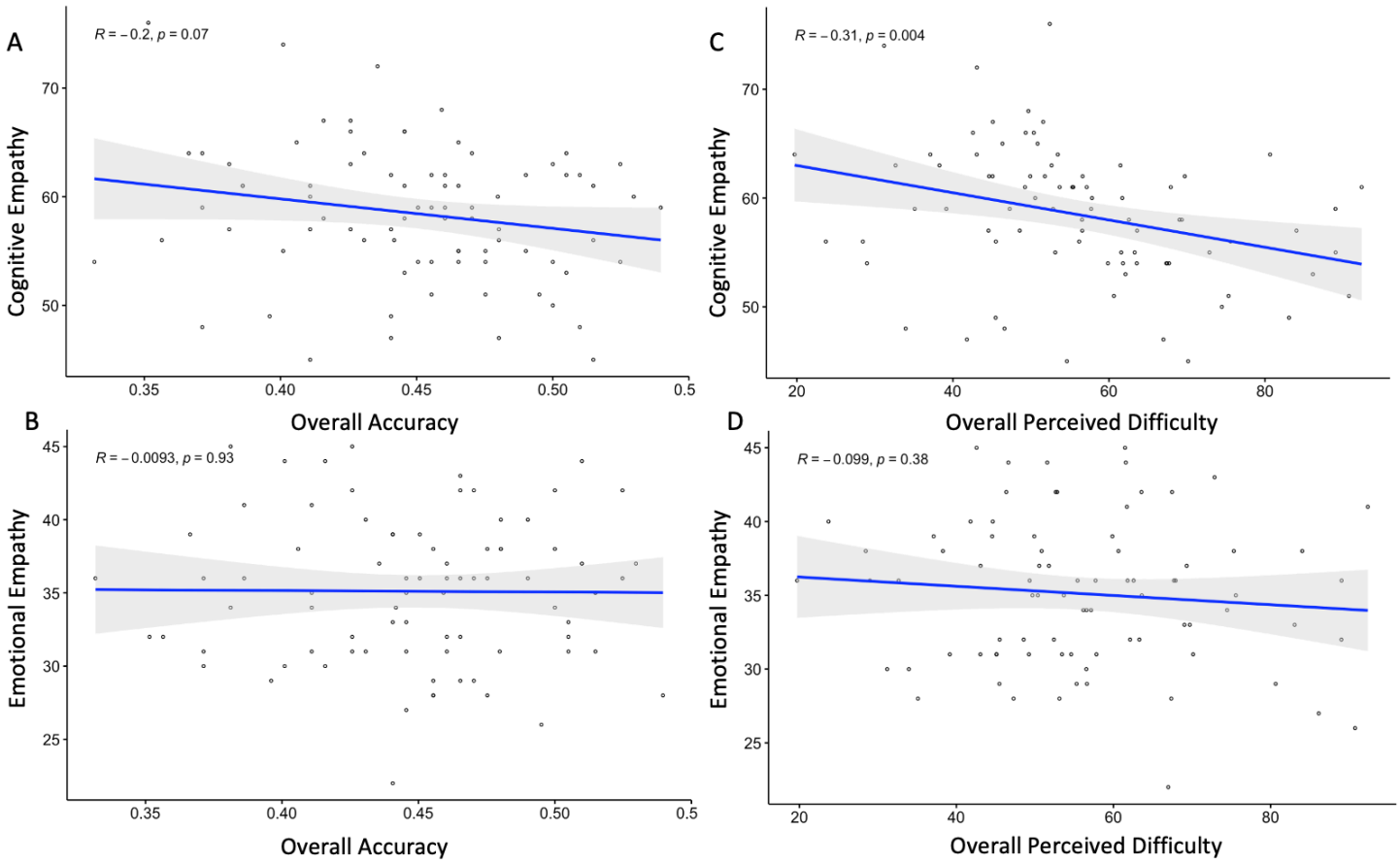


Figure 6 — Cognitive empathy predicts perceived difficulty, but not accuracy, while emotional empathy predicts neither. (**A, B**) As with the general empathy score, there was no significant correlation between participants' empathy scores and their accuracy in assessing assassins and targets. (**C**) Cognitive empathy predicts perceived difficulty, while (**D**) emotional empathy does not.

effect appears to have been driven almost entirely by cognitive empathy (r = -0.31, p = 0.004; Figure 6B).

Discussion.

Here, I proposed to develop stimuli which assessed mental state attribution accuracy and difficulty in neurotypical adults. I intended these stimuli to vary in the accuracy scores and difficulty ratings associated with identifying the assassin and target agents, based on the behaviour of the agents controlling the avatars (rather than the experimenter's intended interpretation of an animation). These stimuli operationalise mental state inference based on participants' ability to identify which shape is operated by an individual with the intention to target another shape (or to elude a shape which is targeting them). As intended, these stimuli capture a large range of variation in participants' accuracy in identifying the correct assassin and target, as well as their perceived difficulty in doing so. This range is reflected in the scores both for individual video clips as well as individual participants. There is also a dissociation between the accuracy for assassins and targets—suggesting that these represent two separate tasks. These results support the claim that these video clips comprise a stimulus set for which the ground-truth with respect to mental state inferences are known, and for which neurotypical adults show variation in their ability to make accurate mental state inferences.

Moreover, I found that the perceived difficulty of identifying the target and assassin—but not the accuracy of making this judgment—was associated specifically with the cognitive component of trait empathy. These results suggest that probing people's motivation to undertake difficult theory of mind tasks may be important for understanding variability in social cognition, rather than just accuracy as often tested in the extant literature (Contreras-Huerta, Pisauro, & Apps, 2020). In terms of implications for the existing questionnaire (QCAE), one interpretation

for this dissociation between difficulty and accuracy is that when a participant answers a question from the assessment—such as "I try to look at everybody's side of a disagreement before I make a decision"—her answer reflects the perceived difficulty in engaging in this process, not her capacity for carrying it out. This is consistent with the distinction made in the Chapters 2 through 4 of this thesis between the capacity and motivational components of exerting effort. Another possibility is that both of these measures—the questionnaire and the perceived difficulty rating—rely on self-report and are therefore correlated based on whether or not a participant tends to exaggerate her reports. (Though this cannot provide the entire explanation, given that there is only a correlation with cognitive, and not emotional, empathy.) Ultimately, the QCAE is intended to measure a trait, whereas the present stimuli are task designed to elicit processes related to that trait; it is therefore a useful check that there is a correlation in the expected direction and with the relevant trait (i.e., cognitive, rather than emotional, empathy).

Study 5.2: Theory of Mind Stimuli Test-Retest Reliability Assessment

In order to ensure the reliability of the initial assessment of the stimuli, I retested a subsample of the participants tested in Study 5.1 to see whether they achieved a comparable score when retested.

Participants. Forty-one participants from the initial cohort of 85 underwent the same testing protocol as detailed in Study 5.1. Participants were invited to take part in what they were told was another version of the original study. The initial round of testing was conducted on 19-21 July 2021 and 8 August 2021. The second round of testing was made available to participants on 2 November 2021 (approximately three months after the initial test). The majority of

participants who submitted re-test responses did so within one week of the re-test study becoming available, with the final participants submitting by 23 November 2021.

Method.

Design. The design of this task was identical to the one described in the initial assessment, with the sole difference that participants were told they were participating in another "version" of original study—though it was in fact precisely the same protocol.

Results.

The video clips showed strong test-retest reliability.

Figure 7 shows test-retest reliability between the two assessments. In the subset of participants who completed both sessions, there was a strong correlation between the average accuracy scores across all video clips ($r = 0.98$, $p < 0.001$; Figure 7A). The difficulty ratings were correlated to a similar degree ($r = 0.95$, $p < 0.001$; Figure 7B).

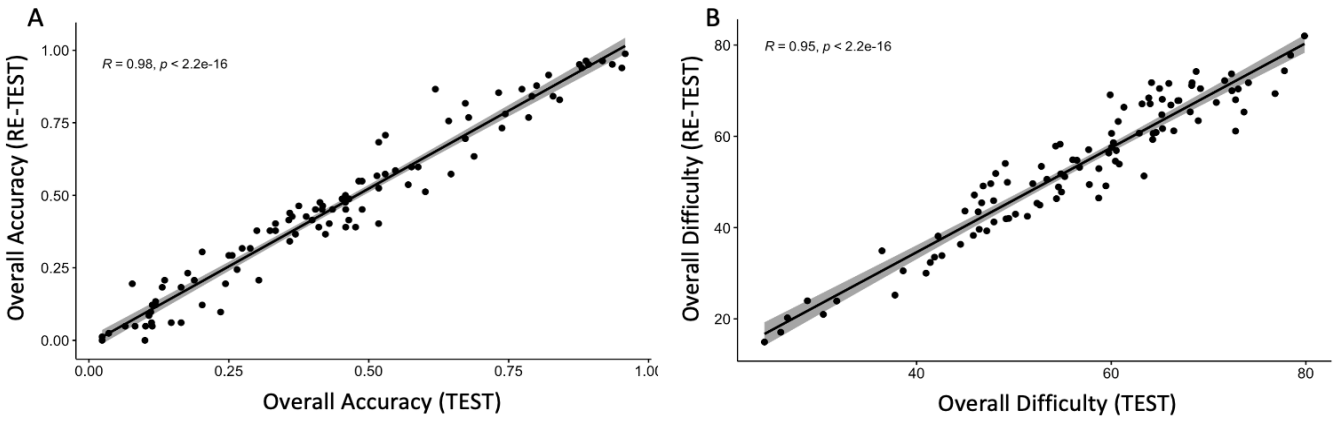


Figure 7 — Test-retest reliability for each video clips' accuracy scores and difficulty ratings. (**A**) The percentage of participants who got the assassin correct on each video clip between test and retest was highly correlated. (**B**) Similarly, participants' difficulty ratings for each video showed high test-retest reliability.

These correlations held across accuracy scores for both assassins ($r = 0.97$, $p < 0.001$) and targets ($r = 0.98$, $p < 0.001$), as well as the difficulty scores for assassins ($r = 0.94$, $p < 0.001$) and targets ($r = 0.94$, $p < 0.001$).

Participants' performance was also reliable between tests.

Figure 8 shows the test-retest reliability of scores from individual participants. There were significant correlations for both accuracy scores ($r = 0.58$, $p < 0.001$; Figure 8A) and difficulty ratings ($r = 0.80$, $p < 0.001$; Figure 8B). Overall, test-retest reliability was higher for targets than assassins in both accuracy scores (target accuracy: $r = 0.55$, $p < 0.001$; assassin accuracy: $r = 0.45$, $p = 0.006$) and difficulty ratings (target difficulty: $r = 0.83$, $p < 0.001$; assassin difficulty: $r = 0.79$, $p < 0.001$).

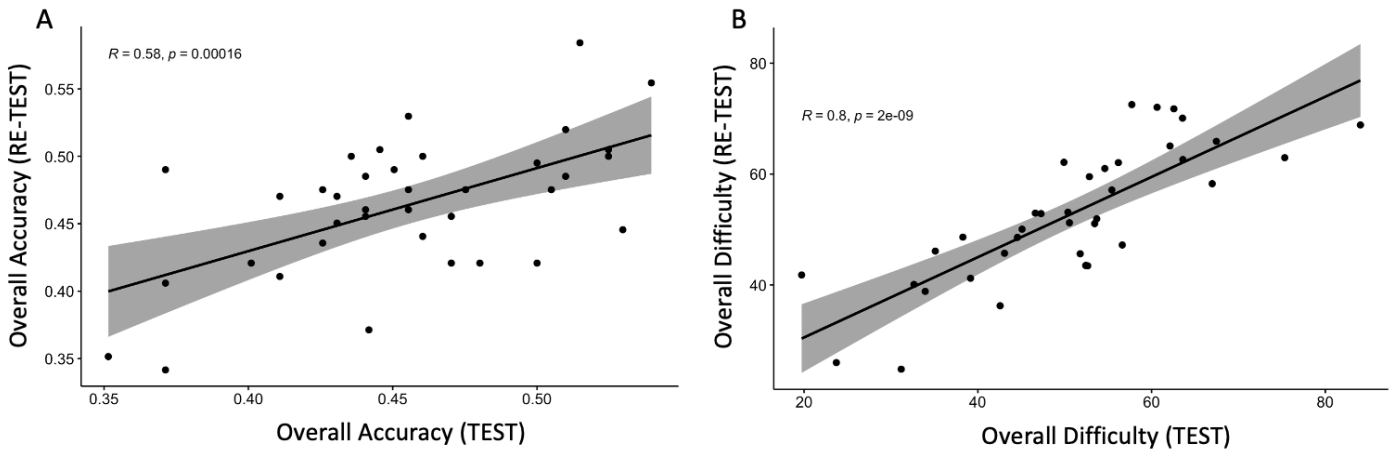


Figure 8 — Participants' individual scores were also reliable across tests. (**A**) Participants' test-retest correlations for average accuracy scores. (**B**) Participants' test-retest correlations for average difficulty ratings.

The participants' scores on the video clips were comparably reliable to their QCAE scores—though reliability in one did not predict reliability in the other.

Figure 9 shows participants' test-retest reliability on the QCAE self-report assessment of empathy. The degree of test-retest reliability on the QCAE fell between that of the difficulty ratings and accuracy scores for the video clips ($r = 0.73$, $p < 0.001$; Figure 9).

I also calculated the intra-individual difference in scores between test and retest to investigate whether reliability differed between these tests in consistent ways. For example, if there was systematic variation between the cohorts in two sessions (e.g., if all participants improved in their empathic abilities between late summer and early autumn), then one would expect to see intra-individual differences vary between two tests in the same way. However, I found no such correlations between the intra-individual difference in QCAE and accuracy scores ($r = -0.054$, $p = 0.75$), the intra-individual difference in QCAE scores and difficulty ratings ($r =$

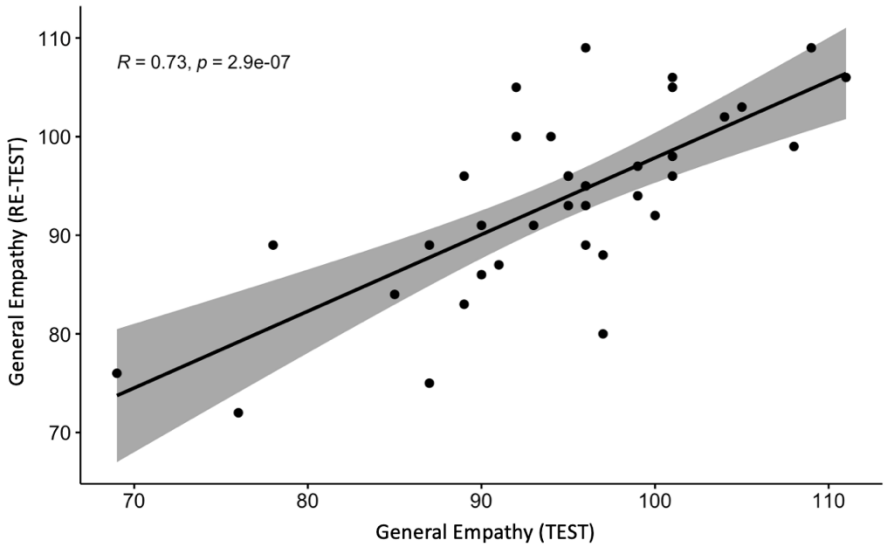


Figure 9 — Participants' test-retest reliability for the QCAE. Participants' test-retest reliability on this validated assessment of empathy fell somewhere between reliability for accuracy scores and difficulty ratings in the videos clips.

-0.11, $p = 0.51$), or even the intra-individual difference in accuracy scores and difficulty ratings ($r = 0.054$, $p = 0.75$).

Discussion.

These video clips show strong test-retest reliability in terms of both the probability that an individual will make an accurate mental state inference and the difficulty reported by participants in making that inference. The accuracy scores and difficulty ratings for individual video clips were almost identical between sessions. This suggests that the videos themselves reflect actual differences associated with the mental states of the agents: the percentage of the time the assassin is accurately identified represents the extent to which people can reliably infer their goals based on their behaviour.

The reliability for test-retest within an individual was comparable to their scores on the QCAE—with accuracy scores a little less reliable, and difficulty ratings a little more reliable. The fact that there was no systematic variation between the intra-individual differences in these scores suggests that any test-retest differences were due to natural variation in the assessments themselves, rather than any substantive differences in the cohort between the two timepoints.

General Discussion

This novel stimuli set represents the first assessment of theory of mind which can identify variation in accuracy and perceived difficulty in neurotypical adults based on observation of behaviour which is causally linked to the underlying mental states. Previous tasks intended to assess theory of mind, as I described in this chapter's Introduction, have tended to suffer from one of two drawbacks: lack of ground-truth knowledge of the mental states in question, or ceiling

effects in neurotypical adults. These stimuli address both of those issues. Real humans control the avatars to reflect the goal-directed mental states "catch the target" or "don't get caught by the assassin." They also elicit a range of scores among neurotypical adults, both in terms of participants' accuracy and their perceived difficulty.

The role of effort in mental state inference.

These data are consistent with an account in which this task—which depends on theory of mind to complete—is perceived as cognitively effortful. Across the two test-retest sessions, participants reliably made similar reports of how difficult they perceived each clip to be. Furthermore, while perceived difficulty and accuracy scores were correlated, there was quite a dissociation between the two. This implies that participants' perceptions of difficulty were largely independent of their ability to complete the task, which is consistent with the distinction between capacity and motivation I have drawn throughout this thesis (Contreras-Huerta, Pisauro, & Apps, 2020). As Contreras-Huerta and colleagues have argued: while the baseline perception of other agents as intentional is likely automatic, it is likely that figuring out what their specific mental states might be can require a greater amount of effort. In the following chapter, I will extend these preliminary findings, which are based on self-report, to investigate evidence of effort discounting with these stimuli using a version of the Prosocial Effort Task (as in Chapters 2 through 4). As Contreras-Huerta and colleagues note in their review, since all existing evidence for effort in theory of mind is indirect, the present work provides the first direct evidence of such effort costs.

Notably there was variability in perceived difficulty even for stimuli that participants were highly accurate in getting correct. This suggests that the stimuli are akin to those used in

physical effort paradigms, where levels of grip force are below participants maximum grip strength, and yet people rate them as effortful and choose to avoid them (Chong et al., 2017). This is an intriguing finding, as for many stimuli used in cognitive effort experiments people do not always succeed at the supposedly higher effort levels (Westbrook, Kester, & Braver, 2013). Thus people might rate it as more effortful because they simply think they will not succeed at it. Here, I show that even at the highest level of accuracy there is variability in difficulty ratings, suggesting that difficulty is not just associated with accuracy for these mental state attribution stimuli.

While in the present work there are differential effort costs associated with mental state inference in the various clips, it is not possible based on the present work to say where exactly that difference in effort cost comes from. This is both a strength and a weakness. It is a strength because it has proved difficult in previous work to theoretically define what exactly about theory of mind might be cognitively effortful. These stimuli are able to measure differences in effort required by different mental state inferences by seeking to identify those differences empirically, rather than specifying a priori where those differences might come from (e.g., in a more or less elaborate false belief scenario). It will be up to future research to provide a detailed examination of where precisely these effort costs come from. An important question consists in whether the mental state inferences in this task represent the activity of an automatic, heuristic-based system or a deliberative, reason-based system (Low & Perner, 2012). Perhaps the most direct test of this question would be to have participants perform these judgments under time pressure (e.g., Phillips, Morris, & Cushman, 2019); if there is significant deviation, then they likely recruit a deliberative, rather than automatic, system. However, based on the present data, it seems likely that these inferences recruit a deliberative, reason-based process, as they seem to require a non-

negligible amount of cognitive effort. In the General Discussion (Chapter 7), I will provide a detailed proposal of what future experiments might answer this question most effectively.

These findings are consonant with my hypotheses presented both in Chapter 1 and at the beginning of this chapter: namely, that theory of mind might incur similar effort costs as has been shown in recent work on empathy. The most direct test of effort costs in empathy has come from the empathy selection task (Cameron et al., 2018). The empathy selection task provides participants with a choice between two decks, one in which they are asked to empathise with the feelings of a pictured individual and one in which they are asked to describe external characteristics of this individual. It has been shown that people consistently choose the external deck over the empathy deck (Ferguson & Inzlicht, 2022). One limitation of the empathy selection task is that it is not possible to know whether participants are less likely to select the empathy deck because they find the affective identification or resonance component of empathy to be more cognitively effortful—is it feeling the same feelings are the target agent, or figuring out what the target agent is feeling that is mentally taxing? In contrast, the present study is able to isolate the mental state inference component of theory of mind, in which participants are only identifying the contents of another agent's mind. The present results suggest that this inference process may be cognitively effortful. Cameron and colleagues (2018) also provide results from a version of the empathy selection task in which participants are asked to reflect on their own emotional state, rather than that of another. Combined with the present results, this suggests that what may be cognitive effortful is the inference problem which requires interpreting the contents of a person's unobservable cognitive and affective states based on their observable expressions and behaviours.

Preliminary cognitive empathy findings.

Additional support for the validity of these stimuli is that a separate measure of "cognitive empathy" (i.e., theory of mind; the Questionnaire of Cognitive and Affective Empathy) predicts perceived difficulty. By contrast, if the results had yielded a correlation with affective empathy, then this would have raised concerns about why stimuli which are designed to assess inferences about cognitive states appear to be picking up a signal about emotional ones. However, it should be stated up front that these correlations are preliminary. Further studies would be necessary to provide support for the finding or to tease apart any causal structure. With that caveat in mind, it is potentially of interest that the particular outcome the cognitive empathy measure correlates with is perceived difficulty and not accuracy. One possibility is that, when completing the QCAE, participants respond according to how difficult they find it to make theory of mind inferences rather than their perceived accuracy (e.g., "I sometimes find it difficult to see things from the 'other guy's' point of view."). In other words, people who score relatively high on the cognitive empathy component of the QCAE may be perfectly fine at making inferences about other minds; they just find it, for whatever reason, difficult to do so. If, as I argue throughout this chapter, there are not currently any tests of theory of mind that reflect variation in ability among neurotypical adults, it is plausible to assert that this pattern of responding in the QCAE has not yet been identified.

How much theory of mind does this task require?

As with any social-cognitive process, it is possible to ask whether the cognitive capacities at hand are *truly* social (in this case, where participants are relying on judgments about mental states), rather than representative of more domain-general processes (e.g., Lockwood, Apps, &

Chang, 2020). To this end, does this paradigm satisfy Quesque and Rossetti's (2020) dual criteria of mentalising (that the same inferences could not be explained by lower-level processing) and non-merging (that there must be a clear distinction between one's own mental states and the mental states of others)? One possible concern is a potential explanation via lower-level processing. Could participants make the same pattern of inferences by assuming the physical stance? For instance, perhaps a simple heuristic could be at play, such as computing the trajectory of the agent in their final moments of the clip and selecting whichever agent is first in line with that trajectory as the target. However, lower-level cues are always going to be present when assessing visual stimuli; and there is still the issue of mapping these lower-level cues onto the attendant mental states. Either way, this potential concern is consistent with the literature on goal-directed definitions of theory of mind surveyed in the introductory section of this chapter. For example, the Baker, Saxe, and Tenenbaum (2009) model clearly relies on a calculation of trajectory (i.e., which target destination is the agent pointed toward?) in their experimental assessment of theory of mind. This same criticism by appeal to lower-level visual processing applies to any paradigm employ simplistic avatars (such as a shapes) moving and interacting on a screen. That does not dismiss the criticism as invalid; but it does contextualise it as a potential drawback of this (longstanding) strategy for constructing paradigms, rather than a drawback specific to this set of stimuli on its own.

Additionally, for Quesque and Rossetti's non-merging criteria, one potential concern is that perhaps participants might not fully represent the assassin's mental states, but rather to ask what moves they themselves would make if they were going after a particular target. However, I would argue that this concern does not hold due to the counterfactual nature of the attributed mental states. That is, participants do not believe that they themselves are the assassin going after

a target. They are (potentially) asking what actions they *would* take if it were the case that they were the intentional agent controlling an avatar that they do not in fact control.

Another reason to believe why this task requires true mind theoretic capabilities is the potentially large range of different strategies the assassin can employ in their attempt to win the round. The simplest strategy is "heat-seeking"—going straight for the target as soon as the round begins. However, the assassin can also follow a more subtle strategy. For example, an assassin could move in the opposite direction from a target, in an effort to lull the target into a false sense of security. Alternatively, the assassin could "stalk" the target by introducing a sufficient amount of noise to their heat-seeking to conceal their intentions (Gao & Scholl, 2011). Yet another strategy would be to pretend to follow a shape that is not the target, deviating only as the real target unsuspectingly draws near. Meanwhile, a player who is not the assassin might pretend to "go after" another player; if that player scampers off in fear, they are unlikely to be the assassin. There are therefore many behaviours afforded by the intention of the assassin in their pursuit of the target, but all motivated by the same underlying goal.

In summary.

I have argued that these stimuli represent a novel assessment of accuracy and perceived difficulty in a theory of mind task. This novel assessment overcomes three limitations of previous theory of mind paradigms: that they do not have a true causal connection between the observed behaviour and the mental states in question; that they are unable to distinguish between individual differences in neurotypical adults; and that they do not sufficiently distinguish between capacity and motivation. Furthermore, the behavioural data based on these stimuli suggest that people do find this theory of mind task effortful, in a way that is dissociable from

their ability to perform it. In the following chapter, I will explore this potential finding further by using these theory of mind stimuli in an effort task based on those used in Chapters 2 through 4.

CHAPTER 6:

EFFORTFUL THEORY OF MIND AS A FUNCTION OF GROUP MEMBERSHIP

***Abstract**. This chapter investigates possible effects of intergroup bias in theory of mind. I begin by surveying current literature evaluating the evidence of whether intergroup bias might play a role in theory of mind—particularly, with people tending to exert less effort to make inferences about out-group minds. In Study 6.1, I begin by using my novel theory of mind stimuli developed in the previous chapter to show that reliable effort aversion effects can be observed, as they are in previous versions of the Prosocial Effort Task (e.g., Study 2.2). In the two subsequent studies, I attempt to introduce a minimal groups manipulation to affect participants' judgments about the theory of mind stimuli. Neither study yielded significant results. In the Discussion, I explore the limitations of these paradigms for addressing my questions of interest and what the current results indicate about potential intergroup bias in theory of mind.*

Introduction

People tend to avoid cognitive effort (Kool & Botvinick, 2018). In the previous chapter, I provided evidence, based on self-report measures, that participants find the mental state inferences required by my novel theory of mind stimuli difficult. Will these stimuli also elicit standard effects of effort aversion, such as those shown in the Intergroup Prosocial Effort Task employed in Chapters 2 through 4? I will also explore whether the mental state inferences required by these stimuli elicit greater or lesser amounts of perceived effort by framing them in an intergroup scenario. Chapters 2 and 4 showed strong effects of group membership on prosocial motivation. Does that intergroup bias extend to motivation to engage in theory of mind as well?

Motivation and capacity in theory of mind.

A key distinction drawn throughout this thesis has been that deficits in social-cognitive processes can be a product of either diminished capacity or motivation (Contreras-Huerta, Pisauro, & Apps, 2020). Following this line of argument, existing evidence suggests two ways in which a particular mental state inference might be harder than another. The first is that one mental state inference might require more capacity than another. In this case, a given mental state inference would be harder (or result in lower accuracy) for the same reason that one might struggle with a tricky calculus problem over an easy algebraic one: it is simply a more difficult reasoning problem (Navarro, 2022; Navarro, Goring, & Conway, 2021). Another analogy would be with physical effort. If the distance between two points in space is greater, then it takes more physical effort to walk the longer distance than the shorter one. If the distance between two minds is greater, then it might take more cognitive effort to traverse the distance between oneself

and a target mind. This is the explanation suggested by the Mind Space framework, as discussed in Chapter 1 (Conway & Bird, 2018; Conway, Catmur, & Bird, 2019; Conway et al., 2020). This framework proposes that people represent the minds of others as points in a high-dimensional space (such as the dimensions used to quantify facial representations in the face recognition literature, or distributed semantic representations in the computational linguistics literature; Kommers et al., 2015). A key claim of the Mind Space framework is that the representation of minds in this high-dimensional space is related to one's ability to make inferences about the specific mental states generated by the mind in question. At a first approximation: if you have a more accurate Mind Space location for a given mind, you are better equipped to make accurate inferences about its mental states (Conway et al., 2020). Supporting this line of argument, evidence suggests that people understand actions by reducing a high-dimensional space of possible action-relevant dimensions down to a low-dimensional space of socially-relevant dimensions (Thornton & Tamir, 2019).

The second possibility is that individual differences in theory of mind are due not to capacity but motivation. This is consistent with evidence, which I have reviewed in Chapters 1 and 5, about effort costs associated with empathy (Zaki, 2014; Keysers & Gazzola, 2014; Cameron et al., 2018; Ferguson & Inzlicht, 2022). In the empathy selection task (Cameron et al., 2018), participants find it more effortful to perform an empathy-based task over an equally difficult control task, and therefore are less likely to choose to engage in empathy. However, these two possible accounts (capacity and motivation) are not mutually exclusive. It is likely that, depending on the circumstances, both can play a role in an individual's mind theoretic ability, since most social-cognitive processes have both a capacity and motivational component (Contreras-Huerta, Pisauro, & Apps, 2020). The question, then, is under what circumstances do

motivation or capacity to perform mental state inferences impact an individual's theory of mind abilities? One line of inquiry, which I will explore in the following section, is that people are worse at making inferences about out-group minds than in-group ones. Few studies have looked at this kind of intergroup bias in the context of theory of mind, and it is not known whether any such intergroup biases might be due to capacity or motivation.

Evidence for attenuated theory of mind abilities in an intergroup context.

Are people worse at making inferences about the mental states of out-group members? While few studies have looked at this question directly, existing evidence supports the possibilities of intergroup bias in both motivation and capacity to engage in theory of mind toward one's out-group. For the motivational component, one reason to suppose that people are less motivated to engage in theory of mind toward out-group members is that people are consistently less likely to engage in prosocial actions general toward out-group members—for example, as evidenced by the studies in Chapter 2 and 4. Understanding an individual's mental states has been shown to facilitate appropriate prosocial actions towards them (Gaesser, Keeler, & Young, 2018), and people show significant variation in their ability to track reward-related processes in others, such as motivation (Apps, Rushworth, & Chang, 2016). Under this account, intergroup bias in theory of mind could exist because people are simply less motivated to understand out-group minds.

Previous studies have investigated a similar account of attenuated motivation for empathy in intergroup contexts (Krebs, 1975; Cikara, Bruneau, & Saxe, 2011; Montalan et al., 2012; Eres & Molenberghs, 2013; Berry & Brown, 2017). Early work on the connection between prosociality and empathy suggested that the more someone was able to intuit another person's

mental state (for example, by being in a shared social group) the more likely they would be to help that other person (Krebs, 1975). Though perhaps motivated by similar premises, these studies have limited applicability to any potential mechanisms of intergroup theory of mind. Most such studies are concerned with negative behaviour in intergroup scenarios and seek to use empathic concern as a mediator of when people act more or less negatively toward out-group members. In their influential account, Cikara, Bruneau, and Saxe (2011) claim that empathic responses are "dampened" in response to out-group suffering, and this dampening works to make the presence of out-group pain less aversive than in-group pain. In line with the long history of intergroup contact theory (Allport, 1954; Everett & Onu, 2013), most research in this vein explores the conditions under which people are likely to take an active role in adopting or appreciating the affective states of out-group members—a kind of affect-augmented contact. It does appear that mutual understanding—not just of affective states, but of cognitive states—can an important be an important part of effective intergroup contact (Livingstone, Fernández Rodríguez, & Rothers, 2019; Livingstone et al., 2022). Indeed the very idea of a stereotype consists in reducing a complex, multifaceted person or group of people to a simpler representation (e.g., Harris & Fiske, 2006). As Harris and Fiske claim, viewing people solely through the lens of a stereotype can culminates in a kind of dehumanisation, getting further away from an intentional agent whose beliefs and motivations we might seek to understand. More closely related to mental state inference, some recent evidence suggests that people are more likely to use mental state terms to describe in- vs out-group members (McClung & Reicher, 2018), that bilinguals outperform monolinguals in the director task (Navarro & Conway, 2021), and that engagement in theory of mind helped to facilitate prosocial actions (Gaesser, Keeler, &

Young, 2018). However, to my knowledge, no existing experimental studies address the specific question of whether people are less motivated to make inferences about out-group minds.

It may also be the case that inferences about out-group minds may in fact be intrinsically more difficult to make than inferences about in-group minds. This would represent an account of intergroup bias in theory of mind due to capacity: it may require greater capacity for making mind theoretic inferences in order to successful interpret the mental states of out-group members. In the above section, I argued that one reason to suspect capacity limitations might be relevant to differences in theory of mind ability comes from the Mind Space framework (Conway & Bird, 2018; Conway, Catmur, & Bird, 2019; Conway et al., 2020). This framework further suggests why out-group minds might pose intrinsically harder inference problems: out-group minds are further away in mind space, and people are therefore less competent at intuiting how mental states arise from a given out-group mind. Empirical evidence suggests that people represent their own mental states with greater nuance than those of others (Thornton et al., 2019). As a consequence, it is reasonable to hypothesise that people are better able to apply this nuanced understanding to mental states that more closely resemble their own. This hypothesis is supported by a separate line of empirical evidence which also suggests that people tend to perceive the minds of in-group members as more heterogenous than out-group members (Linville, Fischer, & Salovey, 1983; Konovalova & Le Mens, 2020). In mind space terms, this suggests that one's in-group members are (or at least are perceived to be) closer to oneself. If inferential difficulty corresponds to distance in mind space, then it follows that there will be a higher effort cost associated with inferences about out-group minds. For example, imagine a Christian observing the action of another person and making inferences about their underlying beliefs and desires. Depending on the situation, the Christian may be better able to make accurate

inferences about a Christian than a Muslim target agent. Under this framework, the explanation would be that minds that are further apart (i.e., out-groups) are more difficult to understand. It is an explanation based on capacity to perform a given mental state inference.

What kind of out-group mental states might be most difficult to infer?

For both possible explanations (capacity or motivation), it should be noted that the kinds of operationalisations used by the present studies and most of the cited literature are representative of only the most simplistic forms of mental state inference in which humans engage. Of the individual studies I review in Chapter 5, each one essentially relies on a point estimate of another individual's beliefs—for example, when they are reaching toward a cup or following the path of another agent (though, for a countervailing experimental approach, using experience sampling, in the context of affective empathy, see Depow, Francis, & Inzlicht, 2021). At a broader scale, the mechanisms of mental state inference must also include much more ambiguous, abstract, or complicated forms of reasoning and deliberation; for example, this is captured in the "strange stories" paradigm (Happé, 1994) and its visual counterpart (Dziobek et al., 2006) mentioned in the previous chapter. There is not a specific belief-to-action mapping, but a morass of behaviour and an arbitrarily large set of explanatory beliefs and desires one could posit to explain any identifiable behavioural patterns (Phillips, Morris, & Cushman, 2019).

In Chapter 1, I defined theory of mind according to the formal inverse planning model developed by Baker, Saxe, and Tenenbaum (2009). Consistent with usage in the literature on theory of mind, the primary mechanisms of that model are dedicated to inferences having to do with situationally-relevant cues: inferring an agent's unobservable beliefs and desires based on their observable actions. Almost all theory of mind tasks focus on the specifics of a situation—

for example, the Director Task is similarly based on inferences about an agent's knowledge or perspective based on what they see within their immediate environment (Keysar, Lin, & Barr, 2003). However, if one does hypothesise that there are potentially capacity-based differences in inferences about out-group minds, it seems unlikely that they would be due to low-level processes such as not being able to see an object behind an occlusion. Rather, these differences would be due to the fact that an out-group member may have vastly different beliefs, preferences, desires, values, morals, or social norms than one's own. In Baker and colleague's model, this is referred to as "world knowledge." The primary function of this node in the model is to make sure the agents can share baseline assumptions, such as approaching an object is evidence of preference for it. It seems likely that any intergroup differences in mind theoretic ability would be due to these more complicated kinds of inferences—ultimately relying on sophisticated inferences about an agent's potential differences in world knowledge. The present set of studies is intended to represent an initial bridge between the current scope of highly situationally bound inferences and the more complex world knowledge inferences which may prove most salient for intergroup bias.

Hypotheses and summary of studies.

My hypothesis is that people are worse at mental inference for out-group members. This could be because of capacity differences (i.e., inferences about out-group minds represent legitimately harder inference problems), or it could be because of motivational differences (i.e., people are less motivated to make inferences in out-group minds). In this chapter, I will present three studies. The first is designed to test whether the novel theory of mind stimuli presented in the previous chapter can be used to demonstrate standard signs of effort aversion. The first study

uses the original (non-intergroup, self-vs-other) Prosocial Effort Task (Lockwood et al., 2018), but the effort operationalisation uses cognitive effort (via mental state inferences using the stimuli developed in Chapter 5) instead of physical effort. In two follow-up studies, I explored this effect in an intergroup context: do people find inferences about out-group minds to be more effortful? The first intergroup study relied on the same structure as the accuracy assessments in Studies 5.1 and 5.2. The second intergroup study used a procedure more closely related to the Prosocial Effort Task. I conclude with a discussion of the novel finding of effort aversion in a theory of mind task, as well as the limitations of my current paradigm to answer the larger question posed about potential intergroup bias in theory of mind.

Study 6.1: Prosocial Effort Task with Theory of Mind Stimuli

I designed this study as a mental state inference analogue of the original Prosocial Effort Task (Lockwood et al., 2017; Lockwood et al., 2021). Whereas most versions of the Prosocial Effort Task in this thesis presented participants with choices about earning points for themselves or one of two other agents, the present version looks only at decisions for self vs other. This is because the present study is not concerned with *prosocial* motivation per se, but rather whether a the novel theory of mind stimuli presented in the Chapter 5 are able to elicit standard effects of effort aversion. With that exception, the basic decision-making structure is the same as other versions presented throughout this thesis. Participants could choose across many trials to perform a task of a certain effort level in exchange for a given reward level, with the effort task operationalised as mental state inference rather than physical exertion (Figure 1).

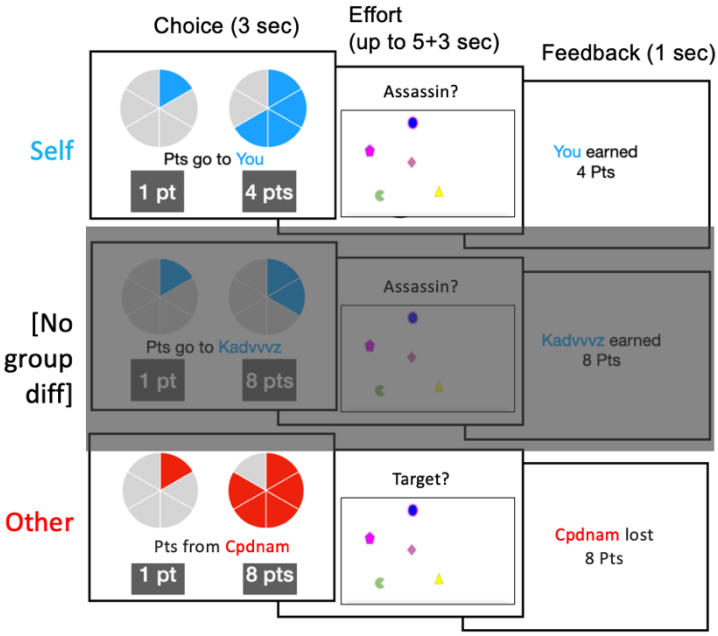


Figure 1 — Original self vs other Prosocial Effort Task with theory-of-mind-based cognitive effort manipulation. The task structure was identical to the intergroup and similarity-manipulated versions of the Prosocial Effort Task presented in previous chapters. Each trial offered an exchange of a given level of effort in order to earn points away from a target agent. Participants were presented with the opportunity to earn points for themselves or another (non-group affiliated) agent.

Participants. Thirty participants were recruited on the Prolific subject recruitment platform (12 female; mean age = 24.5). All participants provided written informed consent. The study was approved by the local Ethics Committee. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately forty-five minutes to complete.

Method.

Design. As much of the structure of the self-vs-other Prosocial Effort Task (Lockwood et al., 2017) was retained in this paradigm to maintain similar decision dynamics. On each trial,

participants made decisions to perform a lower-effort option in exchange for one point or a higher-effort option in exchange for more points. The exact amount of effort required (five effort levels: 2, 3, 4, 5, 6) and the number of points on offer (four reward levels: 2, 4, 6, 8) varied on each trial, as did whether the points would go to the participant or another agent. I reduced the number of reward levels from five to four in order to reduce the number of times each clip would be shown to a participant. Participants were told that the effort levels represented difficulty ratings from a previous group of participants. On each trial, effort levels were indicated by a pie chart, as in the physical effort version of the task.

Thirty-one theory of mind stimuli video clips were used for the higher-effort option. I selected them based on a threshold for participants' accuracy scores in the assessment of the stimuli from the previous chapter (Study 5.1). Clips which had 75% accuracy or higher were included. Both prompts—identifying the assassin or the target—were included, as long as participants answered the relevant question with sufficient accuracy. (Note that on a single trial of the effort task participants would be shown only one of the two prompts.) Despite participants' consistently high accuracy in selecting the correct answer, these clips elicited a range of self-reported perceived difficulty. The clips were segmented into five effort levels with the following ranges of effort scores, as determined by Study 5.1: below 25 (five clips), between 25 and 35 (seven clips), between 35 and 45 (seven clips), between 45 and 55 (seven clips), and above 55 (five clips). Accordingly, if a participant chose the higher-effort option on every trial, they would see each clip a total of eight times (once for each of the four reward levels, in both the self and other conditions). Seven of the clips featured both assassin and target prompts—which means participants would have viewed these clips up to sixteen times.

Whereas in the original Prosocial Effort Task, the lower-effort option consisted of no exertion of physical effort (e.g., no button presses), in this task the lower-effort option presented participants with one of ten video clips. These clips featured the lowest self-reported difficulty in the stimuli validation, all of which were among the highest accuracy scores (clips that were included in this lower-effort set were not also included in the higher-set described above). I chose to present clips in the lower-effort option, because otherwise participants may have chosen the higher-effort option not because of any effort trade-off but because they preferred to do something rather than nothing (Wu, Ferguson, & Inzlicht, 2021). Without any low-effort task to perform, participants who chose the lower-effort option would face a blank screen for up to nine seconds. The ten lower-effort clips were pseudo-randomised across all trials, so that on any given trial it was equiprobable any one of them would be shown to a participant choosing the lower-effort option. If participants failed to make a choice between higher- and lower-effort options within the three second window, they received no points and were presented with nine seconds of a fixation cross to await the onset of the next trial.

Once the participant had selected between the higher- and lower-effort options, they were immediately shown the five second video clip. The clip was shown in conjunction with the prompt—asking the participant to identify either the assassin or the target—as well as a button featuring each of the five players on screen. Participants could select their answer for the target/assassin at any time throughout the video. After the video finished, participants were given an additional three seconds to provide their answer—with a countdown clock indicating the final three seconds. Participants were then given feedback on the number of points earned. Participants earned no points if they failed to answer or gave an incorrect answer.

In total, participants completed 248 trials (31 clips, in each of four reward levels, for self and other). These trials were split into three blocks, between which participants were given the opportunity to rest for up to two minutes. The distribution of clips was counterbalanced between blocks and randomised within them—that is: participants were always shown the same clips within each block, but in a different order. The initial allocation of clips into different blocks was randomised.

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Prolific subject recruitment platform.

Questionnaire. Immediately after providing consent, each participant completed the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011).

Post-task self-report judgments. After completing the main experiment, each participant was asked to use a sliding scale to report how "interesting" they found the task, how "difficult" they found it, and how often they felt they forgot their answers before answering the questions.

Attention checks. Participants were asked at the end of the experiment to report the extent to which they paid attention during the entirety of the experiment—both in a slider and a verbal description. They were ensured this report would not affect their payment. There were also three catch-trials during the main task, in which text appeared in the video instructing the participant to disregard the movement of the shapes and instead select a specific shape as the assassin in the subsequent prompt.

Results.

Participants were more averse to the higher-effort theory of mind clips.

To test whether participants were more likely to choose the higher-effort option when earning points for themselves versus others, I ran a linear mixed-effects regression with choice probability as an outcome variable (the percentage of higher-effort choices made by each participant) and agent as a predictor variable (self vs other), including a subject-level random intercept (Figure 2). The model yielded a significant difference between self and other choice probabilities ($b = 0.098$ [0.045 – 0.152], $z = 3.2585$, $p < 0.001$).

I also tested whether participants were more likely to choose clips from the lowest effort level than the highest effort level. This was done using a linear mixed-effects regression with choice probability as an outcome variable and effort level as a predictor variable (lowest versus highest), including a subject-level random intercept (Figure 3A). The model, again, yielded a

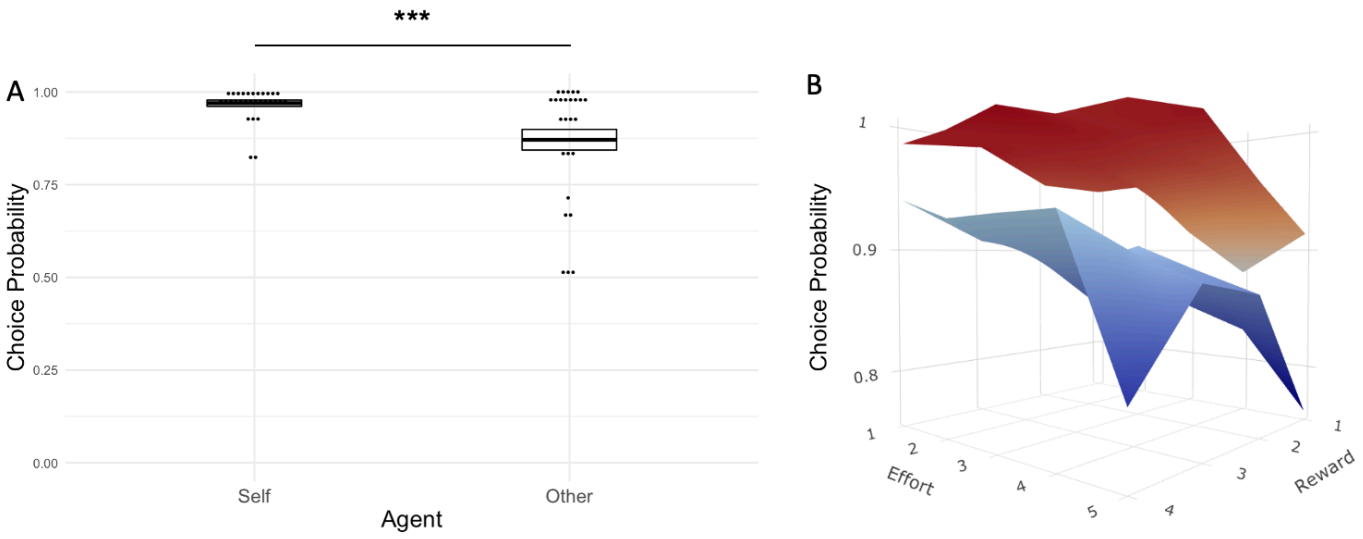


Figure 2 — Difference in choice probability of higher-effort when choosing for self versus other. (**A**) Participants were significantly more likely to choose the higher-effort option when earning points for themselves than others. (**B**) Choice probability for self (red) and other (blue) across effort and reward levels.

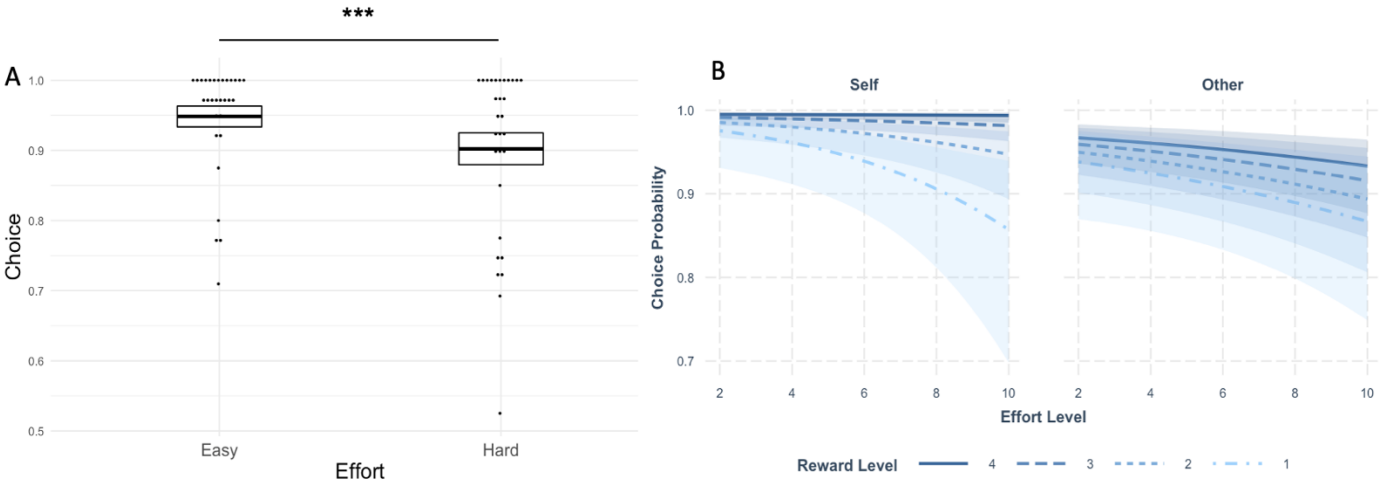


Figure 3 — Participants were more averse to higher-effort theory of mind stimuli. (**A**) Participants were significantly more likely to choose the higher-effort option when presented with a clip from the lowest effort level than the highest. (**B**) General linear model showing trial-by-trial choice probability as a function of agent, effort, and reward levels.

significant difference between self and other choice probabilities (b = -0.046 [-0.081 – 0.012], z = -2.613, p = 0.009).

This pattern of results was consistent with the results of a generalised linear mixed-effects model, designed to test for possible interactions (Figure 3B). The model showed significant main effects of agent (self vs other; X^2 = 148.38, p < 0.001), effort (X^2 = 25.99, p < 0.001), and reward (X^2 = 52.35, p < 0.001). There was also a significant agent-by-reward interaction (X^2 = 23.29, p < 0.001), but no interaction between effort and either agent (X^2 = 0.03, p = 0.865) or reward (X^2 = 0.84, p = 0.359).

This aversion was due to variation in perceived effort, not variation in accuracy.

Figure 4 shows average accuracy as a function of effort levels. Accuracy across clips did not monotonically decrease as effort levels increased. With respect to the first effort level,

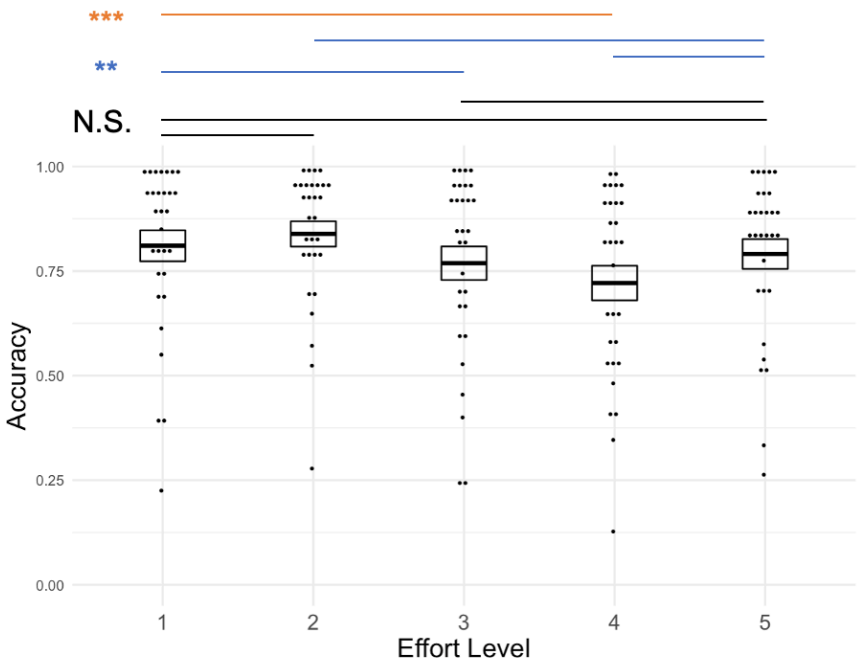


Figure 4 — Clip accuracy did not vary systematically as a function of increasing effort levels. Across the five effort levels, higher effort levels did not yield lower accuracy.

according to a linear mixed-effects regression there was no significant difference with effort levels two (b = 0.028 [-0.007 – 0.063], z = 1.590, p = 0.112) or effort level five (b = 0.020 [-0.055 – 0.015], z = 1.098, p = 0.272) while there was a significant difference with effort levels three (b = -0.042 [-0.077 – -0.007], z = -2.335, p = 0.020) and effort level four (b = -0.089 [-0.124 – -0.054], z = -4.992, p < 0.001). With respect to the fifth effort level, there was no significant difference with effort level three (b = 0.022 [-0.024 – 0.068], z = 0.949, p = 0.343) while there was a significant difference with effort levels two (b = -0.048 [-0.082 – -0.014], z = -2.756, p = 0.006) and effort level four (b = 0.069 [0.024 – 0.115], z = 2.991, p = 0.003).

Discussion.

In this study, I showed that participants were more likely to avoid the more effortful version of the theory of mind task. This effect was present both when obtaining rewards for self

and other. This pattern of results suggests a similar pattern of effort discounting was observed for cognitive effort as observed for physical effort. Specifically, participants were significantly more likely to choose to perform easier mental state inferences than harder ones. To my knowledge, this is the first direct demonstration of effort aversion in a theory of mind paradigm. One possible counter-explanation was ruled out: participants could have been averse to doing the "harder" stimuli not because they required more effort, but because they were more likely to get them wrong. However, I showed that there was no systematic variation in accuracy across the different stimulus effort levels. This suggests that the effects are due primarily to the perceived difficulty of the stimuli in the task and not participants' capacity to perform the required inference.

Study 6.2: Theory of Mind Assessment with "Intergroup" Stimuli

This study was designed to extend the theory of mind assessment paradigm used in Chapter 5 into an intergroup paradigm. The central question of this study was to test whether participants' accuracy on these theory of mind stimuli could be modulated through a minimal groups paradigm: were participants less accurate when they believed a mind belonged to an out-group member?

Participants. Forty participants were recruited on the Prolific subject recruitment platform (14 female; mean age = 30.38). All participants provided written informed consent. The study was approved by the local Ethics Committee and National Health Service Ethics. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately forty-five minutes to complete.

Method.

Design. In each trial, participants were shown one of the five-second theory of mind stimuli video clips developed in Chapter 5. On the following screen, they were asked to identify the assassin. Then they were asked to report how difficult it was to "figure out the assassin" on a sliding scale of 0-100 from "very easy" to "very hard." During the presentation of the initial video, participants were also informed whether the assassin was a member of their in- or out-group.

Each participant completed 82 trials. Each trial featured a unique video, drawing on a subset of the original 104 which removed any clips for which there was an average accuracy of less than 10% accuracy or over 90% so that floor/ceiling effects did not mask any intergroup effect. The presentation of videos was counterbalanced so that both a within-subjects and between-subjects analysis was possible. For any participant, they saw half of the videos featuring "in-group" assassin and half of the videos as "out-group" assassins. Ideally, participants would be presented with clips offering the same distribution of accuracy scores and difficulty ratings for in-group and out-group condition. In the ideal case, this would indicate that any difference in performance between in-group and out-group trials would then reflect only effects of group membership. To achieve this counterbalance, I arranged the 82 clips in order of accuracy scores. Clips with the same accuracy scores were ordered by differences in difficulty ratings. Videos with an odd-numbered ordering were assigned to Group A, and even-numbered order videos were assigned to Group B, with half of the participants assigned Group A as the in-group and Group B the out-group, and the other half with the reverse assignment. Thus, a within-subjects analysis could be performed by comparing in-group versus out-group video clips of matched

accuracy and difficulty levels; and a between-subject analysis could be performed by comparing the same clip across in-group and out-group assignments.

Group induction. The group induction was identical to the one used in other intergroup tasks from previous chapters (described in Study 2.2 and 4.2). The only modification was the connection between the group induction and the main task. In line with the previous group inductions used in the intergroup studies throughout this thesis, participants learned the names of five in-group and five out-group members. Then participants were told that on any given clip, the shape assigned to be the assassin would be controlled by one of these ten individuals. They were told that while the shapes stay the same from trial-to-trial, the players controlling the trials changed—so the assassin, whether an in-group or out-group member, could be any one of the shapes. To indicate group membership, the background of the screen (behind the clip) turned red or blue and the participant was shown the prompt "The assassin is on YOUR/the OTHER team."

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Testable Minds participant recruitment platform.

Questionnaire. Immediately after providing consent, each participant completed the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011).

Post-task self-report judgments. After completing the main experiment, each participant was asked to report how similar they felt to members of their own team and how similar they felt to members of the other team (from "not similar at all" to "completely similar"). They were also asked how strongly they felt a part of their group (from "not strongly at all" to "very strongly"). Participants were also asked whether they felt they did a better job identifying the assassin for their team or the other team (on a sliding scale from "other team" to "my team" with instructions to leave the slider in the middle for no difference). I also asked participants: "To what extent do

you agree with the following statement: “It’s important to understand the perspectives of people from different backgrounds, because it helps you understand what they want and how to interact with them in the real world?’” Finally participants were asked how interesting and difficult they found the task.

Attention checks. Participants were asked at the end of the experiment to report the extent to which they paid attention during the entirety of the experiment—both in a slider and a verbal description. They were ensured this report would not affect their payment. There were also three catch-trials during the main task, in which text appeared in the video instructing the participant to disregard the movement of the shapes and instead select a specific shape as the assassin in the subsequent prompt.

Results.

I analysed these data using an identical suite of tests as in the initial theory of mind stimuli assessments of Chapter 5—with the only exceptions being the addition of group, with levels of in-group and out-group.

Participants did not vary in theory of mind accuracy as a function of group membership.

The between-subject analysis compared each clip’s percentage correct based on whether they were presented to participants as in-group or out-group minds. I ran a linear mixed-effects regression with per-clip accuracy as an outcome variable (the percentage of participants who got a particular video clip correct) and agent as a predictor variable (in-group, out-group), including a subject-level random intercept (Figure 5). The model yielded an non-significant difference

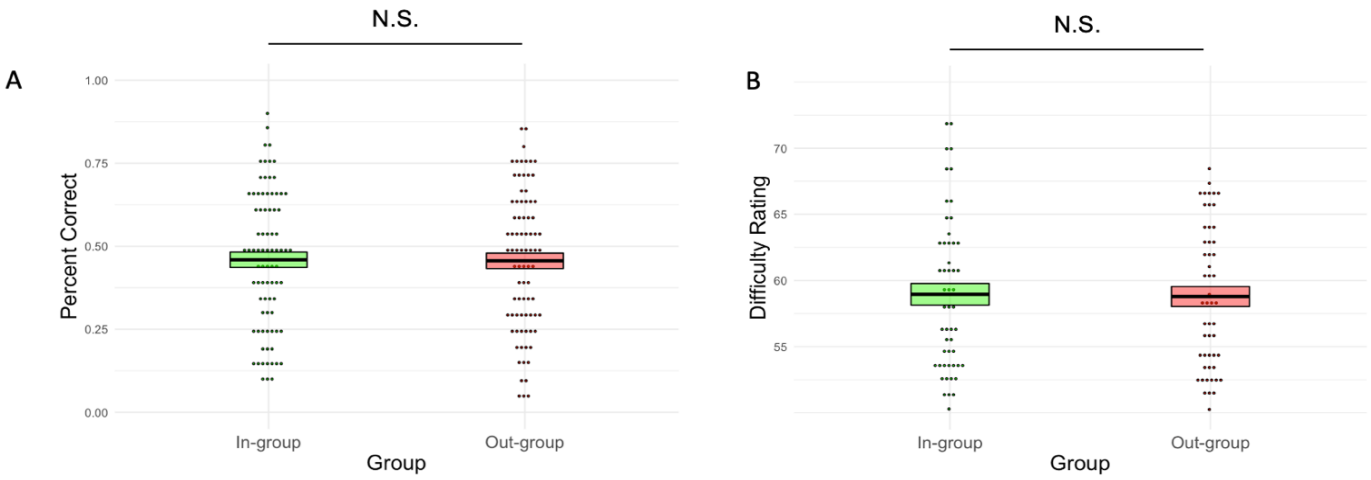


Figure 5 — No between-subjects difference in inferential accuracy. (A) Participants showed no systematic difference in their accuracy of in-group versus out-group clips. (B) Participants did not rate in-group clips as more difficult than out-group clips.

between in-group and out-group accuracy (b = -0.003 [-0.032 – 0.026], z = -0.222, p = 0.824; Figure 5A). Neither was there a significant difference in between-subject difficulty ratings (b = -0.908 [-2.795 – 0.978], z = -0.944, p = 0.345; Figure 5B).

Nor did they vary in theory of mind difficulty as a function of group membership.

The within-subject analysis compared videos of matched accuracy and difficulty presented as either in-group or out-group minds. I ran a linear mixed-effects regression with per-subject accuracy as an outcome variable (the percentage of clips a participant got correct) and agent as a predictor variable (in-group vs out-group), including a subject-level random intercept (Figure 6). The model showed no difference in either accuracy (b = -0.004 [-0.036 – 0.029], z = -0.213, p = 0.832; Figure 6A) or difficulty (b = -0.929 [-2.681 – 0.822], z = -1.040, p = 0.298; Figure 6B).

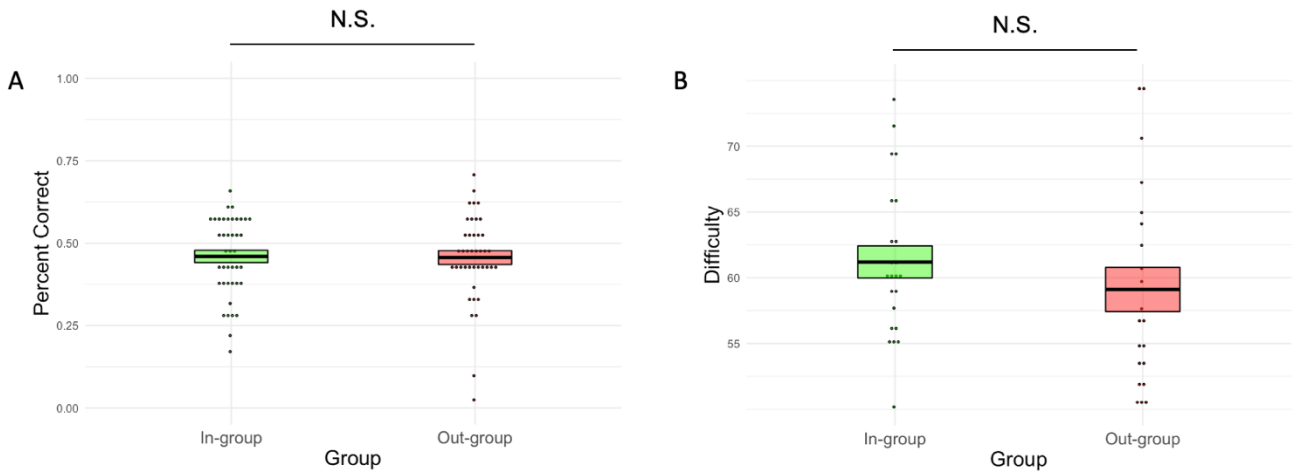


Figure 6 — No within-subjects difference in inferential accuracy or difficulty. (A) Accuracy scores did not differ for in-group vs out-group clips. (B) Difficulty ratings did not differ for in-group vs out-group clips.

This was not explained by variation in participants' belief in the group manipulation.

One possible explanation for these null results was that some participants believed more strongly in the group manipulation than others. Perhaps in participants for whom the group "buy-in" was stronger, there may have been significant differences. Figure 7 shows the correlation between intergroup accuracy (how much more accurate a participant was for in-group videos than out-group videos) and self-reported group "buy-in" (as assessed by a debrief question asking the extent to which participants felt strongly to be a member of their group). The correlation between these measures was not significant ($r = 0.056$, $p = 0.73$).

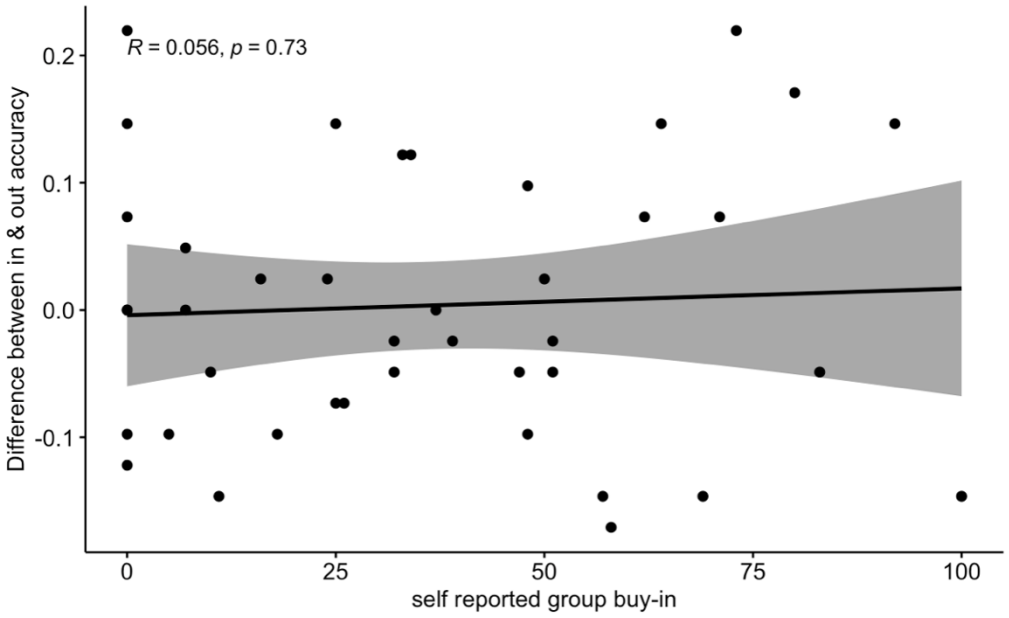


Figure 7 — No correlation between intergroup accuracy and self-reported strength of "buy in" to group manipulation.

Discussion.

No intergroup effects were found in these data. One issue with this paradigm that seems clear in retrospect is the ambiguous correspondence between the theory of mind inferences and the intergroup dynamics. Participants were told that the assassin would be a member of their in- or out-group. But there were no visual cues to distinguish in- from out-group members. It is also unclear why an in-group assassin would behave differently than an out-group one. Perhaps a particular dynamic could be inferred if there were one in-group assassin and many out-group potential targets—or some other permutation of these instructions. But as presented, information about social groups clearly did not factor into participants' responses.

Another potential issue is with the high test-retest reliably of the stimuli clips demonstrated in Study 5.2. If these clips attain more than 95% test-retest reliability, as the Chapter 5 results suggest, then it seems possible they are insufficiently ambiguous to elicit

significant deviations in accuracy or perceived effort with the addition of verbal instruction about group-level information—even if the instruction itself had been more compelling.

With these considerations in mind, a possible revision to my motivating hypothesis—that people tend to be worse at mental state inference for out-group members—based on these data is that people are not less accurate when it comes to out-group theory of mind. Instead, they find it more effortful. This account would align with the preliminary correlations presented in Study 5.2, which showed that the cognitive empathy component of the QCAE tracked perceived difficulty, but not accuracy, in the theory of mind stimuli. In the following experiment, I sought to more directly test this revised hypothesis, which the present stimuli may be better suited to addressing.

Study 6.3: Prosocial Effort Task with "Intergroup" Theory of Mind Stimuli

Throughout this thesis, a key distinction has been made between the capacity and the motivation to engage in a prosocial action (Contreras-Huerta, Pisauro, & Apps, 2020). Attenuation of either can lead to an observed decrease in prosocial action. In the previous study (6.2), the question I investigated was related to capacity: Are people less accurate when making inferences about out-group minds? With the experimental paradigm and stimuli used, there was no evidence to support this hypothesis. However, the first study in this chapter (6.1) showed significant effects of effort aversion using these stimuli. Thus, for this final experiment I sought to maintain the structure of that experimental paradigm as closely as possible, with the addition of an intergroup induction to test whether there might be a differential effect of effort aversion for in- vs out-group agents. Thus, the experiment was designed to determine whether people are less willing to exert effort to make mental state inferences about out-group than in-group minds?

In particular, I expected that any observed effects of group membership on effort aversion in this paradigm would manifest in one specific contrast. For the hardest clips, offering the lowest points, when those points go to the other agent: is there a difference between in- and out-group clips? These are the conditions under which the previous effort aversion effects have been strongest. Therefore if any intergroup effect on effort aversion is present then it will most likely be present under these conditions. If any such difference between choice probability for in- vs out-group clips is observed, it would imply that participants are unwilling to exert effort solely based on aversion to making inferences about the mental states of perceived out-group members.

Participants. Twenty participants were recruited on the Prolific subject recruitment platform (4 female; mean age = 32.9). All participants provided written informed consent. The study was approved by the local Ethics Committee. Participants were paid at a rate of £7.50 per hour and told they would receive up to a £1 bonus based on their performance. Performances earning greater than 0 points in the main task received the full £1 bonus. The full study took approximately forty-five minutes to complete.

Method.

Minimal Groups Induction. Participants underwent the same minimal groups induction as described in Study 6.2.

Design. The design for this experiment was identical to the same self-vs-other Prosocial Effort Task structure used in Study 6.1—with the exception of the additional minimal group induction and relevant changes to the instructions.

Participants were told that they were matched with a random player from a previous version of the experiment (this is the "other" agent for whom they would potentially be earning

points). The same structure of five effort and four reward levels were used, as described above. Decision time for the initial choice (between high and low effort) was increased by 500ms (to 3500ms) to accommodate the additional information of group membership. Overall, in each trial participants were told: effort level (pie chart, out of five), reward level (number of points), the agent to whom the points would go (self or other), and whether the clip featured a cohort of their in- or out-group (i.e., whether the following clip would show players from the red or blue team).

The same stimuli and effort levels were used as described in Study 6.1. Thirty-one theory of mind stimuli video clips were used for the higher-effort option. I selected them based on participant accuracy in the initial validation of the stimuli; clips which had 75% accuracy or higher were included. The inclusion threshold was made based on accuracy, because clips that had less than 75% accuracy on average would be aversive to participants not because they are perceived as difficult but because they are less likely to get them correct and earn the associated reward. Participants completed a total of 248 trials—31 clips, in each of four reward levels, for self and other—across three blocks.

Colour-matched stimuli. In the previous intergroup study with my novel theory of mind stimuli (6.2) participants were told that some players were in-group members and some were out-group members. However, there were no visual cues as to which player belonged to which group. This may have contributed to the lack of salience of the intergroup manipulation in Study 6.2. In this version, I sought to address that possible issue by altering the stimuli. In the original stimuli, the shapes were all different colours. I had the original videos edited to create two sets of stimuli: one in which all the shapes on the screen were red, and one in which all the shapes were blue. Accordingly, in the present version of the experiment, participants viewed clips where the colour of the shapes corresponded with that player's group membership (e.g., if a clip featured

in-group members for a participant on the red "team" then the following clip would feature the same shapes moving in the same trajectories as the original clip, but all in red).

Apparatus. Stimuli were programmed using the Gorilla experiment builder. Data were collected via the Testable Minds subject recruitment platform.

Questionnaire. Immediately after providing consent, each participant completed the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011).

Post-task self-report judgments. After completing the main experiment, each participant was asked to use a sliding scale to report how "interesting" they found the task, how "difficult" they found it, and how often they felt they forgot their answers before answering the questions.

Attention checks. Participants were asked at the end of the experiment to report the extent to which they paid attention during the entirety of the experiment—both in a slider and a verbal description. They were ensured this report would not affect their payment. There were also three catch-trials during the main task, in which text appeared in the video instructing the participant to disregard the movement of the shapes and instead select a specific shape as the assassin in the subsequent prompt.

Results.

I analysed these data with an identical suite of analyses to Study 6.1—with the only difference due to the additional levels of in-group and out-group agents (instead of self).

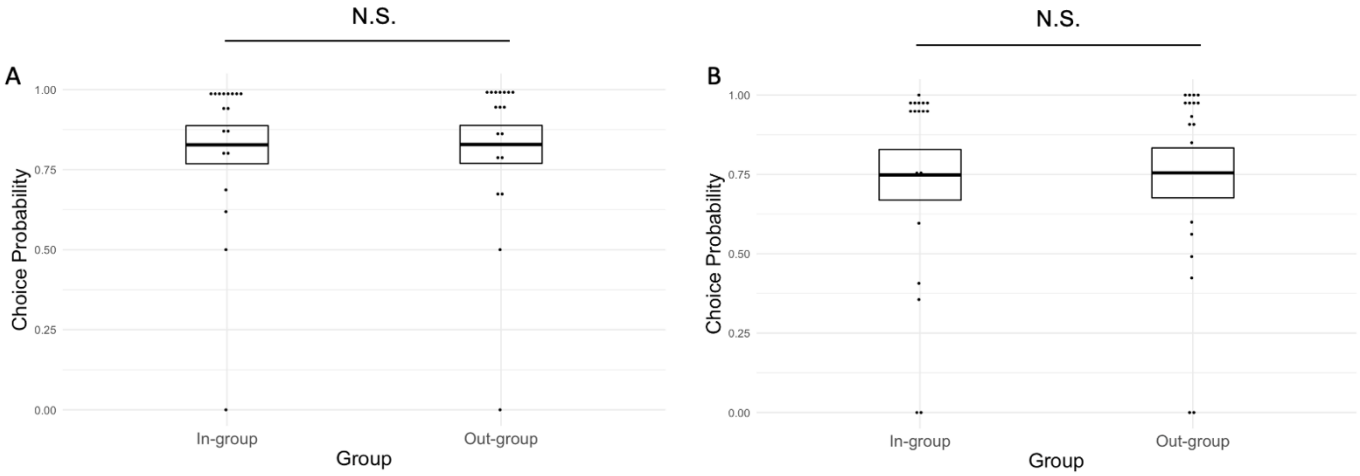


Figure 8 — No observed intergroup effects of effort aversion. Participants' probability of choosing the high effort for in-groups vs out-group clips did not significant differ. (A) shows participant choices for both self and other; while (B) shows only choices for others.

Participants did not choose differently based on whether clips were generated by in- or out-group members.

To test whether participants were less likely to choose the higher-effort option when presented with out- than in-group clips, I ran a linear mixed-effects regression with choice probability as an outcome variable (the percentage of higher-effort choices made by each participant) and group featured in the clip as a predictor variable (in-group vs out-group), including a subject-level random intercept (Figure 8). The model yielded no significant difference between in- and out-group choice probabilities (b = 0.001 [-0.012 – 0.013], z = 0.146, p = 0.884; Figure 8A). To some extent this result is expected, as in previous versions of this paradigm participants tended to choose the high-effort option when the points go to themselves. However, even when looking only at trials on which points went to the other agent, the model yielded no significant difference between in- and out-group choice probabilities (b = 0.006 [-0.020 – 0.032], z = 0.476, p = 0.634; Figure 8B).

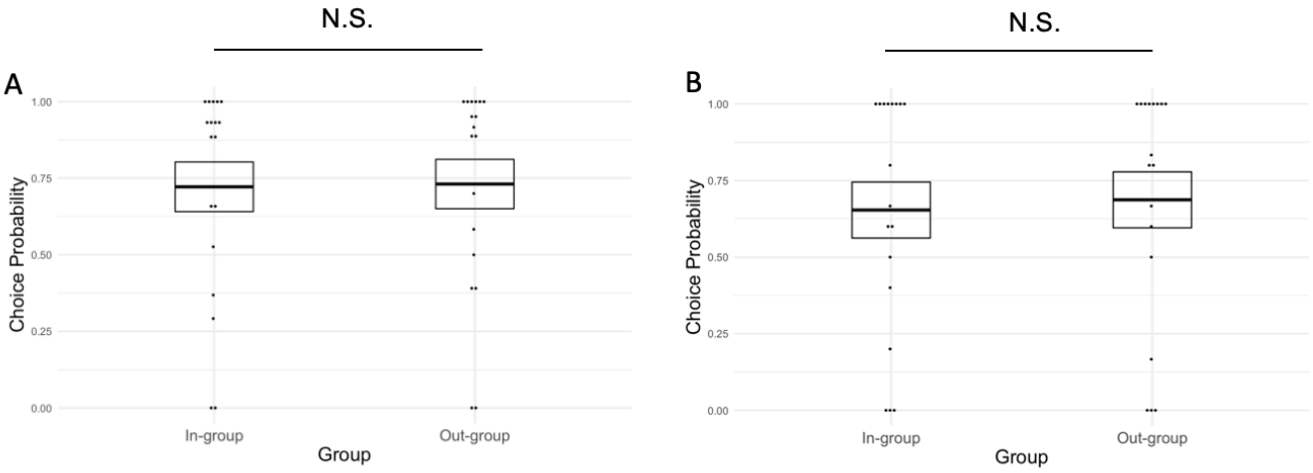


Figure 9 — No observed intergroup effects of effort aversion. (A) Trials with highest two (of five) effort levels when points going to other agent. Similarly, no significant effects were found for the lowest two reward trials. (B) The conditions hypothesised to show strongest effort aversion: high effort, low reward, points to other.

Not even at the highest levels of effort or lowest levels of reward.

Based on previous versions of the task, evidence of effort aversion is most likely to be seen in trials with the highest effort or lowest reward levels. To identify any possible effects at this finer grain of inquiry, I modified the above linear mixed model (predicting overall choice probability differential for the different group conditions) to include relevant subsets of the trials—and only for trials where the points went to the other agent (Figure 9). Yet significant intergroup effects were found neither in trials with the highest two (of five) effort levels (b = 0.006 [-0.020 – 0.032], z = 0.588, p = 0.557; Figure 9A) nor in trials with the lowest two (of four) reward levels (b = 0.018 [-0.019 – 0.055], z = 0.949, p = 0.343). I also modified the model to include trials with the highest two effort levels and lowest two effort levels; this, too, was not significant (b = 0.033 [-0.037 – 0.104], z = 0.925, p = 0.355; Figure 9B).

Not even in the earliest trials.

Finally, one possibility was that participants were learning, eventually, that there was no additional difference in difficulty between in-group and out-group clips. But perhaps participants begin the experiment with an initial bias against out-group clips? Yet even when limiting the linear mixed model to the first block of (80) trials—with high effort, low reward, and points to other—still no significant differences were found (b = 0.040 [-0.163 – 0.249], z = 0.387, p = 0.699; Figure 10A). However, because the strongest effects in the previous section were due not to the combined high-effort, low-reward trials but only the low-reward trials (across all effort levels), I also looked at the first block of trials featuring low-reward, all-effort, and points to the other agent. While still not significant, this was strongest result of all the analyses (b = 0.057 [0.000 – 0.115], z = 1.951, p = 0.051; Figure 10B). However, any possible effect appears to be trending in the opposite direct from the hypothesis, with choice probability increasing for out-group clips.

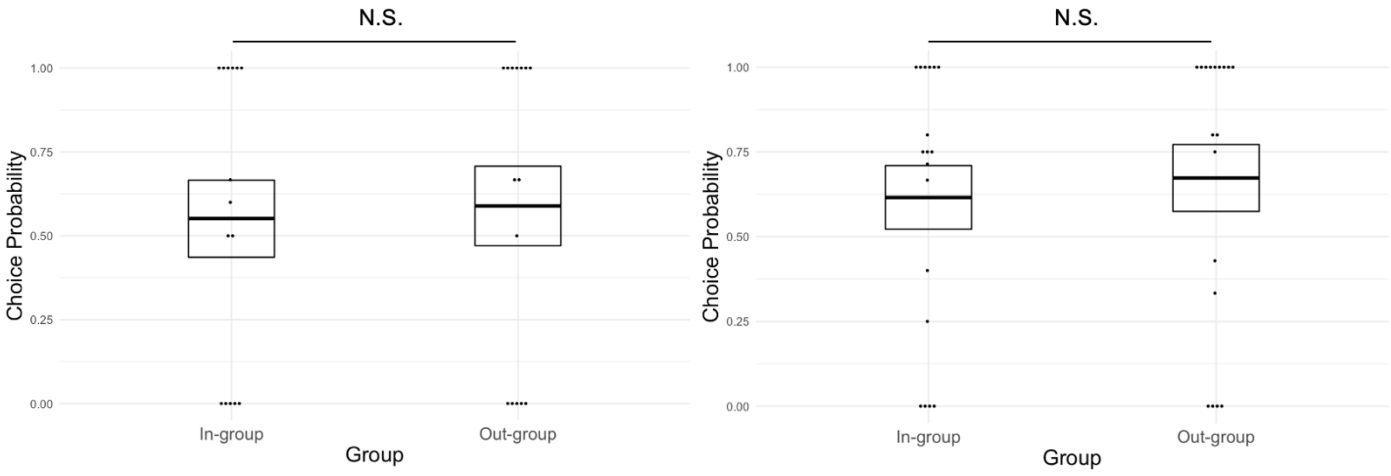


Figure 10 — No observed intergroup effects of effort aversion. Participants did not differ in the probability of choosing the higher effort clips for in-group vs out-group clips. (A) Shows only choices in the first block of trials: high effort, low reward, other agent. (B) Shows the same, but across all effort levels.

Discussion.

The above studies failed to provide evidence that participants found inferences about out-group members more effortful than inferences about in-group minds. Participants showed no difference in accuracy between the clips for different groups.

The initial logic behind this experiment was that if effort aversion effects were found for these stimuli when varying effort and reward levels (as well as the recipient of the points), then perhaps adding additional information about in- or out-group minds might elicit such effort effects as well. One possible point of failure in the previous attempt (Study 6.2), as discussed above, was that there was an unclear connection between the stimuli and in- vs out-group minds. This study addressed that to some degree. A visual cue (colour, as well as an explicit verbal label of in- or out-group) clearly denoted which stimuli were generated by which group. However, one potential issue that remains is that there is a mixed message in relaying information about both a clip's difficulty and which group generated it. If participants are explicitly told how difficult a clip is going to be (e.g., effort level four), are they going to be likely to significantly modulate their expectation of effort based on additional information about group membership? On the one hand, group framing effects were very large in the intergroup versions of the Prosocial Effort Task used in previous studies (particularly those in Chapters 2 and 4). This suggests that perhaps such a modulation is possible. However, based on these results, it seems that what participants were supposed to infer from the group membership of the players in the clips was not sufficiently clear enough to alter their expectation (or perception) about the difficulty of the video clips. And in this respect, this was a rational conclusion for participants to come to—given that there were in fact no actual differences between in- and out-group clips. A secondary issue with the present task is that the majority of participants in this sample are either choosing the high effort option

for the other agent close to 100% of the time or 0% of the time. This leaves almost no variance within which an intergroup effect could be observed. Ultimately, the strength of the Prosocial Effort Task is that it allows for nuanced dissection of a participant's decision-making strategy across many trials; but the effects I hypothesised as presently framed did not elicit a strong enough signal to create an identifiable patter in participants' behaviour.

General Discussion

In the first study from this chapter (Study 6.1), I have shown that people avoid more difficult theory of mind problems when earning rewards for both self and other. Combined with self-reported differences in perceived similarity, this provides strong support for the claim that these novel stimuli represent theory of mind tasks that embody a range of quantifiable difficulty levels. These differences in choices to avoid more difficulty theory of mind problems are not due solely to accuracy, as there is a dissociation between difficulty and accuracy in Study 5.1 and 5.2, and each of the clips used in Study 6.1 was above a threshold of 75% accuracy. Therefore, the differences appear to be due to differences in cognitive effort associated with the inferences required by each stimulus. To my knowledge, no previous theory of mind paradigm is able to quantify such differences in perceived cognitive effort in a task based on mental state inference.

This finding extends existing work on how physical and cognitive effort are processed in the context of a decision-making cost-benefit analysis (Apps, et al., 2015; Kool & Botvinick, 2018). While cognitive effort costs have been found for many different kinds of tasks, none have yet shown such a cost associated with theory of mind. In this task, I showed that there is a cognitive effort cost associated with understanding someone else's mental state, and incentives can motivate an individual to be more willing to overcome the effort cost associated with

inferences about that mental state. One reason this question has proved difficult to answer is that exertion of cognitive effort is harder to quantify precisely than exertion of physical effort (Kool & Botvinick, 2018). For example, in the versions of the Prosocial Effort Task used in the first three empirical chapters of this thesis, each effort level is represented with an exact number of button clicks. Because the physical requirement from the participant is precise, the psychometric responses of how difficult they perceive a given level of exertion to be can be readily tracked as a function of these physical levels. But what makes a given task cognitively effortful is much less straightforward to quantify, because the overall concept of cognitive effort does not easily map onto well-defined psychological concepts, such as working memory load, but instead draws on many distinct cognitive systems (Kool & Botvinick, 2018). For example, perceived difficulty can be increased by either requiring greater speed or greater accuracy in a cognitively effortful task (Manohar et al., 2015). And while recent work has begun to get at questions of cognitive effort in relation to interpreting the minds of others (Cameron et al., 2019; Ferguson, Cameron, & Inzlicht, 2020; Weisz et al., 2020; Ferguson & Inzlicht 2022), most of these studies have attempted to quantify the cognitive costs of empathy rather than theory of mind.

What—if not inferential difficulty—makes some stimuli more difficult than others?

It has proven difficult to say what exactly about theory of mind—or, more generally, any social-cognitive process—might be difficult that is not clearly reducible to another cognitive process such as working memory (Contreras-Heurta, Pisauro, & Apps, 2020). This is one benefit of these stimuli: they empirical vary in their perceived and actual difficulty in a reliable way. Leveraging this empirically-observable aspect of these stimuli allows them to be used to measure effort in theory of mind, without requiring a strong explicit stance as to what exactly the

provenance of the effort is. Cognitive effort avoidance has been shown to be associated with many cognitive processes, such as attention switching (Apps et al., 2015), working memory (Kardan et al, 2020), cognitive control, (Shenhav, Botvinick, & Cohen, 2013), and task switching (Kool et al., 2010)—any of which could potentially contribute to the varied difficulty of these stimuli. The present studies with these novel stimuli represent a first stage in this line of research, showing a basic effect of effort avoidance. However, it should still be noted that the present study does draw a clear distinction between capacity and motivation to engage in theory of mind. Participants clearly have the capacity to correctly answer the majority (if not all) of the stimuli presented to them. As I have shown, there are cases in which the available reward does not justify putting in the effort to determine this correct answer. Future studies will be necessary to disentangle the contributions from various potential sources.

What effort costs might be associated with interferences about out-group minds?

In the Introduction to this chapter, I discussed the role that "world knowledge" from the Baker and colleagues (2009) model might play in making out-group mental state inferences more difficult than in-group ones. One possible limitation is that the present study does not extend quite far enough into the kinds of theory of mind tasks which might be more difficult in an intergroup context. As I have suggested, perhaps out-group minds are truly more difficult to make inferences about when it requires understanding of beliefs, morals, or social norms that one does not share with the target agent. However, even if this was the case, there is should still be the possibility for an effect via differential motivation to engage in out-group versus in-group mental state inferences.

Another consideration is the use of the minimal groups paradigm. Every intergroup context experimentally invoked in this thesis has relied on a minimal groups paradigm. One way of expanding the reach of questions addressable with my theory of mind stimuli is open them up to real-world groups. For example, it is well-established that one important representational space of stereotype content are the axes of warmth and competence (Fiske et al., 2002). Perhaps activating a bias about a stereotyped out-group's competence would modulate participants' perceptions of others' actions—even within the simple world of the assassin game. The trade-off is that while using real-world group might be more likely to yield a significant effect, only minimal-groups allow for empirically supported conclusions about capacity vs motivation in theory of mind. In other words, as soon as real-world groups are introduced, there is a possibility of actual differences in mind-related traits (as discussed in the Introduction to this chapter). Therefore, there may actually be a legitimate sense in which the inferences are in fact harder, which therefore confounds any increase in perceived effort cost. In addition to the consistency of using the same minimal groups paradigm throughout all my experiments, I felt the most important questions to attempt to address for the present work were best suited to minimal group manipulations. I hope to explore variations on the present experiments with real-world groups in future research—which I will outline in greater length in Chapter 7.

What kind of reward is associated with mental state inference?

In the Introduction to this chapter, I surveyed the literature to provide evidence from existing studies to support my hypothesis that people would be worse at mental state inferences about out-group than in-group members, in particular because of increased perceived effort cost (and therefore decreased motivation) associated with doing so. Thus, my first application of the

novel theory of mind stimuli developed in Chapter 5 was to find potential ways of testing this intergroup hypothesis. Both studies (6.2 & 6.3) in which I attempted such an experimental test yielded no evidence of intergroup effects.

One of the successes in research on cognitive effort (and decision-making in general) is the ability to manipulate participants' valuations of rewards (Daw & Shohamy, 2008). To take examples in this thesis, situations that participants deemed worth exerting effort for out-group rewards in the studies from Chapter 2 were used in a way that participants deemed worth exerting effort to take away out-group rewards in the studies from Chapter 4. This sort of goal-directed behaviour is a foundation stone in decision-making research (Sutton & Barto, 1999; Schultz, Dayan, & Montague, 1997). This is a boon for paradigms like the Prosocial Effort Task: the costs are quantifiable, as are the rewards. Therefore, they can both be varied parametrically and the subsequent effects on behaviour can be observed. However, it is unclear the extent to which theory of mind constitutes a goal-directed behaviour (Contreras-Huerta, Pisauro, & Apps, 2020). Certainly, the definition of theory of mind used in this thesis is that a major part of it concerns inferences *about* goal-directed behaviour (e.g., Dennett, 1987; see discussion in Introduction to Chapter 5). And theory of mind would seem to be goal-directed in that its goal is to successfully perform this inference. On the other hand, it is not archetypal goal-directed behaviour in the sense that whatever theory of mind may be in its totality, it is relies on a wide and complex range of neural systems and structures (Abu-Akel & Shamay-Tsoory, 2011). The neural response to a successful inference about another's mental state does not resemble the classical reward responses associated with rewards—for example, of the kind shown by Schultz, Dayan, and Montague (1997). Overall, the reward structure of social behaviour is much more difficult to disentangle than the standard points-for-performance mechanism employed in non-

social decision tasks (Tamir & Hughes, 2018; Adolphs, 2009). The problem this poses for the current paradigm is it is difficult to convey the value of successful mental state inferences without introducing the confound of rewarding it via a standard points mechanism of explicitly rewarding correct answers. How can a paradigm encourage participants to view mental state inferences as intrinsically rewarding without activating the goal-directed psychology of point maximation?

On the other hand, it is likely still possible to find circumstances which elicit intergroup effects without mental state inference being intrinsically rewarding. It is also known that people can find effort itself rewarding (Inzlicht, Shenhav, & Olivola, 2018), and therefore any stimuli which can elicit effort cost might be, under the right framing, able to motivate reward-seeking behaviour. In a real world context, those points might, for instance, be the response of a person one is interacting with. If one correctly interprets another's mental state, they may respond with a positive remark or gesture. If one's one interpretation is incorrect, the interlocutor may correct you. It is therefore an open question about what kind of experimental designs reveal goal-directed behaviour in mental state inference.

One direction of interest for future directions in this research agenda is toward the literature on the value of information (Behrens et al., 2007). For example, a recent spate of studies have developed nuanced paradigms for assessing decision-making strategies related to the explore-exploit trade-off (Blanchard & Gershman, 2018; Gershman, 2018; Schulz & Gershman, 2019; Gershman, 2019; Wilson et al., 2021). These studies investigate the circumstances under which humans prioritise the acquisition of further information over immediate reward—typically under the auspices of maximising long-term value. Perhaps a fruitful way of assessing motivation to engage in theory of mind is as a kind of exploration:

when do people believe they need to acquire further information about a mind in question? For example, in the context of the assassin game, perhaps a version of the stimuli could be developed in which people can view the stimuli for longer while trading-off reward. How long do people watch the stimuli for before stopping the clip to make a judgment about the assassin? This is one potential revision of these theory of mind stimuli for use in addressing my questions of interest.

In summary.

I have provided strong evidence for the claim that people avoid certain mental state inferences because they are perceived as more cognitively effortful. However, I did not find support for my hypothesis that intergroup bias can contribute to these effort costs. In the following chapter, I will discuss my results across all experiments in this thesis, link these findings back to my overarching questions posed in Chapter 1, and provide a rough sketch of what the most promising avenues of future work in this direction might look like.

CHAPTER 7: GENERAL DISCUSSION

In this thesis I have examined how motivation shapes prosocial behaviour—in terms of helping, hurting, and understanding others—as a function of the social information we have about the target agent. Here I will summarise the findings I have presented in the preceding chapters, relate them to the overarching questions of interest I posed in Chapter 1, develop a potential theoretical framework for interpreting these results, and outline several studies which constitute promising future directions for building on the present work.

Summary of Findings and Contributions

An intergroup extension of the Prosocial Effort Task: What drives intergroup bias in prosocial motivation? The Prosocial Effort Task paradigm, developed by Lockwood and colleagues (2017; 2021), represents a novel and important method for distinguishing between capacity and motivation to engage in a prosocial act. This paradigm is fundamentally different from others studying prosociality in that participants not only make a decision about who gets what (such as in economic games) but actually have to carry out an associated effortful action to ensure this allocation comes to fruition. However, previous versions of the task had only investigated prosocial behaviour directed toward an anonymous other agent. I extended this task to understand how prosocial motivation varies as a function of target agent: both in an intergroup setting and as a function of perceived similarity. In particular, I showed that intergroup effects were due to a diminished discounting of in-group rewards and greater discounting of out-group rewards, whereby participants were treating decisions for in-group target agents more like decisions for themselves than decisions for an out-group agent (Study 2.1 & 2.2).

A direct manipulation of perceived similarity and its effects on prosocial behaviour: To what extent does unconfounded perceived similarity drive prosocial motivation? Previous studies have investigated the effects of perceived similarity on prosocial or altruistic behaviour (Krebs, 1975). However, most research purporting to study perceived similarity employs only correlational methods with self-reported perceived similarity or confounds the manipulated similarity with information about standard social groupings (such as gender); therefore any observed effects may be due to the intergroup processes rather than perceived similarity. Studies 3.2 and 3.3, to my knowledge, represent the first to examine the effect of perceived similarity on prosocial behaviour by manipulating perceived similarity directly and without implying confounding information about social groups. I found that perceived similarity alone likely accounts for a very small percentage of overall prosocial motivation compared to intergroup processes.

Evidence of (unincentivized) willingness to exert effort to harm out-groups: Will people exert effort to harm out-group members, even when they are given no reason for doing so? Many studies have shown that people are more likely to direct negative behaviours toward out-group members (Cikara, Bruneau, & Saxe, 2011; Cikara, 2015). However, most of these studies provide an incentive or justification for causing such out-group harm—for example, a zero-sum economic game or a moral infraction committed by an out-group member. Moreover, most studies operationalise punishment through allocation of reward, requiring the participant simply to make a decision, but not to put in effort to see that decision carried out. Studies 4.1 and 4.2 showed that participants were willing to exert effort to take points away from out-group members, even in a minimal-groups paradigm where all that is known about the out-group target is their preference between the colours blue and red. This is particularly striking given that in

other findings (Study 2.1 & 2.2) participants were willing to put in the same kind of effort to earn points for out-group members rather than take them away.

A novel assessment of effort and accuracy in theory of mind: How can individual differences in capacity and motivation to engage in mental state inference be measured? As I argue in the Introduction to Chapter 5, existing paradigms for studying theory of mind are not well-suited to assessing individual differences in the ability of neurotypical adults to perform mental state inference (Bird & Conway, 2018). I argue that this is because most paradigms either do not feature "ground-truth" mental states (i.e., there is no causal connection between the observed behaviour and the purported mental states) or are completed with near perfect accuracy by most neurotypical adults. To address these methodological deficiencies, I provide a new set of stimuli consisting of 104 video clips, which show highly reliable variation in accuracy and difficulty for the mental state inference featured in each clip (Study 5.1). While correlated, the accuracy and effort scores for these clips have unique variance—meaning that some clips may, for example, be perceived as difficult but are typically answered correctly. These clips also appear to reliably capture individual differences in performance, in both accuracy and perceived difficulty (Study 5.2). An individual's perceived difficulty ratings on these clips correlates with an existing self-report measure of cognitive empathy (i.e., theory of mind). Furthermore, these stimuli were shown to show standard signs of effort aversion (Study 6.1)

Intergroup theory of mind: A promising new frontier of psychological research. Using the novel stimuli set presented in Chapter 5, I attempted to address questions about intergroup theory of mind which previous paradigms have been unable to investigate. With a reliable measure of accuracy and perceived difficulty in a theory of mind task, I was able to explore novel questions about whether participants' accuracy or perceived difficulty decreased when a

target mind was framed as an out-group member. I found no supporting evidence for this hypothesis (Study 6.2 & 6.3), though I believe that these studies provide a foundation for promising future directions of research. In the Introduction to Chapter 6, I provide evidence for why one might hypothesise a potential intergroup effect in theory of mind. In short: (1) intergroup biases are present in almost all social-cognitive phenomena, such as empathy; (2) there are strong empirical and theoretical reasons to believe inferences about out-group minds may either be fundamentally more difficult (and therefore require more effort) or that people may be less motivated to engage in them; and (3) any such intergroup effect is a signal that current theory of mind paradigms are not equipped to detect, and that is why contemporary psychological research has not yet detected it. I hope to build on these foundations for my future research.

Evaluating Overall Questions of Interest

In Chapter 1, I introduced two overarching questions which underlie the work presented throughout this thesis. Here I reconsider these questions in light of the findings I have presented in the preceding chapters.

Question I: How does social identity affect the cost-benefit analysis of motivation?

A key concept in this thesis has been motivation: beyond simply deciding on a desired outcome, under what conditions are people willing to exert effort to see that outcome carried out? The background for this concept is an extensive neuro-computational literature on goal-directed behaviour (Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1999; Critchfield & Kollins, 2001; Botvinick, Huffstetler, & McGuire, 2009; Hartmann et al., 2015; Kobayashi &

Schultz, 2008; Hamid et al., 2016; Ruggeri et al., 2022). A recent line of inquiry within this larger research context has sought to apply this neuro-computational infrastructure toward understanding motivation of prosocial acts: when effort is directed toward the attainment of rewards which go to an agent other than oneself (Lockwood et al., 2017; Inzlicht & Hutcherson, 2017; Le Heron, Apps, & Husain, 2018; Inzlicht, Shenhav, & Olivola, 2018; Müller & Apps, 2019; Contreras-Huerta, Pisaruo, & Apps, 2020; Lockwood et al., 2021; Umemoto, Lin, & Inzlicht, 2021; Depow, Lin, & Inzlicht, 2021; Pollerhoff et al., 2022; Batson, 2022). In the following subsections, I explore several contributions I believe the research presented in this thesis make to this broader literature.

Social identity shapes reward discounting. A key concept in the psychological literature on decision-making is "discounting": the projected devaluation of a reward according to considerations such as time until attainment, prerequisite effort, and related factors (Sutton & Barto, 1999; Critchfield & Kollins, 2001; Botvinick, Huffstetler, & McGuire, 2009; Hartmann et al., 2015; Kobayashi & Schultz, 2008; Hamid et al., 2016; Lin et al., 2022; Ruggeri et al., 2022). Such decision-making models often include parameters to account for temporal and effort-based discounting (Botvinick, Huffstetler, & McGuire, 2009; Hartmann et al., 2015; Westbrook, Kester, & Braver, 2013). In this thesis, I have used this approach to explore the discounting of rewards for others, often referred to as social discounting (Jones & Rachlin, 2006; Margittai et al., 2015; Vekaria et al., 2017; Tamir & Hughes, 2018): how much is a reward devalued when it goes to someone other than oneself?

Previous work has shown that people are sensitive to valuation of reward signals based on whether that reward goes to a member of an in- or out-group (Hackel, Zaki, & Van Bavel; 2017). Hackel and colleagues showed that participants showed different neural responses when

choosing to allocate money on the basis of real groups; furthermore the extent to which participants reported identifying with their group predicted the strength of differences in neural response and prosocial decisions. It has also been shown that time pressure—which reduces one's ability to engage in cognitive control—reduced allocation of funds in an economic trust game for out-group, but not in-group, members (Hughes, Ambady, & Zaki, 2017). Additionally, the same study showed that rewards for in-group members were shown to activate neural centres typically associated with reward (ventral striatum) while rewards for out-groups were shown to activate centres associated with cognitive control (dorsal anterior cingulate cortex). While studies such as these show that people devalue rewards for out-groups more than in-groups, they do not show that social identify modulates the subjective value associated with the costs of an action—as in these experiments prosocial behaviour is operationalised as decisions about monetary allocation which do not require effort. Additionally, it is known that people are more likely to engage in emergency interventions for in-group than out-group members (Levine et al., 2005). It has also been shown that people are more motivated to help parents over close friends (Guassi Moreira et al., 2020; Guassi Moreira et al., 2021) and charities over strangers (Depow, Lin, & Inzlicht, 2022). However, prior to the studies reported in this thesis, the effects of social identity on willingness to exert effort have yet to be studied directly.

In this thesis, I have shown that people are more willing to work for in-group than out-group members (Study 2.2), a finding which is consistent with the overall direction of the above studies. However, by looking directly at the interaction between effort and reward for in-group and out-group members, I have shown that while people are sensitive to effort overall (i.e., they find harder actions more aversive than easier ones) they are not *more* sensitive to effort costs for out-group than in-group members (Study 2.2). My findings suggest that people act prosocially in

an intergroup context primarily on the basis of desired reward outcomes, with the value of rewards increasing more steeply for in-group members. In other words, the subjective value of a reward for an in-group member increases steeply as the value of the reward itself increases; whereas that increase in subjective value is not as steep for out-group members. This difference in valuation is primarily what motivates individuals to engage (or not engage) in prosocial action.

While social identity appears to play a key role in prosocial decisions, the effect it can have varies greatly depending on the situation. For example, a prosocial intergroup manipulation resulted in participants working on behalf of out-group member's rewards (Study 2.1 and 2.2) while a more antagonistic framing resulted in participants exerting the same form of effort to take points away from out-groups (Study 4.1 and 4.2). This shows the flexibility of social identity to either motivate positive or negative outcomes, depending on the decision-making affordances presented to participants. Furthermore, it shows that although people are less willing to help those other than themselves (Lockwood et al., 2017; Depow, Lin, & Inzlicht, 2022) the extent to which this is the case can be modulated greatly by variations in social context. Indeed, as shown in Study 2.2, the self-bias—in which people value rewards for themselves more than others—is almost entirely ameliorated, with people willing to work equally as hard for in-group members than they are for themselves. However, it should be stated that in these experiments, as in previous work (e.g., Vekaria et al., 2017), there are large individual differences in the extent to which people discount rewards for out-group members. Future work is necessary to disentangle potential explanations for these individual differences.

Perceived similarity also shapes discounting—and can be manipulated. Previous studies have sought to demonstrate how perceived similarity might affect prosocial behaviour (Oveis,

Horberg, & Keltner, 2010; You & Robert Jr., 2008; Hamlin et al., 2013; López-Rodríguez, Cuadrado, & Navas, 2017; Fischer, 2009; Mobbs et al., 2009). However, these studies have tended to succumb to one of two drawbacks in their experimental design: perceived similarity is either confounded with group membership (i.e., similarity is presented in terms of cues that inform typical real-world groups, such as ethnicity or socio-economic status) or is not manipulated directly (i.e., is measured via self-report and correlated with behavioural outcomes of interest). In one account, the motivational importance of similarity was found to be less important than that of social identity for the valuation of rewards obtained by other agents (Hackel, Zaki, & Van Bavel, 2017). However, this was argued on the basis that self-reported investment in one's group predicted both prosocial behaviour and attendant neural responses; thus, the authors contend that how much one cares about one's group matters more than what one has in common with the members of those groups.

Several other studies have attempted to manipulate perceived similarity directly. An early study by Krebs (1975) attempted to show differential effects on prosocial behaviour based on directly manipulating perceived similarity, however the similarity was based on overt cues about social groupings such as age, ethnicity, and gender. Sceptical that similarity plays any significant role in prosocial behaviour, Batson and colleagues (2005) argued that what matters most in prosocial decisions is feelings of nurturance (e.g., those a parent feels for a child) rather than perceived similarity. Their experiment sought to manipulate similarity directly as well, but also confounded that manipulation with cues about social groups, such as age, ethnicity, and gender. Accordingly, previous studies have not satisfactorily disentangled the varying contributions of group membership and perceived similarity on prosocial motivation.

In this thesis, I showed that both group membership (Study 2.2) and perceived similarity (Study 3.3) can affect prosocial motivation. As I manipulated perceived similarity with a personality quiz, similarity was not confounded with any overt cues associated with existing social groups—such as age, ethnicity, or gender. This finding is unlikely to be due to demand effects, as Study 3.2 showed that the same similarity manipulation with a lesser difference between the two agents did not elicit an effect. Therefore, the significant difference between agents found in Study 3.3 is likely due to participants perceiving sufficient differences in similarity between themselves and the two other agents. And while an effect of similarity was found, it was more subtle than the effects elicited in intergroup paradigms. This suggests that previous claims about substantive effects of similarity on prosocial behaviour are due mostly to insinuation of cues about social identity or group membership.

Social discounting can explain antisocial behaviour toward out-groups—without appeal to intergroup schadenfreude. Previous work has shown that people are especially sensitive to negative rewards for others—that is, they are willing to incur a cost of physical pain to lessen the pain of others, both toward anonymous other individuals (Crockett et al., 2014) and with an intergroup bias (Hein et al., 2010). In Chapter 1, I extensively reviewed a prominent framework for understanding harm directed at out-groups, called intergroup schadenfreude, in which individuals attribute a higher subjective value to negative outcomes experienced by out-group members (Cikara, Bruneau, & Saxe, 2011; Cikara, 2015; Cikara & Fiske, 2012; Cikara & Fiske, 2013; Cikara et al., 2014).

A key aspect of Cikara and colleagues' argument is that intergroup schadenfreude is brought on by competitive contexts. However, in Study 4.1 and 4.2, I show that participants were willing to cause out-group harm in a context where no competition was present. In these studies,

participants chose to exert effort to take points away from out-group members, regardless of effort costs. This suggests that participants did not find the physical effort costs associated with the task aversive, when they were incurred in service of causing out-group harm. While this is consistent with the overarching tenor of Cikara and colleagues' intergroup schadenfreude account, it shows that a competitive context is not necessary to elicit strong effects of out-group harm (as was argued in Cikara et al., 2014).

Moreover, the contrast between studies from Chapter 2 and those from Chapter 4 show the key role effort plays in intergroup decisions. In a prosocial context (Study 2.1 and 2.2), participants showed a main effect of effort, suggesting they were sensitive to the effort costs associated with earning points for in- and out-group agents. However, in an antisocial context (Study 4.1 and 4.2), participants showed no main effect of effort. They did not take effort costs into account when choosing to take points away from out-group members. This suggests that a standard account of effort aversion (e.g., Kool & Botvinick, 2018; Lockwood et al., 2017) cannot be uniformly applied across decisions in an intergroup context.

Work which is aversively effortful in one context is less so in another. The standard assumption about effort in psychological research is that it constitutes a kind of cost (Hull, 1943; Zipf, 1949; Kool & Botvinick, 2018). Studies from Chapters 2 and 3 are well-aligned with this conventional narrative. However, a nascent line of research suggests that people prefer to exert a modest amount of effort over doing nothing (Wu, Ferguson, & Inzlicht, 2021). This questioning of received assumptions about effort is furthered by the studies presented in Chapter 4. In particular, whereas participants are willing to exert help to benefit out-group members in studies from Chapter 2 (at least some of the time), studies from Chapter 4 show that participants are willing to exert the same form of effort to take points away from participants.

One possible explanation in that this antisocial reframing of the same effort manipulation alters participants' perceptions of the cost associated with exerting physical effort. The subjective value of effort costs is highly flexible: for example, people tend to rate the same acts as more effortful when they have received a higher reward for them (Pooresmaeili, Wannig, & Dolan, 2015), and people also perceive rewards into which they have invested more effort as more valuable (Inzlicht, Shenhav, & Olivola, 2018). Inzlicht and colleagues refer to this as the "effort paradox," in which effort can be both costly and valued. Examples of domains in which invested effort has been found to be increase the value of a reward, action, or object include the IKEA effect (Norton et al., 2012), the sunk cost fallacy (Thaler, 1980), earned income versus windfall gains (Muehlbacher & Kirchler, 2009), need for cognition (Cacioppo & Petty, 1982), and flow (Csikszentmihalyi, 2014). It is possible that working to take points from out-groups activates this same kind of effort-as-value mechanism.

By way of illustration, an individual playing a video game is exerting a similar form of effort to the studies presented in the first three empirical chapters (i.e., button-pressing). And depending on the game, much of this effort goes toward the purpose not only of furthering one own's aims, but detracting from the aims of one's opponents. But, colloquially, video games are not thought of as work—rather they are considered fun. Perhaps the antisocial intergroup framing of the button-pressing effort operationalisation activates a similar psychology as those of video games, as observed in the lack of a main effect of effort in Study 4.3 where it was found in Studies 2.2 and 3.3. This suggests that a promising avenue of future research would not just be to contrast the perceived costs of exerting effort versus doing nothing, as investigated by Wu and colleagues, but to contrast the perceived costs of exerting effort as a form of work versus a form of "fun"—particularly in the context of working to the detriment of out-group members.

Group-based differences in reward valuation can both drive intergroup bias in prosocial decisions—but to differing extents. I explore two countervailing forces which can influence prosocial behaviour in an intergroup context; these are the social discounting effects of an in-group "bump" and out-group "dip" (Jones & Rachlin, 2006). The formers consists in the increased valuation of rewards for in-group members, such that they are discounted to the same extent as rewards for self. The latter consists in the decreased valuation of rewards for out-group members, such that they are significantly more discounted than rewards for in-group members. Either of these effects results in the valuation of rewards for in-group members as higher than those for out-group members, more generally known as "in-group favouritism" (Everett et al., 2015). However, they result in different patterns of behaviour, which can be identified by different experimental paradigms. Many studies have found various applications in which either increased valuation of positive outcomes for an in-group or negative outcomes for an out-group provides a better explanation for the observed data (Brewer, 1999; Hewstone et al., 2002; De Dreu et al., 2014; Durrheim et al., 2016; Kok et al., 2020). It is likely that both forces are able to play an important role depending on the situation at hand. However, to my knowledge, no studies have pitted these two forces against each directly to study their differential effects on prosocial motivation: to what extent are people willing to exert effort to carry out goal-directed behaviour in line with the discounting strategies of an in-group bump or an out-group dip? In this thesis, I provide evidence that both are important factors in social discounting (Study 2.2) and that perceived similarity on its own can motivate prosocial behaviour resembling parochial altruism (Study 3.3). In Studies 2.1 and 2.2, I provide evidence from both task behaviour and computational modelling which suggests participants were using a similar strategy to make

decisions for themselves and in-group members, while using a different strategy for out-group members.

One strength of the present work is its highly structured, high internal validity experimental design. Rewards and effort are operationalised in a straightforward, abstract manner, as are the minimal groups members. Thus the various effects of an in-group bump or out-group dip are straightforward to interpret in the data. Contrast this with the field study by Kok and colleagues (2020), which featured a less structured but high external validity study. Because both the in-group bump and out-group dip likely both play a role in parochial altruism and prosocial motivation to various degrees in various contexts, a study focused on one specific context will be better suited to describing the precise dynamics of that situation. On the other hand, my studies using the Intergroup Prosocial Effort Task are almost completely devoid of contextual cues. Therefore, they likely have high generalisability to many contexts in which prosocial decisions are made but only explain a modest amount of variance in decision-making, whereas a more ecologically-valid study like Kok and colleagues' would have low generalisability but in the cases where it does apply explain a higher amount of decision variance.

A recent theoretical account of Batson (2022) has suggested that prosocial motivation can be understood as four components: egoism (increasing one's own welfare), altruism (increasing the welfare of individuals who are not oneself), collectivism (increasing the welfare of a group), and principlism (defining welfare in terms of some a priori moral principle, standard, or ideal). As noted by Dovidio (2022), this conceptualisation of prosocial motivation emphasises the role of groups and social identity more than most existing accounts of prosocial motivation. This is ultimately in line with my goal in this thesis to further the field's understanding of how prosocial

motivation changes depending on is what is known about the person toward whom the behaviour is directed. My studies (particular those presented in the first three empirical chapters) are particularly well-suited to exploring the interactions between Batson's first three components. In each experiment, egoism is compared directly against altruism—with various manipulations of perceived collectivism to influence the actor's decisions toward the other agent. However, in the present experiments there was both a distinct lack of principlism, as well as a lack of contribution toward a collective, rather than an altruistic contribution toward an individual based on collective interests. In summary, this thesis has contributed to furthering current understanding of prosocial motivation—how it varies in intergroup contexts, what kinds of information about target agents influence such motivation, and how framing effects can dramatically alter the costs of effort and the valuation of reward.

Question II: What effort costs might be associated with mental state inference?

In the work presented in Chapters 5 and 6 of this thesis, I developed a novel set of theory of mind stimuli, based on recorded rounds of participants playing the "Assassin Game." In Chapter 5, I surveyed existing theory of mind paradigms and argued that these novel stimuli rectify two important issues which have marked previous theory of mind paradigms: most existing paradigms either do not feature "ground-truth" mental states (i.e., there is no causal connection between the observed behaviour and the purported mental states) or are completed with near perfect accuracy by most neurotypical adults. These novel stimuli feature "ground-truth" mental states, since there is a causal connection between the mental states of the agents controlling the shapes in the Assassin Game and the behaviour observed in the video clips. This stands in contrast to many theory of mind paradigms, in which the "correct" mental states to

ascribe to the observed behaviour are defined by the researchers' own perceptions. I have also shown across Studies 5.1, 5.2, and 6.1 that these stimuli feature: high test-retest reliably for accuracy scores and difficulty ratings associated with the individual clips; individual-level test-retest reliable comparable to other accepted measures of cognitive empathy, such as the QCAE; a dissociation between a clips' accuracy scores (what percentage of participants get the answer to a clip correct) and difficulty ratings (how difficult participants perceive that clip to be); and standard markers of effort aversion, which are not explained by confounds with inferential accuracy. I maintain that this represents a significant and useful addition to the set of existing theory of mind paradigms.

One question, which I pre-emptively raised in the Introduction to Chapter 5, is to what extent do these stimuli really capture a process which truly represents theory of mind? This is a concerned that is commonly levelled against theory of mind tasks (Quesque & Rosetti, 2020). One possible criticism of these stimuli is that the motion of simple shape-based agents on a screen is a sufficiently impoverished instance of intentional behaviour that it not useful for making generalisations about theory of mind in any meaningful broad sense (for similar concerns in paradigms studying theory of mind in infants, see Packer & Moreno-Dulcey, 2022). However, as a first line of rebuttal against this criticism, it worthwhile to note that unlike the majority of theory of mind stimuli, these clips feature actual behaviour performed by actual intentional agents. Therefore any inferences made about the behaviour of the shapes are by definition inferences about intentional agents.

A more nuanced form of this criticism might suggest that it is not actually necessary to engage in the "intentional stance" (Dennett, 1987) to make accurate inferences about the goal-directed dynamics of dot-motion agents, but only the "physical stance." One possibility is that

heuristic based on physical trajectory (without appeal to goals, beliefs, or desires) is all that is necessary to model participants' judgments in this task. I believe this alternate account is worth exploring, and cannot be entirely ruled out by the present data. I outline several different ways of testing the intentional versus physical accounts of judgments with these stimuli in the section on future directions below. Nonetheless, the present usage of dot-motion as an operationalisation of theory of mind is consistent with decades of use of animated avatars to represent intentional agents (Heider & Simmel, 1944; Baker, Saxe, & Tenenbaum, 2009; Quesque & Rossetti, 2020; Houlihan, Tenenbaum, & Saxe, 2021; Ho, Saxe, & Cushman, 2022). It is nonetheless true that this operationalisation of theory of mind is specific to inferences about the unobservable beliefs and desires of goal-directed agents conditional on their publicly observable behaviours, which although a key aspect of theory of mind (Ho, Saxe, & Cushman, 2022) does not include a range of phenomena which could be considered as important aspects of social cognition (Batson, 2009). Therefore, my specific claim is that these stimuli capture a relevant and important part of effort and accuracy in mental state inference—and one which represents a significant advance over existing methods in the field—but not that they capture every possible aspect of interest related to theory of mind.

Theory of mind is effortful in way that is similar to empathy—but dissociable from it. Most previous research measuring the cognitive costs of assessing another agent's mental states focuses on affective states (i.e., empathy) rather than cognitive states (i.e., theory of mind; Cameron, 2018; Cameron et al., 2019; Ferguson, Cameron, & Inzlicht, 2020; Weisz et al., 2020; Contreras-Huerta, Pisauro, & Apps, 2020; Ferguson & Inzlicht, 2022). Consequently, there is nascent evidence to support the proposition that empathy is an effortful process, but much less evidence to support the existence of a similar effort cost associated with theory of mind (Lin,

Keysar, & Epley, 2010). Indeed, theory of mind is often treated as a kind of reflexive process, one that is only absent in the presence of some significant cognitive deficiency (Senju, 2012).

However, in the work presented throughout this thesis, I show that stimuli based on theory of mind elicit standard signs of effort aversion (Study 6.1). This discounting of harder theory of mind stimuli cannot be accounted for by differences in length (e.g., where harder clips take longer), inferential complexity (e.g., where harder clips feature more agents or more complicated scenarios), or efficacy (e.g., where people choose not to do clips because they are concerned they will get them incorrect). Furthermore, the mental state inferences are sufficiently constrained to be able to isolate the relevant mental state in question: that is, to find the agent who is attempting to approach (or avoid) another agent. This stands in contrast to paradigms such as the empathy selection task (Cameron et al., 2018), which instructs participants to empathise with a pictured individual but cannot demonstrate the extent to which the participants successfully identified or authentically felt the emotions shown in the stimuli. Study 5.1 and 5.2 provide correlational evidence that the effortful component of these stimuli is related to cognitive—but not emotional—empathy as measured by the QCAE self-report scale (Reniers et al., 2011). This is consistent with the account that these stimuli activate standard processes associated with mental state inference. I maintain that this is the first paradigm which can show strong evidence for effort aversion associated with mental state inference in a way that satisfactorily dissociates theory of mind from affective empathy.

Individuals differ in their willingness to exert effort for mental state inferences. Individual differences in theory of mind ability have proved difficult to study, because most current theory of mind paradigms are able to be completed by neurotypical adults with near-perfect accuracy (Conway & Bird, 2018; Apperly, 2012). When individual differences in

participants' ability to understand other minds are assessed, these measurements are usually done indirectly by putting participants under time pressure (Lin, Keysar, & Epley, 2010; Hughes, Ambady, & Zaki, 2017). By analogy, this is like assessing individual differences in students' mathematical abilities by giving each of them the same set of problems under different levels of time pressure. An alternative, and more standard, approach would be to present them with problems featuring a range of difficulties and evaluate which of the problems they get right and which they get wrong. However, because most theory of mind paradigms do not present participants with stimuli that require non-trivial mental state inferences (Conway & Bird, 2018), this approach has remained underexplored.

In the present work, I show that people vary in their ability to accurately assess these novel theory of mind stimuli (Study 5.1) and that this variation is reliably assessed across the span of several months (Study 5.2). Moreover, participant accuracy is only loosely correlated with the self-reported perceived difficulty of each stimuli. This suggests a dissociation between motivation and capacity to perform mental state inference, which has been shown to be a consistently observed dissociation across social-cognitive phenomena—but previously only indirectly in theory of mind (Contreras-Huerta, Pisauro, & Apps, 2020). Accordingly, the present research provides evidence that people both vary in their theory of mind abilities, and that this variation can be due to either diminished capacity or motivation.

Inconclusive results on intergroup bias in willingness to make mental state inferences. In Chapter 6, I hypothesised that participants would find inferences about out-group mental states more cognitively effortful than inferences about in-group members. I hypothesised that participants would be more averse to out-group mental state inferences because people show attenuated empathy toward out-group members (Cikara, Bruneau, & Saxe, 2011); people are

more likely to use mental state terms to describe in-group than out-group members (McClung & Reicher, 2018); people are generally less prosocially inclined toward out-group members (Study 2.1 and 2.2), and mental state inference helps to facilitate prosociality (Gaesser, Keeler, & Young, 2018); and out-group beliefs are more easily misjudged than in-group ones (Lees & Cikara, 2020). However, across two studies I failed to find support for this hypothesis, either in participants' self-reported difficulty for in-group or out-group clips (Study 6.2) or in people's choices to make inferences about in-group versus out-group clips in an effort task (Study 6.3).

While these experiments did not find support for the hypothesis in question, there are several factors that may have limited the present work's ability to satisfactorily address it. One possibility is that out-group minds might be harder to understand because they are more different from one's own than in-group minds, and theory of mind has been shown to rely on a kind of ego-centric simulation (Epley et al., 2004; Gallese & Goldman, 1998; Apperly, 2008; see also the discussion in Chapter 1). However, this would not necessarily be the case for a minimal groups paradigm in general, or for the particular minimal groups paradigm employed in Studies 6.2 and 6.3—as the stimuli were not created by members of one's in- or out-group group, but manipulated post-hoc to be framed as reflecting the behaviour of in- or out-group members. There is also a concern with Study 6.2 that the connection between group membership and stimuli was not sufficiently clear to elicit intergroup effects (i.e., participants were told whether the assassin was an in- or out-group member, but were not informed about the membership of the other four players or given any group-level cues beyond inferences about which player is the assassin). Thus, it is not possible to rule out this hypothesis, as it was not the primary one these experiments were designed to test.

Another possibility is that out-group minds might be harder to understand because people are less motivated to make inferences about them. This hypothetical account was more directly tested in Studies 6.2 and 6.3. The rationale beyond this hypothesis was based on evidence that people find theory of mind similarly effortful to empathy (Studies 5.1, 5.2, & 6.1) and empathic abilities have been shown to be attenuated for out-group members (Cikara, Bruneau, & Saxe, 2011). Both Studies 6.2 and 6.3 suffered from limitations that could have prevented them from shedding light on this potential account. As noted above, Study 6.2 featured a tenuous connection between group membership and mental states, which likely inhibited any difference in motivation that might have been found. While the paradigm used in Study 6.3 was designed to make a stronger connection between mental states and group membership (i.e., each animated avatar was colour-coded for group membership, though again this was done post hoc), one potential limitation was that participants were still informed of the effort level associated with each clip. Clips were presented as being made by in- or out-group, but in retrospect it seems unlikely that participants were given sufficient ambiguity about the clips to make a decision based on whether or not they were created by an out-group member. In other words, if the researchers have told you how effortful a clip will be, why would you update your anticipated effort costs based on extraneous information such as group affiliation? Any possible effect which might have been found would have been drowned out by the explicit information about a clip's effort cost.

Ultimately, these hypotheses about intergroup differences in theory of mind were based on more exploratory, more novel paradigms than the other studies in this thesis. Prosocial motivation with physical effort is well-studied (Lockwood et al., 2017; Lockwood et al., 2021), as is the cost associated with cognitive effort generally (Kool & Botvinick, 2018) and with

social-cognitive processes specifically (Contreras-Huerta, Pisauro, & Apps, 2020). Studies 6.2 and 6.3 represented a larger departure from existing literature, investigating hypotheses which have not been extensively explored by existing paradigms. Thus, I contend that the questions these studies were intended to address are still worth asking—and that this is an appropriate purview for future research. In the following section, I discuss a potential theoretical framework which could accommodate future studies in this vein.

The Intuitive Anthropologist Framework

Understanding others' mental states is typically thought of as a form of "intuitive psychology" (Ross, 1977). This framing of how we understand others has led to the development a productive research paradigm, as surveyed throughout this thesis in the context of theory of mind. However, psychology is just one of many behavioural or social sciences which seeks to understand the way people work. While research has explored the extent to which people resemble informal versions of other scientific domains such as physics (e.g., Sosa et al., 2021), fewer accounts have been given about how human cognition and behaviour might be compared other social and behavioural sciences.

In particular, one historical shortcoming of psychology has been a focus on one particular demographic group of individuals, often summarised as Western, Educated, Industrialised, Rich, and Democratic (Henrich, Heine, & Norenzayan, 2010). Given that this is the same demographic as the majority of the researchers who historically have conducted this research, this could be considered an in-group bias—in which studies of the in-group have been conducted in order to generalise not just to people like oneself but to people in general. Therefore, an intuitive

psychologist might suffer from the same shortcomings as formal psychology historically has: the assumption that out-group minds work in more or less the same way as in-group minds.

By contrast, there is another behavioural science in which this assumption is not present: anthropology. The baseline assumption of cultural and social anthropologists is that understanding people from a different social group than one's own requires a very different set of skills than those useful for understanding one's in-group (Malinowski, 1922; Geertz, 1973; Geertz, 1983). While there is overlap between cultural anthropology and social psychology (e.g., Shweder, 1991), they are largely separate fields with separate methods. Thus it is potentially useful to extend the intuitive psychology metaphor into the key differences between these separate fields. To what extent do people resemble "intuitive anthropologists" when they attempt to understand the behaviour of out-group members?

At a first approximation, a major methodological difference between these two fields is that psychology traditionally relies on experimental methods while cultural anthropology traditionally relies on qualitative fieldwork. Psychologists set up experiments to isolate variables, impose interventions, and conduct statistical tests in order to understand the causal role those variables play in the behaviour observed in the different conditions of the intervention. By analogy, the intuitive psychologist also performs inferences about the causes of behaviour, though without the help of formal experiments and statistical tests. Anthropologists, by contrast, do not conduct experiments, but instead spend long stretches with a particular group of people in a process known as participant-observation (Geertz, 1983). A primary form of analysis for anthropologists engaged in participant-observation is called "thick description" (Geertz, 1973).[1]

[1] At the time of this writing, the seminal essay collection describing thick description (Geertz, 1973) has been cited 136,089 times, which shows just how widespread a concept it is in the qualitative social sciences. It is curious that the term is entirely absent from psychology.

Whereas an anthropologist would consider a statistical analysis a "thin" description, thick description is closer to literary analysis—indeed, its progenitor compared the methods appropriate for interpretations of cultural practices with the methods for interpretation of a literary text (Geertz, 1973). The purview of the intuitive psychologist is essentially point-estimates of another individual's belief. The intuitive anthropologist's aim would be something closer to a thick description of an out-group member's worldview.

In terms of psychological mechanisms, the intuitive psychologist is defined by the inferential mechanisms she employs (Baker, Saxe, & Tenenbaum, 2009; Baker et al., 2017). By contrast, anthropologists do not rely on complex inferential mechanisms to support the reliability of their findings (Geertz, 1983). Rather, a key concept is "being there" (Geertz, 1988). This is the shorthand for the idea that understanding another culture is not about the sophistication of one's ability to put themselves "in the other person's shoes" but rather it is a function of the extent to which they were immersed in the other person's culture and were able to develop first-hand knowledge of the individual's lived experience. By no means is this a fool-proof method—just as psychological methods have also had their own setbacks with issues like replication and generalisability—but it is at the very least a completely separate way of going about understanding behaviour than the typical methods used by psychologists. Furthermore, the key psychological construct behind "being there" is motivation. It is about one's willingness to put in the effort to experience a way of life different from one's own (Geertz, 1983). Whereas the intuitive psychologist model focuses on inferential mechanisms used for understanding other minds, the intuitive anthropologist model focuses on motivation to gather relevant data in service for understanding other (specifically out-group) minds. In short, intuitive psychology is about inference whereas intuitive anthropology is about motivation.

One term often employed within the intuitive psychologist framework is "perspective-taking" (Epley et al., 2004). Colloquially, perspective-taking is often formulated as putting oneself in the other person's shoes—i.e., introspection into another individual's mind (Epley & Caruso, 2012). Anthropologists have gone to great lengths to find ways of conveying their participant-observation data which do not rely on the mental gymnastics of perspective-taking (Geertz, 1983). For example, one of the alternative framings is the distinction between experience-near and experience-far descriptions (Geertz, 1983). To describe an individual's perspective in an experience-near manner is to describe it in emotional or logical terms similar to those they would use themselves, reflecting a first-person perspective. Experience-far descriptions are those that would be employed in an analysis by a specialist, reflecting a third-person perspective. Both are necessary for a successful thick description (Geertz, 1983). As an alternative to perspective-taking, one account has proposed the idea of "perspective-getting" (Eyal, Steffel, & Epley, 2018). This is the simple idea that the most effective way of understanding the contents of another person's mind is to ask them for it. While the concept has not been much utilised in further psychological research, perspective-getting is consonant with the approach of an intuitive anthropologist framework. A key question for future research would be the extent to which people are willing to "get" the perspective of in-group versus out-group members.

The studies presented in this thesis (particularly in Chapters 5 and 6) explore key concepts of a potential intuitive anthropologist framework—specifically, the role of motivation in intergroup theory of mind—using methods consistent with the framework of the intuitive psychologist. While I have demonstrated novel findings related to the effort costs associated with mental state inference, I contend that future research should begin to incorporate approaches of

this novel framework. Extending the current work's focus on high internal validity studies with many reward and effort levels, future studies could prioritise a more ecological approach: for example, exploring how people use perspective-taking versus perspective-getting to understand differences in belief, morals, or social norms among in-group versus out-group members. This framework provides a longer-term direction for this research agenda to work toward, which may prove to be a useful way to think about the role of motivation to engage in intergroup theory of mind within the larger psychological literature.

Future Directions

The present work on effort in theory of mind—as well as my novel theory of mind stimuli—represent a potential avenue for establishing my own unique research agenda, investigating a space of questions which fewer studies has addressed. In this final section, I describe some of the future directions I would like to pursue in my research on effort and theory of mind in an intergroup context.

Exploring a physics/mechanics-based model of the theory of mind stimuli. One question I raised in the discussion of my novel theory of mind stimuli is the extent to which judgments about the stimuli truly depend on the social-cognitive process of theory of mind. In the terms of Dennett (1987), the question is whether the intentional stance is needed to make sense of the videos or instead whether the physical stance would suffice. In other words, how much do motion cues account for difficulty ratings and accuracy scores? For example, if a model accounted for the motion vectors of each shape in the Assassin Game, could it predict participants' selection of the assassin by looking at the direction each player is going in a clip's final frames and selecting the agent most directly moving toward another? While the model for

this account would be fairly straightforward to implement—it only needs to know about motion vectors, since for example all the players are moving at the same constant velocity—it would be non-trivial to translate the existing video clips into a format which quantitatively encodes players' motion trajectories in a way that could be utilised by the model. However, if it can be shown that this model does not account fully for participants' judgments, that would provide strong evidence that these stimuli actually do depend on mind theoretic inferences and not simply lower-level processes.

Experiments with non-mental state framings. In addition to using a physics-based model to explain variance in participants' judgments, an alternative would be to present the same experiment with a different backstory. An explanation of participants' judgments of these stimuli via lower-level processing (and not mental states) would predict that changing the narrative motivating the shapes' motion in the clips should *not* change participants' judgments. For example, participants could be told that that the shapes in the clips are not agents playing a game, but planetary bodies in motion, each with its own potential gravitational pull. Each body has a different density, and in this universe only one dyad (a "planet" and a "moon") is affected by one another's gravitational pull. At some point within the game, the planet can turn on its gravity to attract the moon. The participant is asked: which shape is the planet and which is the moon? This is an identical question to the original one about assassins (the moon) and targets (the planet), but it makes no reference to mental states. If participants showed different accuracy scores and difficulty ratings on the stimuli with this alternative framing, then that would suggest the framing with mental states makes a crucial difference—and therefore supports the argument that these stimuli indeed draw on mind theoretic resources.

However, it should be noted that for both the physics-based model and experimental framing a deviation from current results would provide support for the argument that these stimuli are fundamentally about theory of mind, the opposite result would not necessarily prove otherwise. For example, even if a physics-based model predicts judgments on the task, agent trajectory could still play a key role in mental state inferences. Likewise, in the alternative experimental framing, participants could be attributing agency to the planetary bodies (with gravitation working as a kind of "desire") and basing their inferences accordingly.

Working memory and other cognitive capacities. Given the test-retest reliability for the stimuli corresponds with existing assessments such as the QCAE, it would potentially be useful to compare participants' scores on these theory of mind stimuli with their performance in other cognitive tasks, for example on tests of working memory capacity. This would be another way of helping to explain the variance underlying which aspects of the stimuli are specific to social-cognitive processes versus other non-social cognitive processes.

Assessing the mind theoretic capabilities of the Assassin Game players. What makes one clip harder than another? One explanation could come from the varying behaviours of the players who controlled the shapes featured in the clip. Certain players may be more effective assassins than others. Perhaps there is a way to measure these behaviours, inclinations, or strategies in a way that predicts how difficulty their clips will be perceived as later on.

Eye-tracking studies. How do people go from tracking all five players to homing in on one pair in particular? Eye-tracking has proven useful in assessing how people judge causal relationships between intentional agents represented by simple shapes (Gerstenberg et al., 2017). A similar approach could be used in these stimuli. What sort of motion most influences judgments of who is an assassin and who is a target?

Evidence accumulation studies. In the present version of the stimuli, all clips last exactly five seconds. However, another way to structure them would be to make them variable in length—or more precisely, to create longer clips, then present them to participants while allowing them to terminate the clip whenever they are sufficiently certain of who the assassin is. The sooner they terminate the clip, the greater the amount of points they would receive for a correct judgment. If they make an incorrect judgment, they would receive no reward or even a negative reward. If the clips also were encoded (as discussed in the physical-based model proposal above) to capture the quantitative motion vectors of each agent, then those data could be used to inform an evidence accumulation drift diffusion model (e.g., Krajbich & Rangell, 2011). This would also align well with my proposals about theory of mind as a kind of exploration, in the discussion of Chapter 6, and could also be usefully studied by models of optimal foraging (Gabay & Apps, 2021).

Autism studies. As currently framed for intergroup theory of mind, the present stimuli feature clips that do not actually vary in their behaviour as a function of group membership. But what if group-based clips were in fact created by groups with substantively different minds? For example, one set of Assassin Game clips could be generated by participants with Autism Spectrum Disorder and one set generated by neurotypical individuals. Do participants find one group harder to make inferences about than the other group? This kind of study could complement the motivational questions about intergroup theory of mind with capacity-based questions—whether it is in fact harder to make inferences about different kinds of mind.

A multi-pronged approach: experiments, modelling, and naturalistic studies. More generally, I believe that the larger questions I have posed around intergroup theory of mind are addressable only by a multi-pronged approach. A highly-structured, high internal validity

experimental approach such as the Prosocial Effort Task is only able to capture a small range of possible findings on this. For example, what one means when one looks across the political aisle and finds the other side's behaviour lacking in rational motivation is indeed a form of mental state inference, but not of the kind typically studied in theory of mind. I believe a combination of highly structured experiments using the present stimuli, naturalistic studies looking at behaviour in more open ended situations, and a rigorous computational framework will all be necessary to piece together the puzzle of intergroup theory of mind.

In the Introduction to this thesis (Chapter 1), I defined the purview of theory of mind as it relates to my experiments according to the Baker and colleagues' (2009; 2017) model. In this model, the primary inferential considerations stem from the interaction of beliefs, goals, and the immediate environment. Outside these considerations of rational intentionality and the immediate situational context, there is a node for "world knowledge." This node is extraneous to the model's inferences, present mostly to ensure basic commonalities in understanding, for example such that both agents know that gravity tends to bring an agent closer to earth's surface rather than further away from it. However, I believe this is a key component in intergroup theory of mind: when the target agent is from an entirely separate social group, they may not share the same world knowledge at all. A dramatic instance comes, for example, from conspiracy theorists (Douglas, Sutton, & Cichocka, 2017). Everyone may believe that gravity pulls objects down, but many people believe that the earth is flat. World knowledge, in this case, is not extraneous to an inference about the mental states underlying the behaviour of an agent who holds this belief; it is a core element that an inference is made about. How is one supposed to make inferences about the actions of an agent who has such fantastically different world knowledge than oneself? This is a crucial question toward which very little research has been directed. One broad area of

expansion of my current research would be toward building a computational framework that focuses on inferences about differences in world knowledge as they pertain to an individual's beliefs and desires.

Another broad area takes an opposing approach: naturalistic studies. If what is important is the messiness of different worldviews, then more naturalistic, ecologically-situated studies might be more appropriate to address some of my questions of interest. Whereas the value of my novel theory of mind stimuli is that they are highly manipulable by the experimenter, the drawback is that they drastically limit the space of possible mental states that can be relevant: essentially, a unidimensional spectrum of approach or avoid. The kinds of beliefs that would differ between members of separate social groups are likely to be much broader than neat laboratory studies are able to control precisely. In future research, I hope to find ways of making contact between the more traditional experimental approach I have taken in this thesis and a wider range of higher ecological validity settings in which explore these questions. In conclusion, the frontier of intergroup theory of mind represents a set of questions that have not been asked as frequently or in as great as depth as other topics covered in this thesis. To a significant degree, they have not been asked at all. Therefore, I believe this represents the most promising direction for my future work.

Ultimately, my personal hypothesis is that future research will show that when it comes to our motivation to understand out-group minds, people resemble Dr Sacks's post-encephalitic patients. We stand there, metaphorically motionless, unable to generate the behavioural force necessary to determine the nuances in worldviews that differ from our own. Theory of mind may be an intrinsic and automatic element of human social cognition, but the success we experience in understanding the separate cognitive universe of another agent depends on the social identity

of the agent toward whom we are directing our inferential prowess. My hope is that this thesis represents a first step toward understanding in greater detail the circumstances under which people are motivated to help, to hurt, and—especially—to understand target agents from different social groups.

REFERENCES

Aaldering, H., Ten Velden, F. S., van Kleef, G. A., & De Dreu, C. K. (2018). Parochial cooperation in nested intergroup dilemmas is reduced when it harms out-groups. *Journal of personality and social psychology*, *114*(6), 909.

Abbink, K., & Harris, D. (2019). In-group favouritism and out-group discrimination in naturally occurring groups. *PloS one*, *14*(9), e0221616.

Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1-16.

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia*, *49*(11), 2971-2984.

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual review of psychology*, *60*, 693.

Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting.

Allport, G. W. (1954). The nature of prejudice.

Altiero, E. C., Kang, Y. J., & Peecher, M. E. (2022). Motivated perspective taking: Why prompting auditors to take an investor's perspective makes them treat identified audit differences as less material. *Contemporary Accounting Research*, *39*(1), 339-370.

Alós-Ferrer, C., García-Segarra, J., & Ritschel, A. (2022). Generous with individuals and selfish to the masses. *Nature human behaviour*, *6*(1), 88-96.

Amira, K., Wright, J. C., & Goya-Tocchetto, D. (2021). In-group love versus out-group hate: Which is more important to partisans and when?. *Political Behavior*, *43*(2), 473-494.

Anderson, J., & Cheers, C. (2018). Does the dark triad predict prejudice?: The role of

machiavellianism, psychopathy, and narcissism in explaining negativity toward asylum seekers. *Australian Psychologist*, *53*(3), 271-281.

Anderson, N. E., & Kiehl, K. A. (2012). The psychopath magnetized: insights from brain imaging. *Trends in cognitive sciences*, *16*(1), 52-60.

Ang, Y. S., Manohar, S., & Apps, M. A. (2015). Commentary: noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. *Frontiers in behavioral neuroscience*, *9*, 310.

Ang, Y. S., Lockwood, P., Apps, M. A., Muhammed, K., & Husain, M. (2017). Distinct subtypes of apathy revealed by the apathy motivation index. *PloS one*, *12*(1), e0169938.

Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: why social cognitive neuroscience should use its own concepts to study "Theory of Mind". *Cognition*, *107*(1), 266-283.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, *65*(5), 825-839.

Apps, M. A., McKay, R., Azevedo, R. T., Whitehouse, H., & Tsakiris, M. (2018). Not on my team: Medial prefrontal cortex responses to ingroup fusion and unfair monetary divisions. *Brain and Behavior*, *8*(8), e01030.

Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The anterior cingulate gyrus and social cognition: tracking the motivation of others. *Neuron*, *90*(4), 692-707.

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of psychosomatic research*, *38*(1), 23-32.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative

attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1-10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological bulletin*, *140*(6), 1556.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325-344.

Baron, J. (2001). Confusion of group interest and self-interest in parochial cooperation on behalf of a group. *Journal of Conflict Resolution*, *45*(3), 283-296.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Batson, C. D. (2009). These things called empathy: eight related but distinct phenomena.

Batson, C. D. (2022). Prosocial motivation: A Lewinian approach. *Motivation Science*, *8*(1), 1.

Batson, C. D., Lishner, D. A., Cook, J., & Sawyer, S. (2005). Similarity and nurturance: Two possible sources of empathy for strangers. *Basic and applied social psychology*, *27*(1), 15-25.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214-1221.

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., & Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology*, *38*(8), 1495-1503.

Berke, M., & Jara-Ettinger, J. (2022). Integrating Experience into Bayesian Theory of Mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912-915.

Berry, D. R., & Brown, K. W. (2017). Reducing separateness with presence: How mindfulness catalyzes intergroup prosociality. In *Mindfulness in social psychology* (pp. 153-166). Routledge.

Bilancini, E., Boncinelli, L., Capraro, V., Celadin, T., & Di Paolo, R. (2020). "Do the right thing" for whom? An experiment on ingroup favouritism, group assorting and moral suasion. *arXiv preprint arXiv:2002.12464*.

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed?. *Frontiers in psychology*, *9*, 2558.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 117-126.

Bortolini, T., Bado, P., Hoefle, S., Engel, A., Zahn, R., de Oliveira Souza, R., ... & Moll, J. (2017). Neural bases of ingroup altruistic motivation in soccer fans. *Scientific reports*, *7*(1), 1-12.

Botvinick, M. M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus

accumbens. *Cognitive, affective, & behavioral neuroscience*, *9*(1), 16-27.

Bruneau, E. G., Cikara, M., & Saxe, R. (2017). Parochial empathy predicts reduced altruism and the endorsement of passive harm. *Social Psychological and Personality Science*, *8*(8), 934-942.

Burton-Chellew, M. N., & Dunbar, R. I. (2015). Hamilton's rule predicts anticipated social support in humans. *Behavioral Ecology*, *26*(1), 130-137.

Cacault, M. P., Goette, L., Lalive, R., & Thoenig, M. (2015). Do we harm others even if we don't need to?. *Frontiers in psychology*, *6*, 729.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116.

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, *148*(6), 962.

Campbell, A. V., Chung, J. M. H., & Inzlicht, M. (2022). Meaningfulness of effort: Deriving purpose from really trying.

Capraro, V., Everett, J. A., & Earp, B. D. (2019). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*, *83*, 142-149.

Carpenter, J. M., Green, M. C., & Vacharkulksemsuk, T. (2016). Beyond perspective-taking: Mind-reading motivation. *Motivation and Emotion*, *40*(3), 358-374.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in cognitive sciences*, *16*(4), 231-239.

Chong, T. T. J., Apps, M., Giehl, K., Sillence, A., Grima, L. L., & Husain, M. (2017). Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS*

biology, *15*(2), e1002598.

Chong, T. T. J., Apps, M. A., Giehl, K., Hall, S., Clifton, C. H., & Husain, M. (2018). Computational modelling reveals distinct patterns of cognitive and physical motivation in elite athletes. *Scientific reports*, *8*(1), 1-11.

Cikara, M. (2015). Intergroup schadenfreude: Motivating participation in collective violence. *Current opinion in behavioral sciences*, *3*, 12-17.

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, *20*(3), 149-153.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of experimental social psychology*, *55*, 110-125.

Cikara, M., & Fiske, S. T. (2012). Stereotypes and schadenfreude: Affective and physiological markers of pleasure at outgroup misfortunes. *Social Psychological and Personality Science*, *3*(1), 63-71.

Cikara, M., & Fiske, S. T. (2013). Their pain, our pleasure: stereotype content and schadenfreude. *Annals of the New York Academy of Sciences*, *1299*(1), 52-59.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science*, *9*(3), 245-274.

Clay, G., Mlynski, C., Korb, F. M., Goschke, T., & Job, V. (2022). Rewarding cognitive effort increases the intrinsic value of mental labor. *Proceedings of the National Academy of Sciences*, *119*(5), e2111785119.

Contreras-Huerta, L. S., Pisauro, M. A., & Apps, M. A. (2020). Effort shapes social cognition and behaviour: A neuro-cognitive framework. *Neuroscience & Biobehavioral*

Reviews, *118*, 426-439.

Conway, J. R., & Bird, G. (2018). Conceptualizing degrees of theory of mind. *Proceedings of the National Academy of Sciences*, *115*(7), 1408-1410.

Conway, J. R., Catmur, C., & Bird, G. (2019). Understanding individual differences in theory of mind via representation of minds, not mental states. *Psychonomic bulletin & \ review*, *26*(3), 798-812.

Conway, J. R., Coll, M. P., Cuve, H. C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (2020). Understanding how minds vary relates to skill in inferring mental states, personality, and intelligence. *Journal of Experimental Psychology: General*, *149*(6), 1032.

Coricelli, G. (2005). Two-levels of mental states attribution: from automaticity to voluntariness. *Neuropsychologia*, *43*(2), 294-300.

Critchfield, T. S., & Kollins, S. H. (2001). Temporal discounting: Basic research and the analysis of socially important behavior. *Journal of applied behavior analysis*, *34*(1), 101-122.

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

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

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*, *25*(14), 1852-1859.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in*

experimental social psychology, *40*, 61-149.

Cutler, J., Nitschke, J. P., Lamm, C., & Lockwood, P. L. (2021). Older adults across the globe exhibit increased prosocial behavior but also greater in-group preferences. *Nature Aging*, *1*(10), 880-888.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience* (Vol. 1990). New York: Harper & Row.

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, *26*(5), 593-620.

Dawkins, R. (1976). *The Selfish Gene*.

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., ... & Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, *328*(5984), 1408-1411.

De Dreu, C. K., Fariña, A., Gross, J., & Romano, A. (2022). Prosociality as a foundation for intergroup conflict. *Current Opinion in Psychology*, *44*, 112-116.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Depow, G. J., Lin, H., & Inzlicht, M. (2022). Cognitive effort for self, strangers, and charities. *Scientific reports*, *12*(1), 1-11.

Dishon, G., & Kafai, Y. B. (2020). Making more of games: Cultivating perspective-taking through game design. *Computers & Education*, *148*, 103810.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current directions in psychological science*, *26*(6), 538-542.

Dovidio, J. F. (2022). Interpersonal relations, group processes, and prosocial motivation: Expanding the perspective. Commentary on Batson (2022).

Drayton, L. A., Santos, L. R., & Baskin-Sommers, A. (2018). Psychopaths fail to automatically take the perspective of others. *Proceedings of the National Academy of Sciences*, *115*(13), 3302-3307.

Duclos, R., & Barasch, A. (2014). Prosocial behavior in intergroup relations: How donor self-construal and recipient group-membership shape generosity. *Journal of Consumer Research*, *41*(1), 93-108.

Dunham, Y. (2013). Balanced identity in the minimal groups paradigm. *PLoS One*, *8*(12), e84205.

Dunn, T. L., Inzlicht, M., & Risko, E. F. (2019). Anticipating cognitive effort: roles of perceived error-likelihood and time demands. *Psychological Research*, *83*(5), 1033-1056.

Durrheim, K., Quayle, M., Tredoux, C. G., Titlestad, K., & Tooke, L. (2016). Investigating the evolution of ingroup favoritism using a minimal group interaction paradigm: The effects of inter-and intragroup interdependence. *PloS one*, *11*(11), e0165974.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., ... & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of autism and developmental disorders*, *36*(5), 623-636.

Eagly, A. H. (2009). The his and hers of prosocial behavior: an examination of the social psychology of gender. *American psychologist*, *64*(8), 644.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871-878.

Eisenberg, N., Cumberland, A., Guthrie, I. K., Murphy, B. C., & Shepard, S. A. (2005). Age

changes in prosocial responding and moral reasoning in adolescence and early adulthood. *Journal of research on adolescence*, *15*(3), 235-260.

Emswiller, T., Deaux, K., & Willits, J. E. (1971). Similarity, Sex, and Requests for Small Favors. *Journal of Applied Social Psychology*, *1*(3), 284-291.

Engel, C. (2011). Dictator games: A meta study. *Experimental economics*, *14*(4), 583-610.

Epley, N., & Caruso, E. M. (2012). Perspective taking: Misstepping into others' shoes. In *Handbook of imagination and mental simulation* (pp. 295-309). Psychology Press.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, *87*(3), 327.

Eres, R., & Molenberghs, P. (2013). The influence of group membership on the neural correlates involved in empathy. *Frontiers in human neuroscience*, *7*, 176.

Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in behavioral neuroscience*, *9*, 15.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of experimental social psychology*, *79*, 200-216.

Everett, J. A., & Onu, D. (2013). Intergroup contact theory: Past, present, and future. *The Inquisitive Mind*, *2*(17), 1-6.

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of personality and social psychology*, *114*(4), 547.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The*

quarterly journal of economics, *114*(3), 817-868.

Fehr, E., & Schmidt, K. M. (2010). On inequity aversion: A reply to Binmore and Shaked. *Journal of economic behavior & organization*, *73*(1), 101-108.

Ferguson, A. M., & Inzlicht, M. (2022). Reliability of the empathy selection task, a novel behavioral measure of empathy avoidance. *Behavior Research Methods*, 1-14.

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

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*(2), 244.

Fischer, I. (2009). Friend or foe: subjective expected relative similarity as a determinant of cooperation. *Journal of Experimental Psychology: General*, *138*(3), 341.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2018). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition* (pp. 162-214). Routledge.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. Mcgraw-Hill Book Company.

Francis, Z., Depow, G., & Inzlicht, M. (2021). Do early birds share their worms? How prosocial behaviour and empathy vary across the day. *Journal of Research in Personality*, *90*, 104055.

Frith, C., & Frith, U. (2005). Theory of mind. *Current biology*, *15*(17), R644-R645.

Gabay, A. S., & Apps, M. A. (2021). Foraging optimally in social neuroscience: computations and methodological considerations. *Social Cognitive and Affective Neuroscience*, *16*(8),

782-794.

Gaesser, B., Keeler, K., & Young, L. (2018). Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind. *Cognition*, *171*, 180-193.

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological science*, *19*(4), 378-384.

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological science*, *17*(12), 1068-1074.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493-501.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive psychology*, *59*(2), 154-179.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of experimental psychology: Human perception and performance*, *37*(3), 669.

Gao, L., Wang, Z., Pansini, R., Li, Y. T., & Wang, R. W. (2015). Collective punishment is more effective than collective reward for promoting cooperation. *Scientific reports*, *5*(1), 1-12.

Gavreliuc, A., Gavreliuc, D., & Semenescu, A. (2021). Beyond the façade of generosity—Regional stereotypes within the same national culture influence prosocial behaviors. *Plos one*, *16*(5), e0250125.

Geertz, C. (1973). *The interpretation of cultures*. Basic books.

Geertz, C. (1983). *Local Knowledge: Further Essays in Interpretive Anthropology*. Basic books.

Geertz, C. (1988). *Works and lives: The anthropologist as author*. Stanford University Press.

Gershman, S. J. (2017). Dopamine, inference, and uncertainty. *Neural Computation*, *29*(12), 3311-3326.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277.

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, *29*(5), 460-466.

Gershman, S., Zhou, J., & Kommers, C. (2017). Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, *29*(12), 2103-2113.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731-1744.

Gillespie, S. M., Mitchell, I. J., Johnson, I., Dawson, E., & Beech, A. R. (2013). Exaggerated intergroup bias in economical decision making games: differential effects of primary and secondary psychopathic traits. *PloS one*, *8*(8), e69565.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Guassi Moreira, J. F., Tashjian, S. M., Galván, A., & Silvers, J. A. (2020). Is social decision making for close others consistent across domains and within individuals?. *Journal of Experimental Psychology: General*, *149*(8), 1509.

Guassi Moreira, J. F., Tashjian, S. M., Galván, A., & Silvers, J. A. (2021). Computational and motivational mechanisms of human social decision making involving close others. *Journal of experimental social psychology*, *93*, 104086.

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: neural evidence of an empathy gap. *Social cognitive and affective neuroscience*, *7*(5), 596-603.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), 367-388.

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.

Hackel, L. M., Zaki, J., & Van Bavel, J. J. (2017). Social identity shapes social valuation: evidence from prosocial behavior and vicarious reward. *Social cognitive and affective neuroscience*, *12*(8), 1219-1228.

Halevy, N., Weisel, O., & Bornstein, G. (2012). "In-group love" and "out-group hate" in repeated interaction between groups. *Journal of Behavioral Decision Making*, *25*(2), 188-195.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me= bad: Infants prefer those who harm dissimilar others. *Psychological science*, *24*(4), 589-594.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., ... & Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature neuroscience*, *19*(1), 117-126.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, *7*(1), 17-52.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*(2), 129-154.

Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child development*, *66*(3), 843-855.

Happé, F., Cook, J. L., & Bird, G. (2017). The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual review of psychology*, *68*(1), 243-267.

Harmon-Jones, E., & van Honk, J. (2012). Introduction to a special issue on the neuroscience of motivation and emotion. *Motivation and Emotion*, *36*(1), 1-3.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological science*, *17*(10), 847-853.

Hartmann, M. N., Hager, O. M., Reimann, A. V., Chumbley, J. R., Kirschner, M., Seifritz, E., ... & Kaiser, S. (2015). Apathy but not diminished expression in schizophrenia is associated with discounting of monetary rewards by physical effort. *Schizophrenia bulletin*, *41*(2), 503-512.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243-259.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

Hepach, R. (2017). Prosocial arousal in children. *Child Development Perspectives*, *11*(1), 50-55.

Hepach, R., Hedley, D., & Nuske, H. J. (2020). Prosocial attention in children with and without autism spectrum disorder: Dissociation between anticipatory gaze and internal arousal. *Journal of Abnormal Child Psychology*, *48*(4), 589-605.

Hepach, R., Vaish, A., Grossmann, T., & Tomasello, M. (2016). Young children want to see others get the help they need. *Child Development*, *87*(6), 1703-1714.

Hepach, R., Vaish, A., & Tomasello, M. (2013). A new look at children's prosocial

motivation. *Infancy*, *18*(1), 67-90.

Hepach, R., Vaish, A., & Tomasello, M. (2017). Children's intrinsic motivation to provide help themselves after accidentally harming others. *Child development*, *88*(4), 1251-1264.

Hepach, R., Vaish, A., Kano, F., Albiach-Serrano, A., Benziad, L., Call, J., & Tomasello, M. (2021). Chimpanzees'(Pan troglodytes) internal arousal remains elevated if they cannot themselves help a conspecific. *Journal of Comparative Psychology*, *135*(2), 196.

Hepach, R., & Warneken, F. (2018). Editorial overview: Early development of prosocial behavior: Revealing the foundation of human prosociality. *Curr Opin Psychol*, *20*, iv-viii.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology*, *53*(1), 575-604.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*(2), 131-143.

Hillman, K. L., & Bilkey, D. K. (2010). Neurons in the rat anterior cingulate cortex dynamically encode cost–benefit in a spatial decision-making task. *Journal of Neuroscience*, *30*(22), 7705-7713.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with Theory of Mind. *Trends in Cognitive Sciences*.

Hobson, N. M., & Inzlicht, M. (2016). The mere presence of an outgroup member disrupts the brain's feedback-monitoring system. *Social Cognitive and Affective Neuroscience*, *11*(11), 1698-1706.

Hodges, S. D., Denning, K. R., & Lieber, S. (2018). Perspective taking: motivation and impediment to shared reality. *Current opinion in psychology*, *23*, 104-108.

Houlihan, S. D., Tenenbaum, J. B., & Saxe, R. (2021). Linking models of Theory of Mind and

measures of human brain activity. In *The Neural Basis of Mentalizing* (pp. 209-235). Springer, Cham.

Hosokawa, T., Kennerley, S. W., Sloan, J., & Wallis, J. D. (2013). Single-neuron mechanisms underlying cost-benefit analysis in frontal cortex. *Journal of Neuroscience*, *33*(44), 17385-17397.

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., & Hodges, J. R. (2013). Validation of the Addenbrooke's Cognitive Examination III in frontotemporal dementia and Alzheimer's disease. *Dementia and geriatric cognitive disorders*, *36*(3-4), 242-250.

Hughes, B. L., Ambady, N., & Zaki, J. (2017). Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. *Social Cognitive and Affective Neuroscience*, *12*(3), 372-381.

Hull, C. L. (1943). Principles of behavior: An introduction to behavior theory.

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental psychology*, *52*(8), 1192.

Inzlicht, M., & Hutcherson, C. A. (2017). Psychology: People work less hard for others. *Nature Human Behaviour*, *1*(7), 1-2.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in cognitive sciences*, *22*(4), 337-349.

Jaswal, V. K., & Akhtar, N. (2019). Being versus appearing socially uninterested: Challenging assumptions about social motivation in autism. *Behavioral and Brain Sciences*, *42*.

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological science*, *17*(4), 283-286.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J.

(2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review*, *125*(2), 131.

Kardan, O., Adam, K. C., Mance, I., Churchill, N. W., Vogel, E. K., & Berman, M. G. (2020). Distinguishing cognitive effort and working memory load using scale-invariance and alpha suppression in EEG. *NeuroImage*, *211*, 116622.

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends in cognitive sciences*, *18*(4), 163-166.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25-41.

Klein-Flügge, M. C., Kennerley, S. W., Friston, K., & Bestmann, S. (2016). Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort-reward trade-off. *Journal of Neuroscience*, *36*(39), 10002-10015.

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *Journal of neuroscience*, *28*(31), 7837-7846.

Kok, L., Oosterbaan, V., Stoker, H., & Vyrastekova, J. (2020). In-group favouritism and social norms: Public goods experiments in Tanzania. *Journal of Behavioral and Experimental Economics*, *85*, 101509.

Kommers, C., Ustun, V., Demski A., & Rosenbloom, P. (2015). Hierarchical reasoning with distributed vector representations. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Konovalova, E., & Le Mens, G. (2020). An information sampling explanation for the in-group heterogeneity effect. *Psychological Review*, *127*(1), 47.

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature human behaviour*, *2*(12), 899-908.

Kumle, L., Võ, M. L. H., & Draschkow, D. (2020). Simulation-based solutions for power analyses for mixed models considering by-subject and by-item variability. *Journal of Vision*, *20*(11), 696-696.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852-13857.

Krebs, D. (1975). Empathy and altruism. *Journal of Personality and Social psychology*, *32*(6), 1134.

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology*, *7*, 67-70.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Lang, M., Xygalatas, D., Kavanagh, C. M., Boccardi, N., Halberstadt, J., Jackson, C., ... & Gomez, A. (2021). Outgroup threat and the emergence of cohesive groups: A cross-cultural examination. *Group Processes & Intergroup Relations*, 13684302211016961.

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, *147*(12), 1881.

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, *18*(1), 18-66.

Le Heron, C., Apps, M. A. J., & Husain, M. (2018). The anatomy of apathy: a neurocognitive framework for amotivated behaviour. *Neuropsychologia*, *118*, 54-67.

Lee, V. K., Kranton, R. E., Conzo, P., & Huettel, S. A. (2021). The hidden cost of humanization: Individuating information reduces prosocial behavior toward in-group members. *Journal of Economic Psychology*, *86*, 102424.

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279-286.

Lees, J. (2022). Political violence and inaccurate metaperceptions. *Proceedings of the National Academy of Sciences*, *119*(19), e2204045119.

Leising, D., Grande, T., & Faber, R. (2009). The Toronto Alexithymia Scale (TAS-20): A measure of general psychological distress. *Journal of research in personality*, *43*(4), 707-710.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of personality and social psychology*, *68*(1), 151.

Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and social psychology bulletin*, *31*(4), 443-453.

Lewis, P. A., Birch, A., Hall, A., & Dunbar, R. I. (2017). Higher order intentionality tasks are cognitively more demanding. *Social cognitive and affective neuroscience*, *12*(7), 1063-1071.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551-556.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of personality and social psychology*, *57*(2), 165.

Livingstone, A. G., Afyouni, A., Vu, N., Bedford, S. L., Kapantai, I., Makanju, D., ... & Tapp, C.

(2022). You get us, so you like us: Feeling understood by an outgroup predicts more positive intergroup relations via perceived positive regard.

Livingstone, A. G., Fernández Rodríguez, L., & Rothers, A. (2020). “They just don't understand us”: The role of felt understanding in intergroup relations. *Journal of personality and social psychology*, *119*(3), 633.

Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a ‘social’brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*(10), 802-813.

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. (2017). Prosocial apathy for helping others when effort is required. *Nature human behaviour*, *1*(7), 1-10.

Lockwood, P. L., Abdurahman, A., Gabay, A. S., Drew, D., Tamm, M., Husain, M., & Apps, M. A. (2021). Aging increases prosocial motivation for effort. *Psychological Science*, *32*(5), 668-681.

Lopez-Gamundi, P., Yao, Y. W., Chong, T. T., Heekeren, H. R., Mas-Herrero, E., & Marco-Pallarés, J. (2021). The neural basis of effort valuation: A meta-analysis of functional magnetic resonance imaging studies. *Neuroscience & Biobehavioral Reviews*, *131*, 1275-1287.

López-Rodríguez, L., Cuadrado, I., & Navas, M. (2017). I will help you because we are similar: Quality of contact mediates the effect of perceived similarity on facilitative behaviour towards immigrants. *International Journal of Psychology*, *52*(4), 273-282.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, *30*(1), 1-13.

Malinowski, B. (1922). Argonauts of the Western Pacific.

Manohar, S. G., Chong, T. T. J., Apps, M. A., Batla, A., Stamelou, M., Jarman, P. R., ... & Husain, M. (2015). Reward pays the cost of noise reduction in motor and cognitive control. *Current Biology*, *25*(13), 1707-1716.

Margittai, Z., Strombach, T., Van Wingerden, M., Joëls, M., Schwabe, L., & Kalenscher, T. (2015). A friend in need: Time-dependent effects of stress on social discounting in men. *Hormones and Behavior*, *73*, 75-82.

McCutcheon, R. A., Abi-Dargham, A., & Howes, O. D. (2019). Schizophrenia, dopamine and the striatum: from biology to symptoms. *Trends in neurosciences*, *42*(3), 205-220.

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist*, *25*(2), 179-210.

McGuigan, S., Zhou, S. H., Brosnan, M. B., Thyagarajan, D., Bellgrove, M. A., & Chong, T. T. (2019). Dopamine restores cognitive motivation in Parkinson's disease. *Brain*, *142*(3), 719-732.

McClung, J. S., & Reicher, S. D. (2018). Representing other minds: Mental state reference is moderated by group membership. *Journal of Experimental Social Psychology*, *76*, 385-392.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415-444.

Meder, D., Herz, D. M., Rowe, J. B., Lehéricy, S., & Siebner, H. R. (2019). The role of dopamine in the brain-lessons learned from Parkinson's disease. *Neuroimage*, *190*, 79-93.

Mei, D., Zhang, W., & Yin, L. (2020). Neural responses of in-group "favoritism" and out-group "discrimination" toward moral behaviors. *Neuropsychologia*, *139*, 107375.

Meyer, M. L., & Lieberman, M. D. (2016). Social working memory training improves

perspective-taking accuracy. *Social Psychological and Personality Science*, *7*(4), 381-389.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., ... & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science*, *324*(5929), 900-900.

Montalan, B., Lelard, T., Godefroy, O., & Mouras, H. (2012). Behavioral investigation of the influence of social categorization on empathy for pain: a minimal group paradigm study. *Frontiers in psychology*, *3*, 389.

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, *25*(6), 889-922.

Moscatelli, S., & Rubini, M. (2017). Ingroup love and outgroup hate. In *Oxford research encyclopedia of communication*.

Muehlbacher, S., & Kirchler, E. (2009). Origin of endowments in public good games: The impact of effort on contributions. *Journal of neuroscience, psychology, and economics*, *2*(1), 59.

Müller, T., & Apps, M. A. (2019). Motivational fatigue: A neurocognitive framework for the impact of effortful exertion on subsequent motivation. *Neuropsychologia*, *123*, 141-151.

Müller, T., Husain, M., & Apps, M. A. (2022). Preferences for seeking effort or reward information bias the willingness to work.

Mummendey, A., & Otten, S. (1998). Positive–negative asymmetry in social discrimination. *European review of social psychology*, *9*(1), 107-143.

Mussweiler, T., & Ockenfels, A. (2013). Similarity increases altruistic punishment in humans. *Proceedings of the National Academy of Sciences*, *110*(48), 19318-19323.

Navarro, E. (2022). What is theory of mind? A psychometric study of theory of mind and intelligence. *Cognitive Psychology*, *136*, 101495.

Navarro, E., & Conway, A. R. (2021). Adult bilinguals outperform monolinguals in theory of mind. *Quarterly Journal of Experimental Psychology*, *74*(11), 1841-1851.

Navarro, E., Goring, S. A., & Conway, A. R. (2021). The relationship between theory of mind and intelligence: A formative g approach. *Journal of Intelligence*, *9*(1), 11.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520.

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of consumer psychology*, *22*(3), 453-460.

Novak, D. W., & Lerner, M. J. (1968). Rejection as a consequence of perceived similarity. *Journal of Personality and Social Psychology*, *9*(2p1), 147.

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*(5485), 1773-1775.

Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of abnormal psychology*, *125*(6), 818.

Ostaszewski, P., & Osiński, J. T. (2011). Social discounting of monetary rewards: The effects of amount and social relationship. *European Psychologist*, *16*(3), 220.

Oveis, C., Horberg, E. J., & Keltner, D. (2010). Compassion, pride, and social intuitions of self-other similarity. *Journal of personality and social psychology*, *98*(4), 618.

Packer, M. J., & Moreno-Dulcey, F. A. (2022). Theory of puppets?: A critique of the use of

puppets as stimulus materials in psychological research with young children. *Cognitive Development*, *61*, 101146.

Park, J. H., & Schaller, M. (2005). Does attitude similarity serve as a heuristic cue for kinship? Evidence of an implicit cognitive association. *Evolution and Human Behavior*, *26*(2), 158-170.

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition.

Pech, G. P., Gishoma, D., & Caspar, E. (2022). A novel EEG-based paradigm to measure intergroup prosociality. An intergenerational study in the aftermath of the genocide in Rwanda.

Pfattheicher, S., Nielsen, Y. A., & Thielmann, I. (2022). Prosocial behavior and altruism: A review of concepts and definitions. *Current opinion in psychology*, *44*, 124-129.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, *23*(12), 1026-1040.

Pinter, B., & Greenwald, A. G. (2011). A comparison of minimal group induction procedures. *Group Processes & Intergroup Relations*, *14*(1), 81-98.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied cognitive psychology*, *19*(3), 291-312.

Pollerhoff, L., Stietz, J., Depow, G. J., Inzlicht, M., Kanske, P., Li, S. C., & Reiter, A. M. (2022). Investigating adult age differences in real-life empathy, prosociality, and well-being using experience sampling. *Scientific reports*, *12*(1), 1-15.

Pooresmaeili, A., Wannig, A., & Dolan, R. J. (2015). Receipt of reward leads to altered estimation of effort. *Proceedings of the National Academy of Sciences*, *112*(43), 13407-13410.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences*, *1*(4), 515-526.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality*, *41*(1), 203-212.

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of personality assessment*, *93*(1), 84-95.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*(1), 52-73.

Romano, A., Sutter, M., Liu, J. H., Yamagishi, T., & Balliet, D. (2021). National parochialism is ubiquitous across 42 nations around the world. *Nature Communications*, *12*(1), 1-8.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173-220). Academic Press.

Ruggeri, K., Panin, A., Vdovic, M., Većkalov, B., Abdul-Salaam, N., Achterberg, J., ... & Toscano, F. (2022). The globalizability of temporal discounting. *Nature Human Behaviour*, 1-12.

Rushton, J. P. (1989). Genetic similarity, human altruism, and group selection. *Behavioral and Brain sciences*, *12*(3), 503-518.

Sacks, O. (1973). *Awakenings*. Pan Macmillan.

Salamone, J. D., & Correa, M. (2002). Motivational views of reinforcement: implications for

understanding the behavioral functions of nucleus accumbens dopamine. *Behavioural brain research*, *137*(1-2), 3-25.

Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2015). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic bulletin & review*, *22*(3), 844-849.

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain research*, *1079*(1), 86-97.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences*, *19*(2), 65-72.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, *55*, 7-14.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., ... & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psych. Bulletin*, *147*(3), 293.

Secord, P. F., & Backman, C. W. (1964). Interpersonal congruency, perceived similarity, and friendship. *Sociometry*, 115-127.

Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders. *The Neuroscientist*, *18*(2), 108-113.

Shaw, A., Choshen-Hillel, S., & Caruso, E. M. (2016). The development of inequity aversion: Understanding when (and why) people give others the bigger piece of the pie. *Psychological Science*, *27*(10), 1352-1359.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*(1), 99-124.

Shenhav, A., Prater Fahey, M., & Grahek, I. (2021). Decomposing the motivation to exert mental effort. *Current Directions in Psychological Science*, *30*(4), 307-314.

Sherif, M., Harvey, B., White, J., Hood, W., Sherif, C. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment.*

Shweder, R. A. (1991). *Thinking through cultures: Expeditions in cultural psychology*. Harvard University Press.

Simon, B., Mlicki, P., Johnston, L., Caetano, A., Warowicki, M., Van Knippenberg, A., & Deridder, R. (1990). The effects of ingroup and outgroup homogeneity on ingroup favouritism, stereotyping and overestimation of relative ingroup size. *European Journal of Social Psychology*, *20*(6), 519-523.

Simpson, E. H., & Balsam, P. D. (2015). The behavioral neuroscience of motivation: an overview of concepts, measures, and translational applications. *Behavioral neuroscience of motivation*, 1-12.

Simpson, A. J., & Todd, A. R. (2017). Intergroup visual perspective-taking: Shared group membership impairs self-perspective inhibition but may facilitate perspective calculation. *Cognition*, *166*, 371-381.

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive

psychology. *Cognition*, *217*, 104890.

Stengelin, R., Hepach, R., & Haun, D. (2019). Being observed increases overimitation in three diverse cultures. *Developmental Psychology*, *55*(12), 2630.

Sutton, R. S., & Barto, A. G. (1998). Introduction to reinforcement learning.

Suzuki, S., Lawlor, V. M., Cooper, J. A., Arulpragasam, A. R., & Treadway, M. T. (2021). Distinct regions of the striatum underlying effort, movement initiation and effort discounting. *Nature human behaviour*, *5*(3), 378-388.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, *1*(2), 149-178.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984-16.

Tamir, D. I., & Hughes, B. L. (2018). Social rewards: from basic social building blocks to complex social behavior. *Perspectives on Psychological Science*, *13*(6), 700-717.

Tanaka, S., O'Doherty, J. P., & Sakagami, M. (2019). The cost of obtaining rewards enhances the reward prediction error signal of midbrain dopamine neurons. *Nature communications*, *10*(1), 1-13.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of economic behavior & organization*, *1*(1), 39-60.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of economic perspectives*, *2*(4), 195-206.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*. *94*(6), 1395–1415.

Thornton, M. A., & Tamir, D. I. (2020). People represent mental states in terms of rationality, social impact, and valence: Validating the 3d Mind Model. *Cortex*, *125*, 44-59.

Thornton, M. A., & Tamir, D. I. (2022). Six dimensions describe action understanding: The ACT-FASTaxonomy. *Journal of personality and social psychology*, *122*(4), 577.

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. *Nature communications*, *10*(1), 1-9.

Thornton, M. A., Wolf, S., Reilly, B. J., Slingerland, E. G., & Tamir, D. I. (2022). The 3d Mind Model characterizes how people understand mental states across modern and historical cultures. *Affective Science*, *3*(1), 93-104.

Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard university press.

Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., & Zald, D. H. (2009). Worth the 'EEfRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. *PloS one*, *4*(8), e6598.

Treadway, M. T., Bossaller, N. A., Shelton, R. C., & Zald, D. H. (2012). Effort-based decision-making in major depressive disorder: a translational model of motivational anhedonia. *Journal of abnormal psychology*, *121*(3), 553.

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European journal of social psychology*, *9*(2), 187-204.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Blackwell.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of business*, *59*(4), S251-S278.

Umemoto, A., Lin, H., & Inzlicht, M. (2021). Cost-benefit analysis in physical effort expenditure: An electrophysiological registered report.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161-204.

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. *Journal of Neuroscience*, *35*(20), 7866-7877.

Vekaria, K. M., Brethel-Haurwitz, K. M., Cardinale, E. M., Stoycos, S. A., & Marsh, A. A. (2017). Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour*, *1*(5), 1-7.

Walton, M. E., Bannerman, D. M., Alterescu, K., & Rushworth, M. F. (2003). Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. *Journal of Neuroscience*, *23*(16), 6475-6479.

Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M., & Rushworth, M. F. (2006). Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural networks*, *19*(8), 1302-1314.

Weisz, E., Ong, D. C., Carlson, R. W., & Zaki, J. (2021). Building empathy through motivation-based interventions. *Emotion*, *21*(5), 990.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395-415.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one*, *8*(7), e68210.

Whitton, A. E., Reinen, J. M., Slifstein, M., Ang, Y. S., McGrath, P. J., Iosifescu, D. V., ... &

Schneier, F. R. (2020). Baseline reward processing and ventrostriatal dopamine function are associated with pramipexole response in depression. *Brain*, *143*(2), 701-710.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences*, *38*, 49-56.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Wu, R., Ferguson, A., & Inzlicht, M. (2021). Do humans prefer cognitive effort over doing nothing?

Yee, D. M., & Braver, T. S. (2018). Interactions of motivation and cognitive control. *Current opinion in behavioral sciences*, *19*, 83-90.

Yee, D. M., Leng, X., Shenhav, A., & Braver, T. S. (2021). Aversive motivation and cognitive control. *Neuroscience & Biobehavioral Reviews*.

You, S., & Robert, L. P. (2018, March). Human–robot similarity and willingness to work with a robotic co-worker. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 251-260). IEEE.

Zaki, J. (2014). Empathy: a motivated account. *Psychological bulletin*, *140*(6), 1608.

Zhao, H., Yang, Z., Chen, Z., & Zhang, X. (2015, January). Automatic chinese personality recognition based on prosodic features. In *International Conference on Multimedia Modeling* (pp. 180-190). Springer, Cham.

Zipf, G. K. (1949). Human behavior and the principle of least effort: an introd. to human ecology.