Computational, psychological and neural mechanisms of prosocial motivation



Luis Sebastian Contreras Huerta

Wolfson College

Department of Experimental Psychology

Hilary Term, 2021

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

Prosocial behaviours are essential for healthy social interactions, with established impacts on life satisfaction, mental and physical health, and economic success. Yet, how willing people are to act prosocially varies considerably between individuals and across contexts. When making decisions that can result in others' pain, people can be highly prosocial, and will typically not profit from other's harm. In contrast, people are much less motivated to exert effort to benefit others compared to themselves. However, the psychological and neural mechanisms of these behaviours remain largely unknown. In this thesis, I examine the affective, cognitive, and neural processes that underpin harm aversion and effortful prosocial behaviours and their variability between people. First, I propose a theoretical framework where individual differences in prosocial behaviour across domains are due to idiosyncratic sensitivities to personal costs and others' benefits, giving a conceptual and methodological overview on what factors might be hypothesised to influence cost-benefit prosocial evaluations. Second, I show that harm aversion and prosocial effortful actions share variance despite their differences at group level, and are linked by responses in the anterior cingulate cortex (ACC) and the anterior insula (AI) to others' benefits. Third, I identify an affective profile characterised by high empathy and low apathy as crucial for predicting prosociality. Fourth, I find that people who are highly hypocritical, i.e. fail to live up to their moral standards, are more sensitive to effort costs, suggesting motivational aspects in moral hypocrisy. Fifth, I disentangle the role of ACC and AI in others' rewards representations, with the dorsal AI linked to interoceptive processes and the ACC linked to motivational aspects of others' rewards. Finally, I propose a new experimental framework to investigate prosocial motivation, examining how the alternative opportunities in the environment shape prosociality. In sum, in this thesis I show that being motivated to benefit others depends on cost-benefit evaluations, which are modulated by affective traits and interoceptive sensibility, with crucial, but differential, roles of ACC and AI in prosocial behaviour.

Acknowledgements

It has been a long and beautiful ride which has led me to writing these words, closing an important chapter of both my personal and academic life. This achievement is dedicated to all the people who have inspired and helped me to be here writing these words, finalising my doctorate in Oxford.

First, I would like to thank my supervisors for their support and guidance in all these years. I would not be writing these words if prof. Molly Crockett had not supported my application in the first place. Thank you, Molly, for guiding me into my first steps as a PhD student, for being supportive and helping me overcome obstacles on the way, and for inspiring me to become a better scientist, rigorous in the methods, strong in theoretical foundations, and creative and passionate in finding and solving relevant social questions. I would also like to thank prof. Geoff Bird, who always found the right words when I felt lost trying to understand results which were not those I expected, always providing new perspectives for interpretation. Thanks, Geoff, for your enthusiasm, your encouragement and support, and thanks for those beers and good moments we shared together throughout these years! Last, but not least, I would like to thank Dr. Matthew Apps, who was my main supervisor for most of my PhD. Matt, thanks for your scientific advice, for those discussions about social theory and neuroscience, for having the patience to explain complex concepts in simple words, for hearing and encouraging my ideas, for reading my drafts countless times, and for being supportive in difficult moments such as this pandemic. Under your supervision, I found new perspectives to approach the social phenomenon, which made me feel similar excitements when I took my first steps in science many years ago. A big part of the scientist I am today I owe to you. Also, thanks for all those moments in the pub, talking about politics, football, and life in general, always with that sarcastic point of view of yours. I find in you not only a mentor, but also a friend, and I am deeply grateful for that.

Different people have also contributed to me being able to write these words today. Thanks to Patricia Lockwood for the scientific and non-scientific discussions, for her support and advice, and for the past, present, and future collaborations. To the Crockett Lab members for being such special people. Special thanks to Hongbo, who always had illuminating scientific and life advice, and Anne-Marie, with whom I shared light and dark days on this PhD trip. Thanks also to the MSN lab members, who were like a family. Special thanks go to Anthony, Andrea and Tanja. We were the very first members of the MSN lab even before it had a real name! Thanks for the technical advice, but more importantly, for the bond we built together.

These words would not have been written without the support of my dear friends, the 'banned gang' from Wolfson College. So many memories, conversations, dances, dinners, drinks... I will always be grateful for having met so many smart, kind, and weird people – Eduardo, Thea, Catarina, Gigi, Alba, Marco, Mesi, Pinar, Giuseppe, Tommaso, Yannis, Lanbo, Mika, Giovanni, Alexis, Sabin and Mitko, I really hope our paths cross again. Special thanks to Edu, with whom I shared not only music and adventures, but also brotherhood. Thanks also to my dear Chilean family, Camila, Margarita, Gonzalo, Camilo, Felipe, Daniela, and little Pascal, for making me feel at home, especially in these last couple of years in which our country is transforming into something new.

I would also like to thank my previous mentors that guided my early steps into science. Thank you prof. Adrian Palacios and Errol Dennis for showing me that the mysteries of the mind could be researched through science, to prof. Ross Cunnington for shaping my interest into social neuroscience, and prof. David Huepe for his support and friendship in my first job as an academic. Importantly, thanks to the Government of Chile, that has funded all my academic career.

These words would have been impossible to write without Cristiana. She not only read and listened to all the themes of this thesis with genuine interest, always providing her valuable critical thoughts, but also sheltered me (metaphorically and literally!) when I

needed it throughout this journey. Cristiana, thank you for your immense support and for encouraging me in every endeavour I planned. Also, thanks for bringing a second family to me, who made me feel like a son, brother and cousin, and who gave me unconditional support. You have all made my life in the UK a lovely experience.

Due to the pandemic, I find myself writing these words where everything started, in a small house at the top of a hill in Valparaiso, Chile. Statistically, I am a complete outlier compared with a typical graduate from Oxford. If statistical predictions were always true, I should have never been writing these words. If I am doing it, it is because my mother and father created a home for this to happen. Thank you papis, Patricia Huerta Cabrera and Luis Contreras Arredondo, for all the love, for all that immense love that made all this possible, for all the sacrifices that you have made throughout all these years to make your sons fly, for creating opportunities that were denied to too many children, including yourselves. And thank you, Rodrigo, for illuminating my paths since we were children. Thank you for your wisdom, for your advice, for opening the world to my eyes and for making me conceive another life. Without you, family, I would not be writing these last words of this thesis - *Pessimism of the intellect, optimism of the will*.

Statement of Contribution

Science is a collaborative human endeavour. As such, the present thesis is the result of the coordinated work of many researchers, mainly members of my supervisors' laboratories. The theoretical ideas proposed in Chapter 1 were partially based on a review paper I published last year together with Dr. Andrea Pisauro and my supervisor Dr. Matthew Apps (https://doi.org/10.1016/j.neubiorev.2020.08.003). Likewise, Chapter 4 was modified from our last year publication in *Emotion* (http://dx.doi.org/10.1037/emo0000813). All co-authors have agreed to include the material in the thesis. For Chapters 3 to 6, the original scripts for the computational models were developed by Dr. Patricia Lockwood, and my supervisors Dr. Molly Crockett and Dr. Matthew Apps. Furthermore, the original script for the Marginal Value Theorem in Chapter 7 was developed by Dr. Nils Kollings and Dr. Campbell Le Heron. I used and modified these scripts for my own purposes for each chapter. Chapters 3 and 5 were designed in collaboration with my supervisors Dr. Molly Crockett and Dr. Matthew Apps, together with Dr. Hongbo Yu, post-doc member of the Crockett lab. The fMRI task in Chapter 6 was designed in collaboration with Dr. Geoffrey Bird, one of my supervisors, and Dr. Michel-Pierre Coll, a postdoc member of his lab, who also contributed to the analysis, writing part of the code used for multivariate analysis. This thesis was developed under the supervision of Dr. Matthew Apps. I led data collection and analysis for all empirical chapters.

Table of Contents

Abstract ii
Acknowledgements iii
Statement of Contribution vi
Table of Contents vii
1. Chapter 1. General Introduction 1
1.1 Introduction 1
1.2 Theories of prosocial behaviour and its variability 4
1.2.1 Prosocial behaviour in economic games 4
1.2.2 Moral principles and moral dilemmas 6
1.2.3 Affective theories of prosocial behaviour 10
1.2.4 Summary and critique of theories of prosocial motivation 18
1.3 Prosocial motivations as a cost-benefit evaluation 21
1.3.1 Prosocial Effort 24
1.3.2 Harm Aversion 27
1.3.3 Distinctive and common aspects of prosocial behaviour in different domains. 29
1.3.4 Dynamic costs in goal-directed behaviour: Opportunity costs. 31
1.4 Neural underpinnings of individual differences in prosociality 33
1.4.1 Anatomical aspects of ACC and AI 34
1.4.2 ACC and AI signal others' outcomes across domains 36
1.4.3 ACC and AI weight others' outcomes in prosocial behaviour 38
1.4.4 Functional distinction between ACC and AI 41
1.4.5 ACC and AI as a hub evaluating prosocial benefits and costs 43
1.5 The present thesis 44
2. Chapter 2. General Methods 45
2.1 Summary 45
2.2 General Procedure 45
2.2.1 Participants 45
2.2.2 Apparatus 46
2.2.3 Social manipulation in lab-based studies: Role Assignment Procedure 47
2.3 Measuring two manifestations of prosocial motivation 51
2.3.1 Prosocial Effort Task 51
2.3.2 Harm Aversion Task 56
2.4 Analysis of behavioural data 60
2.4.1 Computational Modelling 61

2.4.2 Mixed Effects Models 66
3. Chapter 3. Common mechanisms underlie prosocial behaviour in different domains .. 74
3.1 Summary 74
3.2 Introduction 75
3.3 Methods 79
3.3.1 Participants 79
3.3.2 Procedure 80
3.3.3 Analysis of behavioural data 82
3.3.4 fMRI acquisition and preprocessing 85
3.3.5 fMRI analysis 86
3.4 Results 90
3.4.1 Hyperaltruism and prosocial apathy effects are replicated. 90
3.4.2 Prosocial behaviour is correlated across domains 95
3.4.3 fMRI results 98
3.5 Discussion 102
4. Chapter 4. Transdiagnostic markers of affective sensitivity are associated with prosocial behaviour in different domains 106
4.1 Summary 106
4.2 Introduction 107
4.3 Methods 110
4.3.1 Participants 110
4.3.2 Procedure 111
4.3.3 Exclusion criteria and quality assurance 111
4.3.4 Behavioural Tasks 113
4.3.5 Computational Modelling 116
4.3.6 Self-Report scales 118
4.3.7 Statistical analysis of behavioural data. 119
4.4 Results 122
4.4.1 Study 1: Hyperaltruism and prosocial apathy effects are replicated in online samples. 122
4.4.2 Study 2: Replicating Hyperaltruism and Prosocial Apathy effects 123
4.4.3 Prosocial behaviours in both tasks are correlated 126
4.4.4 Transdiagnostic relationship between affective and psychiatric traits and prosocial dispositions 129
4.5 Discussion 133
5. Chapter 5. Motivational aspects of moral hypocrisy 138
5.1 Summary 138
5.2 Introduction 139

5.3 Methods 142
5.3.1 Participants 142
5.3.2 Harm Aversion Task. 142
5.3.3 Moral Judgement Task 143
5.3.4 Prosocial Effort Task 143
5.3.5 Analysis of the behavioural data 144
5.4 Results 148
5.4.1 Moral hypocrisy is common and varies across participants 148
5.4.2 Moral hypocrisy is associated with prosocial apathy 149
5.4.3 Moral hypocrisy is associated with trial performance. 154
5.5 Discussion 161
6. Chapter 6. Neural self/other reward representations are linked to interoception and prosocial motivation 166
6.1 Summary 166
6.2 Introduction 167
6.3 Methods 170
6.3.1 Participants 170
6.3.2 Procedure 171
6.3.3 Analysis of behavioural data 178
6.3.4 fMRI acquisition and preprocessing 181
6.3.5 fMRI analyses 183
6.4 Results 187
6.4.1 Behavioural results: Interoceptive sensibility facilitates prosocial motivation. 187
6.4.2 fMRI results: Different roles of ACCg and RdAI in reward processing. 192
6.5. Discussion 194
7. Chapter 7. Opportunity cost shapes prosocial behaviour: Prosocial Foraging 200
7.1 Summary 200
7.2 Introduction 201
7.3 Methods 205
7.3.1 Participants 205
7.3.2 General Procedure 206
7.3.3 The Social Foraging Task. 207
7.3.4 Optimal predictions using the Marginal Value Theorem 213
7.3.5 Statistical Analyses 215
7.4 Results 216
7.4.1 Study 1 216
7.4.2 Study 2 223
7.5 Discussion 227

8. Chapter 8. General Discussion 233
8.1 Introduction 233
8.2 Mechanisms underlying prosocial motivation 233
8.3 Towards an integrated model of individual differences in prosocial motivation . 238
8.3.1 Previous approaches do not focus on individual differences 238
8.3.2 Individual differences in cost-benefit processing shape prosocial motivation 239
8.3.3 Neural structures at the base of common prosocial disposition..... 242
8.3.4 Dynamic cost-benefit decisions in prosocial behaviour 246
8.4 Concluding Remarks..... 247
8.4.1 Limitations and Future Directions 247
8.4.2 General Conclusion 251
References..... 252
Appendix A. Supplementary tables for whole brain analyses in Chapter 3..... 311
Appendix B. Supplementary tables for whole brain analyses in Chapter 6..... 313

Chapter 1

General Introduction

Sections of this chapter are included in Contreras-Huerta et al., (2020), *Neuroscience and Biobehavioural Reviews*

1.1 Introduction

Prosocial behaviour has been defined as any action that benefits others at a personal cost for the agent (de Waal, 2008; Fehr & Fischbacher, 2003; Kurzban, Burton-Chellew, & West, 2015, also see Padilla-Walker & Carlo, 2015; Schroeder & Graziano, 2015; for further discussions about the definitions of prosociality). From holding a door open for a stranger, to donating a kidney to save another person's life, a wide range of human behaviours are prosocial, manifested across cultures, from early childhood to elderly years. Prosocial behaviour is believed to have a fundamental role in social animals, promoting group cohesion and increasing likelihood of survival (Chudek & Henrich, 2011; de Waal, 2008; Kurzban, Duckworth, Kable, & Myers, 2013; Warneken, 2016). Prosocial tendencies predict increased physical and mental health, which is manifested in lower mortality risks not only for humans but also in other animals (Brown & Brown, 2015; Post, 2005; Stern, 2019). Being prosocial is therefore beneficial to the individual and to society.

Modern science has aimed to understand the factors underlying these behaviours from fields such as Anthropology, Biology, Economics and Psychology. This interdisciplinary endeavour has proven to be fruitful, providing scientific explanations on how these actions have evolved, and the psychological variables that facilitate prosocial behaviour. However, one of the biggest challenges for understanding prosocial behaviour is its high variability across people. People within and between cultures exhibit prosocial behaviours at different frequencies and with greater or lesser generosity. Scientific

theories have helped to illuminate some of the ultimate and proximate causes for prosocial behaviour, but not all of the mechanisms underlying people's prosocial decision-making have been revealed. Moreover, explanations of its high variability have not been forthcoming.

In humans a vast range of experimental paradigms have been used to understand prosocial behaviour (e.g. Batson, 2017; Fehr & Camerer, 2007; Greene, 2015; Haidt, 2007). Strikingly, in the vast majority of these experiments, high variability is observed in participants' prosocial tendencies. Likewise, levels of prosociality can vary from one task to another. For example, people are typically more prosocial in tasks where they need to prevent others' harm than in economic games. Thus, explanations about prosocial behaviour under one framework can be difficult to apply to another, given the high variability within and between contexts.

How can prosocial behaviour be operationalised to understand its variability? Research in decision-making in psychology and neuroscience has often placed emphasis on the value of benefits and costs that guide our behaviour. This has recently been extended to prosocial behaviour, where what is being given up by the agent, and what the benefit is to the other person have become the focus of research. Both the modality, such as pain vs money, and valence (obtaining a benefit vs avoiding a loss) of the personal cost, and of the social benefit, affect the willingness to help within and between people (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Engel, 2011; Hein, Silani, Preuschoff, Batson, & Singer, 2010; Lockwood, Hamonet et al., 2017). For example, a person can be highly prosocial in assisting someone who is suffering from a heart attack, but reluctant to spend time and effort in helping a stranger in moving into a new house. Still, both of these actions involve at least two common aspects – a benefit for another person, and a cost for the agent performing the action. In this thesis, I argue that understanding how prosocial an individual will be across contexts will only come by

studying the variability in the mechanisms underlying the valence and modality of different prosocial acts.

The roots of this view come from research on goal-directed behaviours (Niv, 2007; Walton, Rudebeck, Bannerman, & Rushworth, 2007). Decades of research has examined how people weigh up the costs of actions performed to obtain a benefit for oneself. This work has revealed that the degree which people weigh costs and benefits is crucial to understand variability in levels of daily actions and even for more severe deficits in goal-directed behaviour such as apathy in neurological disorders (Husain & Roiser, 2018; Le Heron, Apps., & Husain, 2018). Here I conceptualise prosocial behaviours as goal-directed behaviours. However, they differ from classical goal-directed actions as the immediate, direct benefit is not for oneself but for another person. How sensitive people are to costs and benefits may explain some of the variability found in prosocial behaviour.

In this thesis, I use a cost-benefit framework to examine individual differences in prosocial behaviours across different valences and modalities. I will define someone's willingness to do something prosocial as their level of '*prosocial motivation'*. Someone who is highly prosocially motivated will frequently and readily engage in acts that incur significant costs to obtain large benefits to others. Using this approach, I aimed to identify the psychological and neural mechanisms that underlie high levels of prosocial behaviour in different domains, illuminating common and distinctive factors underlying variability in prosociality. In the next section, I will outline many of the theories proposed for the antecedents of prosocial behaviour, and how variability within them may lead to variability in prosocial motivation – i.e. why some people are more or less motivated to help others.

1.2 Theories of prosocial behaviour and its variability

Prosocial motivation can vary considerably depending on the agent and the context. Characteristics of the beneficiary can modulate prosocial behaviour, e.g. people are more willing to help members of their kin or social group compared with strangers (Bernhard, Fischbacher, & Fehr, 2006; Hamilton, 1963; Hein et al., 2010; Rachlin & Jones, 2008). Furthermore, characteristics of the agent can also influence prosocial behaviour, with different traits either facilitating or obstructing prosocial behaviours (Hu, Hu, Li, & Zhou, 2021; Lockwood, et al., 2017; White, 2014). Finally, characteristics of the situation can affect motivation to act prosocially. Incurring costs to benefit a person who is not in need can be seen as supererogatory, and people might be less willing to help than in a situation where a person is under a high risk or pain (Crockett et al., 2014; Lockwood, Hamonet, et al., 2017; Volz, Welborn, Gobel, Gazzaniga, & Grafton, 2017). Different approaches and theories of prosocial behaviour have shed light on the factors that underlie these actions. In the following section, I will briefly describe the different research areas and theories which examine prosocial behaviour, highlight their key findings and how they may attempt to explain variability in prosocial behaviour between people.

1.2.1 Prosocial behaviour in economic games

Neoclassical economic theory considers people as rational agents whose decisions are made after a self-interested focused evaluation of costs and benefits, motivated exclusively by their material returns (Glimcher & Fehr, 2013). However, behavioural game paradigms have shown that humans usually deviate from an economic rationale, with most individuals showing prosocial motivation (Camerer, 2011; Fehr & Camerer, 2007; Glimcher & Fehr, 2013; Rilling & Sanfey, 2011). Such motivation has been manifested in different games used in behavioural economics studies, in which people are sensitive to

fairness (*ultimatum game*), cooperation (*prisoner dilemma*), trust and reciprocity (*trust game*), and other prosocial considerations even when it is not economically rational.

The games deployed in behavioural economics have proven useful for studying complex social interactions in an experimental setting, allowing the control and the manipulation of different variables that can affect social decision-making processes. There are, broadly speaking, two classes of games: one-shot games, where each player makes one decision that affects themself and another person, and iterative games, where both players' actions, in repeated interactions, impact upon the payoffs for each other. Iterative games can be used to probe human strategic decisions. However, in simple one-shot games, economically rational agents should maximise their returns, and thus not consider the other player. Yet, repeatedly behaviour in these games shows that people will perform actions to benefit even unrelated strangers (Camerer, 2011; Fehr & Camerer, 2007; Glimcher & Fehr, 2013; Rilling & Sanfey, 2011)

One of the most relevant games for understanding prosocial tendencies is the *dictator game* (DG). In the DG, the *dictator* receives a given amount of money and donates a desired part of it to the *recipient*. The game ends with that decision and the recipient has no chances of punishing the dictator if the amount received is considered too small, in contrast with the *ultimatum game* (UG), where, if the recipient rejects the offer, both players result in zero. Thus, even though in a DG the *dictator* has no incentives to share any portion of the received money, participants tend to donate between 20-30% of their endowment (Camerer, 2011; Engel, 2011; Rilling & Sanfey, 2011). Although people are seemingly willing to give, this also highlights that people value more their own rewards than others when there is no possibility of retaliation (Camerer, 2011; Fehr & Fischbacher, 2003; Rilling & Sanfey, 2011).

Studies have shown that different factors can modulate dictators' behaviour, including social distance, status of the recipient, and sense of ownership of the money provided (Camerer, 2011; Cherry, Frykblom, & Shogren, 2002; Eckel & Grossman, 1996; Hoffman,

McCabe, & Smith, 1996; Rilling & Sanfey, 2011), and thus there is high variability in DG donations. Although fruitful, classic economic games do not provide mechanistic models of this variability nor the psychological basis of it – in general, they either account for mathematical formalisations of strategic decisions, or provide descriptive accounts of how changing the payoffs increase or decrease people's generosity. One of the major barriers to understanding the psychology of behaviour in such economic games is a confounding between the social benefit and personal cost in the prosocial acts – there is a direct trade-off between the benefit to self and benefit to other. As such when a person donates in the dictator game, it is unclear whether they do so because they care about others' welfare, or because they value their own rewards little. Thus, self and other interests are merged in economic games, making it challenging to understand the underlying psychology of prosocial variability.

1.2.2 Moral principles and moral dilemmas

Moral principles are rules or virtues directed towards a universal and impartial good, usually internalised and reinforced through social norms, that commonly prescribe or prevent some actions (Batson, Ahmad, & Tsang, 2002; Eisenberg, VanSchyndel, & Spinrad, 2016; House, 2018). For example, 'do no harm' is a moral principle found across different cultures, and is reinforced in both legal and social contracts (Decety & Cowell, 2017; Gert, 2004; Gray, Young, & Waytz, 2012; Turiel, 2002). Moral psychology traditionally has relied on hypothetical moral dilemmas in the lab (Christensen & Gomila, 2012). One of the most classic dilemmas is the 'trolley problem' (Christensen & Gomila, 2012; Thomson, 1976), where a trolley has gone out of control and it will run over five track workers who are in the rails. In one variant of this dilemma (*switch*), the participant is asked whether they would be willing to flip a switch that deviates the trolley to another track where only one worker would die, saving the other five trackworkers. In another variant of this dilemma (*push*), participants are asked whether they would push a person

onto the tracks to stop the trolley, killing this person but saving the five workers. Is it morally right to kill one person to save five?

Empirical results of the trolley problem illustrate how variations in the nature of the action shape moral decisions. In the switch version, most participants are willing to save five people by killing one, whereas in the push version most people are reluctant to kill one person to save the trackworkers (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007). These results illustrate two major traditions in normative ethics – while in the switch version people predominantly follow *consequentialism*, judging moral actions according to their outcomes (i.e. saving five is better than one), in the push variant most people follow *deontology*, acting according to a set of rules that proscribe actions that are fundamentally immoral irrespective of their consequences (Crockett, 2013; Cushman, 2013). These two moral traditions have been framed in social cognitive psychology within a dual-system cognitive structure (Greene, 2009), or through habitual (deontology) and goal-directed (consequentialism) decisions (Crockett, 2013; Cushman, 2013). From these perspectives, deontological decisions are action-based decisions - harming others is associated with a negative value, regardless of the outcome, consistent with studies showing that actions that harm others, even without any real outcome, triggers negative feeling and distress and therefore are avoided (Cushman, Gray, Gaffey, & Mendes, 2012; Greene et al., 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Miller & Cushman, 2013). On the other hand, in consequentialist decisions the agent represents actions and outcomes, together with their value, to determine which decisions bring the best outcomes available (Cushman, 2013).

Do people act in real life as they report in a trolley problem? ‘Do-no-harm’ in this theory appears as a very strong moral principle. However, anecdotical evidence from everyday life suggests that people sometimes harm others to obtain a benefit themselves. Recent moral decision-making studies, where participants are required to trade-off self-

interest against harming others, have shown that, even though people show harm aversion towards others, they can also cause small harm for high rewards (FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; FeldmanHall, Dalgleish, & Mobbs, 2013). Furthermore, when harm aversion decisions become hypothetical, people are more prosocial than when decisions trigger real outcomes for self and others (FeldmanHall et al., 2012). These results suggest that people vary in the degree in which they are willing to sacrifice self-interest to avoid harm. Classic moral dilemmas might have difficulties to account for these individual differences in harm aversion – they typically involve a binary decision where the magnitude of the social benefit is not parametrically manipulated. Could prosocial behaviour in the trolley problem change if the number of deaths of not acting is increased? Do people have a 'threshold' for the cost that they are willing to accept to benefit others? Moreover, the social benefit (saving more lives) and the self cost (moral transgression) are not clearly disentangled – people could push the large person, or switch the flip, by the cost of their own moral integrity, or rather, they could be insensitive to this cost. Thus, although useful, existing moral dilemmas alone have not fully addressed the mechanisms underlying prosocial variability.

1.2.2.1 When moral actions fail: The case of moral hypocrisy

Moral principles not only guide the actions of an agent, but also how to judge the actions of other people – transgressors are blamed and can be socially (e.g. exclusion), institutionally (e.g. excommunication) and legally (e.g. imprisoned) punished, especially when active actions result in others' harm rather than omissions with equivalent immoral outcomes (Baron & Ritov, 2004; Cushman, Young, & Hauser, 2006; Hauser, Tonnaer, & Cima, 2009; Navarrete, McDonald, Mott, & Asher, 2012). However, people often fail to follow the rules, even though they can keep judging others harshly for similar deviations. This discrepancy between moral actions and judgements have been identified as a form

of moral hypocrisy (Alicke, Gordon, & Rose, 2013; Dover, 2019; Effron, O'Connor, Leroy, & Lucas, 2018; Howe & Monin, 2017; Laurent & Clark, 2019).

Philosophical approaches have proposed hypocrites as having different moral standards for self and other actions, which might imply deceptive intentions (Dover, 2019; Howe & Monin, 2017; Szabados & Soifer, 2004; Todd, 2019; Valdesolo & DeSteno, 2008; Wallace, 2010). For instance, a promiscuous husband who blames others for their extra-marital affairs might hide his true immoral nature through a moral façade, either to increase his reputation (social deceit), or self-esteem (self deceit). In any case, his moral principles might not be genuine, and he may not feel remorse when transgressing the norm (Szabados & Soifer, 2004; Wallace, 2010). Consistently, people commonly consider hypocrites as fundamentally immoral, even in the absence of information about hypocrites' intentions (Barden, Rucker, & Petty, 2005; Jordan, Sommers, Bloom, & Rand, 2017; O'Connor, Effron, & Lucas, 2020).

An alternative view of hypocrisy however proposes that it occurs due to a failure of the will, i.e. *akrasia* (Bartel, 2019; Batson & Thompson, 2001; Dover, 2019; Mele, 1989). In this approach, judgements toward others are not deceptive, and the immoral action is the product of self-interest overpowering sincere moral principles, skipping the cost of conforming to their genuine moral principle. This is consistent with social psychology studies showing that actions deviate from self-reported moral principles when self interest is in dispute (Batson, 2008; Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997; Batson, Thompson, & Chen, 2002; Batson & Thompson, 2001; Batson, Thompson, Seuferling, Whitney, & Strongman, 1999). In these studies, participants acted selfishly while admitting that this action was not the right thing to do. Thus, people could fall into the temptation, transgressing their moral standards.

Crucially, previous studies have shown that people vary in manifesting hypocritical behaviour (Batson et al., 2002, 1999; Carpenter & Marshall, 2009; Dong, van Prooijen, & van Lange, 2019; Lammers, Stapel, & Galinsky, 2010; Polman & Ruttan, 2012; Tang et

al., 2018; Valdesolo & DeSteno, 2008), but they do not address the potential psychological processes underlying it. Furthermore, research has not typically taken hypocrisy as a continuum, but a categorical phenomenon, not allowing people to vary in the degree in which they act hypocritically (Yu, Siegel, & Crockett, 2019). Moreover, the personal costs and social benefits of being moral are usually not disentangled nor their magnitude parametrised in measuring hypocrisy. Likewise, often hypocrisy is measured with people self-reporting moral principles, which are of course subject to social desirability effects. To measure hypocrisy, one needs to measure someone's judgements and their own actual behaviour in the same setting. Therefore, studies researching moral hypocrisy have failed in illuminating the psychological mechanisms underlying moral principles/actions discrepancy, which makes interpreting the results as supporting an akratic or deceptive model of hypocrisy challenging.

1.2.3 Affective theories of prosocial behaviour

1.2.3.1 Empathy as driver for prosocial behaviour

In contrast to action-based moral decisions, it has long been argued that prosocial behaviour can be driven by observing the emotional state of a person (Batson, 2017; Batson, Duncan, Ackerman, Buckley, & Birch, 1981; de Waal, 2008). Perceiving another person in need elicits an empathetic reaction in the observer that drives helping behaviour to alleviate the other person's distress. This is the central argument of what is has been called the *empathy-altruism* hypothesis (Batson, 2017; Batson, Eklund, Chermok, Hoyt, & Ortiz, 2007; Batson, Lishner, & Stocks, 2015).

Empathy is a complex, multilevel concept, and its definition varies depending on the author. Broadly, it can be defined as the ability to share (affective component) and understand (cognitive component) the affective state of another person, discriminating also (either implicitly or explicitly) the source of that emotion – i.e. internal (self) vs

external (other) (de Vignemont & Singer, 2006; de Waal, 2008; Decety & Cowell, 2014b; Shamay-Tsoory, 2011). Furthermore, empathy has been linked to emotional contagion abilities observed in childhood, which has also been found in other mammals, suggesting some evolutionary roots (de Waal, 2008; De Waal & Preston, 2017; Decety & Cowell, 2014b; Dondi, Simion, & Caltran, 1999; Langford et al., 2006).

Empirical studies have supported the empathy-altruism hypothesis. In one classic set of studies, when participants observed others receiving painful electric shocks, they took their place, receiving the remaining set of shocks (Batson, 2010, 2017; Batson et al., 2015). These studies have revealed that inducing higher empathic concern makes people more willing to help and have suggested that alternative explanations do not explain helping behaviour in this paradigm. In this set of experiments, *empathic concern* is a construct somewhat different from the definition of empathy provided above. Here, empathic concern refers to an other-oriented emotional response towards the welfare of others in need. These emotional responses are not necessarily shared with the other person, and are strongly related with concepts such as compassion and sympathy. Thus, people triggered by another person's distress are readily willing to help them.

Other studies have supported these results, expanded them to other aspects of the concept of empathy. For instance, previous research has revealed that people who strongly react, physiologically and behaviourally, to the pain of others, act more prosocially, suggesting that congruent shared affective states between self and other facilitates prosociality (FeldmanHall et al., 2015, 2013; Hein et al., 2010; Hu et al., 2021). Furthermore, emotional contagion, which has also been considered as a component of empathy, has been linked to prosocial behaviour in other mammals - observing another rat receiving painful shocks triggers a behavioural and physiological reaction in the observer rat that is linked with helping behaviour (Hernandez-Lallement et al., 2020). Taken together, empathy seems to be a strong facilitator of prosocial behaviour.

Empathy is often considered as a trait that can be present in different degrees, measured across individuals by self-report scales. Typically, such scales measure the degree to which people affectively and cognitive engage with others' emotional states (Lockwood, 2016). Variability in empathic traits has been associated with variability in prosocial tendencies (e.g. FeldmanHall et al., 2015; Hu et al., 2021; Lockwood, Seara-Cardoso, & Viding, 2014), although other studies have failed to find such link (e.g. Brethel-Haurwitz et al., 2018; Van Der Meulen, Van IJzendoorn, & Crone, 2016). Contradictory results have been also found in the relationship between empathic traits and empathic reactivity towards others' affective states (e.g. Brethel-Haurwitz et al., 2018; Contreras-Huerta, Baker, Reynolds, Batalha, & Cunnington, 2013; Hein et al., 2010; Singer et al., 2004). These results highlight the fact that, while empathic reactions can elicit prosocial behaviour, they are not synonymous, and empathy is not the only factor predicting prosociality. In fact, the distress elicited by others in need can indeed drive prosocial behaviour (*approach* response), but alternatively, can guide escaping behaviour from the source of distress (*avoidance* response; Bloom, 2017; Decety, 2011; Ibáñez et al., 2011; Yamada & Decety, 2009). Therefore, there is still a mechanistic gap between empathic traits, empathic reactions, and prosocial behaviour. This gap could be due to methodological considerations. Empathy-based prosocial studies typically do not separate others' benefits from self costs. Theoretically, individual differences in empathy traits should be linked to sensitivity to others' benefits and not necessarily to insensitivity to self-costs (Batson et al., 2002; de Waal, 2008). Observing another person in need might elicit helping behaviour, but the underlying mechanisms can be variable. Therefore, measuring individual differences in sensitivity to others' states is crucial to understand variability in empathic traits, empathic reactions, and prosocial behaviour.

1.2.3.2 Social rewards and positive affect

Rewards produce a subjective experience of pleasure, which motivates courses of actions that pursue those rewarding experiences (Paulus, 2007; Rangel, Camerer, & Montague, 2008; Rescorla, 1972; Schultz, Dayan, & Montague, 1997). A classic example of a primary reward is food. Animals are willing to expend a significant amount of energy to satiate their hunger and obtain food, and they can incur highly costly actions if that food has a great value. Crucially, several studies have revealed that social cues such as faces expressing positive feelings, positive (affective) touch, and social approval, can elicit subjective rewarding experiences (Brinkmann, Franzen, Rossier, & Gendolla, 2014; Cascio, Moore, & McGlone, 2019; Dichter, Richey, Rittenberg, Sabatino, & Bodfish, 2012; Fareri & Delgado, 2014; Sims, Van Reekum, Johnstone, & Chakrabarti, 2012). Experiments have shown that people are willing to incur significant costs to obtain such social outcomes (Brinkmann et al., 2014; Prévost, Pessiglione, Météreau, Cléry-Melin, & Dreher, 2010; Smith et al., 2010). Importantly, one of the courses of actions that usually brings social rewards is prosocial behaviour (Fareri & Delgado, 2014; Karns, Moore, & Mayr, 2017; Tamir & Hughes, 2018). Social rewards can be provided not only by the person being helped, but also by third-parties observing the prosocial actions. Consistently, experiments have shown that prosocial actions are more likely to occur in public than in private (Izuma, Saito, & Sadato, 2010; Moll et al., 2006), which might help to build a reputation that increases their social esteem (Emler, 1990; King-Casas et al., 2005; Phan, Sripada, Angstadt, & McCabe, 2010). Thus, direct and indirect social rewards elicited by prosocial behaviour can facilitate these actions to occur in the future.

Importantly, these positive affective experiences can also be elicited by observing another person getting a reward (Lockwood, 2016). Previous studies in humans and animals have shown that observing another individual receiving a reward elicit physiological, neural and affective responses in the observer congruent with the observed reward consumption (Chang, Gariepy, & Platt, 2013; Hackel, Zaki, & Van Bavel, 2017;

Lockwood, Apps, Roiser, & Viding, 2015; Lockwood, 2016; Morelli, Sacchet, & Zaki, 2015). This vicarious reward processing has been extensively associated with vicarious learning - agents can anticipate and evaluate others' rewards to then apply that learning into their own set of motor commands (Chang et al., 2013; Lockwood et al., 2015; van Gurp, Hoog, Kalenscher, & van Wingerden, 2020). Crucially, people who display stronger affective and physiological responses to others' rewards are more willing to act prosocially (Fukuda et al., 2019; Harbaugh, Mayr, & Burghart, 2007; Lockwood, Apps, Valton, Viding, & Roiser, 2016). Thus, it is possible that the mechanisms underpinning vicarious and social reward effects on prosocial behaviours are similar. Positive affective experiences given by positive social expressions or merely by passively observing others obtaining rewards may facilitate the agent to act in a way that results in those outcomes.

Intriguingly, the empathy-altruism hypothesis predicts similar mechanisms but with a negative valence – others' pain produces a negative affective state in the agent who acts to alleviate that distress in the other person. Even though similar *other-regarding* mechanisms, based on empathic reactions, might underlie prosocial behaviour across these positive and negative domains, there has not been clear evidence to link both and support this hypothesis. Furthermore, in both pain and reward contexts, a high variability in prosocial motivation has been found between people. Without disentangling costs and benefits in prosocial behaviour, addressing variability in motivation across these contexts is elusive. If individual differences in empathy underlie variability in reward and pain-related prosocial behaviour, higher sensitivity to others' outcomes, regardless of their valence, could explain greater prosocial dispositions.

1.2.3.3 Interoception, affect and prosocial behaviour

Interoception is the ability to process afferent signals from the body, especially those from visceral organs and the vascular system (Craig, 2002, 2003; Critchley & Garfinkel, 2017; Seth, 2013). These signals inform cognitive and affective processes about the state

of the body and facilitates body-ownership experiences and self-awareness. The classic James-Lange theory of emotions proposes that awareness of physiological body arousal facilitates affective experience (Cannon, 1927). Consistently, people who have higher interoceptive abilities show stronger physiological and behavioural reactions to emotional stimuli, and report more intense emotional experiences (Barrett, Quigley, Bliss-Moreau, & Aronson, 2004; Critchley, Wiens, Rotshtein, Öhman, & Dolan, 2004; Dunn et al., 2010; Pollatos, Traut-Mattausch, Schroeder, & Schandry, 2007; Wiens, 2005). Interoception not only informs about the current state of the body, but also predicts its future states, anticipating affective experiences that an outcome could trigger if an action is performed and facilitating goal-directed behaviour (Ondobaka, Kilner, & Friston, 2017; Paulus, 2007; Paulus & Stewart, 2014; Seth, 2013).

Interoceptive ability varies considerably across individuals, which is typically measured through performance in the heartbeat detection test (Critchley & Garfinkel, 2017). This task measures the ability to perceive own heartbeats at rest by either counting, tapping or by estimating heartbeats timing relative to external cues. Although this task has its limitations (Murphy, Brewer, Hobson, Catmur, & Bird, 2018; Ring & Brener, 2018), it has been useful to identify and characterise the high variability existing in interoception. High interoceptive ability has been linked to different social processes (Ferri, Ardizzi, Ambrosecchia, & Gallese, 2013; Löffler, Foell, & Bekrater-Bodmann, 2018; Ondobaka et al., 2017; Palmer & Tsakiris, 2018), including greater affective sharing and cognitive understanding of others' emotions (Grynberg & Pollatos, 2015; Lischke, Weippert, Mau-Moeller, Jacksteit, & Pahnke, 2020; Shah, Catmur, & Bird, 2017; Terasawa, Fukushima, & Umeda, 2013). Thus, interoception may link social contexts together with bodily states, facilitating the evaluation and anticipation of others' outcomes through emotions experienced by the agent.

Given the role of interoception in socio-affective experiences, interoception might be associated with prosocial behaviour. Low interoceptive ability has been strongly linked

with deficits in recognising, identifying and describing own and other's emotions, i.e. alexithymia (Brewer, Cook, & Bird, 2016; Murphy, Catmur, & Bird, 2018; Shah, Hall, Catmur, & Bird, 2016), which in turn has predicted reduced prosocial motivation and empathic reactions (FeldmanHall et al., 2013; Grynberg, Luminet, Corneille, Grèzes, & Berthoz, 2010; Lockwood, Bird, Bridge, & Viding, 2013; Moriguchi et al., 2007). However, only one study, to my knowledge, has tested this hypothesis directly (Piech et al., 2017). This study revealed that greater interoceptive ability predicted higher monetary donations in the dictator game, although training interoceptive capacity did not improve donating behaviour. This suggests that baseline but not trained interoceptive capacity might be linked to prosociality, in contrast with studies showing that meditation training improves levels of interoceptive ability, empathy and prosocial behaviour (Gibson, 2019; Kirk et al., 2016). Thus, it is unclear what the specific link between interoception and prosocial behaviour is.

The link between interoception and prosocial behaviour still needs to be sufficiently tested, as evidence is scarce. Furthermore, it is unknown what mechanisms might be involved in this putative link. On the one hand, interoceptive signals could inform the current agent's body state while deciding to help others, modulating the personal costs of performing the prosocial action. On the contrary, interoceptive signals could inform about the affective states triggered by the benefit to be received by the other person, modulating thereby sensitivities to the social benefit of the prosocial action. However, since the only previous study researching this link has used the dictator game, the personal costs and the social benefits have been confounded. Thus, the current state of the link between interoception and prosocial variability has yet to account for.

1.2.3.4 Psychiatric traits inhibiting prosociality.

Psychiatric disorders are maladaptive psychological and behavioural patterns that bring substantial personal distress in the person who suffers them, affecting the normal

functioning of the individual (American Psychiatric Association, 2013; Bolton, 2008). Despite a high variability in their symptomatology, something common across disorders is social and affect disruption. As prosocial behaviour is known for impacting social relationships, life satisfaction, mental and physical health, and economic success (Brown & Brown, 2015; Post, 2005; Stern, 2019), its alteration might be linked to psychiatric disorders. Indeed, empirical studies have shown that personality, depressive and anxiety disorders are associated with disruptive prosocial behaviour in experimental settings, as well as reduced socio-affective processing (e.g. American Psychiatric Association, 2013; Blair, 2007; Herpertz & Bertsch, 2014; James & Blair, 2013; Mier et al., 2013; Rilling et al., 2007; Robson, Repetto, Gountouna, & Nicodemus, 2019; Unoka, Seres, Áspán, Bódi, & Kéri, 2009).

Subclinical psychiatric traits have been also associated with prosocial disruption. As it was mentioned above, alexithymia, i.e. the inability to identify and describe own emotions and characterised by externally oriented thinking (Bagby, Parker, & Taylor, 1994), has been linked to impaired prosocial behaviour and social cognition (Demers & Koven, 2015; FeldmanHall et al., 2013; Grynberg & Pollatos, 2015; Lockwood et al., 2013; Shah et al., 2017; Valdespino, Antezana, Ghane, & Richey, 2017). Furthermore, the social dimension of apathy, i.e. reduction in goal-directed behaviour characterised by high effort sensitivity (Ang, Lockwood, Apps, Muhammed, & Husain, 2017; Bonnelle et al., 2015; Le Heron et al., 2018; Lockwood, Hamonet, et al., 2017) has been associated with a decrease in the motivation to exert effort to help others, but not to the motivation to benefit oneself (Lockwood, Hamonet, et al., 2017). Consistently, different dimensions of apathy are linked to affective and cognitive empathy (Lockwood, Ang, Husain, & Crockett, 2017), suggesting that people who have reduced motivation may show a disruption in understanding and sharing others' emotions, which could theoretically then lead to a reduction in prosociality under the empathy-altruism hypothesis.

Taken together, different clinical and subclinical psychiatric traits can disrupt prosocial motivation. However, these traits are not independent. Many psychiatric disorders are characterised by high levels of alexithymia and apathy, and low levels of empathy and interoception, while apathy, alexithymia, empathy and interoception are highly interrelated to each other (e.g. American Psychiatric Association, 2013; Ang et al., 2018; Bird & Viding, 2014; Chow, 2000; De Berardis et al., 2017; Hendryx, Haviland, & Shaw, 1991; Husain & Roiser, 2018; Marin, 1990; New et al., 2012; Starkstein & Leentjens, 2008; Taylor & Bagby, 2004; Thobois, Prange, Sgambato-Faure, Tremblay, & Broussolle, 2017). Thus, even though psychiatric traits might disrupt prosocial behaviour, it is still unclear whether this effect is because their symptomatology is linked to affective and subclinical traits, or rather due to their specific contribution to prosociality.

1.2.4 Summary and critique of theories of prosocial motivation

The approaches described above have given important insights on prosocial behaviour. However, they are mostly focused on explaining prosociality at the population level, i.e. which factors facilitate or disrupt prosocial motivation across participants. As such, they partially ignore the high variability in prosocial tendencies that exists between individuals. Importantly, all these theoretical explanations are based on specific methodological approaches, where prosocial behaviour can occur due to specific motives and underlying factors. Thus, if addressing the variability within a task is challenging, accounting for variability between tasks gets even more complex. A common framework may therefore be necessary to have a better understanding of individual differences within and between domains of prosocial behaviour.

This common framework must address some conceptual and methodological limitations that underlie some of the empirical work described above. One of its biggest problems is that self-related costs are not disentangled from other-related benefits. When

an agent decides to act (e.g. travel) for an outcome (e.g. food), they must consider both the value of that outcome and the cost of the action. If, for instance, two animals travel long distances to find food, one animal could do so because it is very incentivised by the food, while the other animal might not be very sensitive to the effort costs of travelling. The same could be true for prosocial behaviour. As such, distinguishing between sensitivity to costs and benefits is fundamental to address individual differences in dispositions to act prosocially across different social situations. For example, if one person helps another, it is unclear whether this is driven by a low sensitivity to their own interest/cost, or rather, due to a high sensitivity to others' benefit. This is not trivial, as action-based decisions, other-regarding mechanisms, and self-biases in value all need to be measured, but without separating sensitivities to costs and benefits, we will not completely understand prosocial behaviour or its variability.

Furthermore, previous theories are based on tasks that only measure willingness to incur self-costs for others' benefits without comparing it with baseline self motivation. For instance, someone may appear less prosocial but only because generally they value their own costs highly, being reluctant to incur personal costs to benefit another person, but likewise, to benefit themselves, e.g. someone who is apathetic may not act prosocially, but only because they generally lack motivation. However, self-benefits and prosocial motivation do not necessarily correlate. It might be the case that a person who is highly averse to self-costs, may be more reluctant to act when benefitting self than others. Thus, the comparison between self and prosocial motivation is crucial to understand the idiosyncratic tendencies of a person to be prosocial relative to other individuals.

Finally, the manipulation of costs and benefits in prosocial behaviour presents important limitations in previous studies. On the one hand, the personal cost is generally financial. This is problematic because, first, many of the prosocial behaviours in everyday life do not involve financial costs, such as a helping a colleague to complete a task or a friend to move into a new house. Secondly, the value of a reward can vary depending on

external factors such as the socioeconomic status of the agent (Tricomi, Rangel, Camerer, & Doherty, 2010). Moreover, often the studies outlined above do not manipulate costs and benefits parametrically. It is unclear whether people treat small or large costs and benefits equally, and how their sensitivities vary to these degrees in magnitude. For instance, it is clearly different to give away a few dollars rather than large sums of money, or alleviating another person's mild rather than significant pain. People may only show prosociality or stop being prosocial at a certain magnitude of cost or benefit. Therefore, parametric manipulation of costs and benefits are likely to alter prosocial motivation and its variability across people.

In recent years, decision-making paradigms from cognitive neuroscience have been employed, accounting for many of the limitations described above. Among others, recent studies have used *temporal discounting* models, where self rewards get increasingly devalued as time delay increases, to research prosocial behaviour (Critchfield & Kollins, 2001). In these *social discounting* models, people usually prefer rewards for themselves than for others, with rewards getting increasingly devalued as social distance augments, in a similar manner to time delays (Bickel et al., 2012; Story et al., 2020; Strombach et al., 2015). Other studies have approached prosocial behaviour using *reinforcement learning* models that explain the probabilistic contingencies that an agent considers to pair decisions with an outcome value over time, researching how people learn actions that reward or harm others compared with self (Lockwood et al., 2016; Lockwood, Klein-Flügge, Abdurahman, & Crockett, 2020; Lockwood & Klein-Flügge, 2020). Even though these models account for most of the methodological flaws described above, their main focus has still been to give a mechanistic explanation of group effects on prosocial behaviour.

Here, I propose a model for prosociality using paradigms that disentangle cost and benefits, aiming to provide a common framework for individual differences in prosocial motivation across prosocial domains. For this purpose, I will use effort-based decisions as

a model for prosocial behaviour, to then extend it to other domains such as moral harm aversion. Different from financial losses, effort is an aversive cost across humans and other species, which is personal and subjective but not directly influenced by socio-economic status (Chong et al., 2017; Hull, 1943; Kool & Botvinick, 2018). Surprisingly, effort has been somewhat ignored by prosocial studies, even though many prosocial actions require, in fact, the exertion of effort. In the next section, I will outline this cost-benefit framework for prosocial motivation.

1.3 Prosocial motivations as a cost-benefit evaluation

Goal-directed behaviour is defined as how willing an individual is to incur a cost, such as effort, to obtain a benefit, such as food rewards (Le Heron et al., 2018; Pessiglione, Vinckier, Bouret, Daunizeau, & Le Bouc, 2018; Salamone & Correa, 2018). From this perspective, motivation to act prosocially can be understood as the willingness to incur personal costs to benefit others, as a function of cost-benefit evaluations (Fig.1.1A). A highly prosocially motivated person is therefore someone who is willing to incur significant personal costs to obtain only small benefits for another person. For instance, when helping a person to move to a new house, the benefit is to them, not the agent, in terms of workload reductions. The agent is only going to be prosocial if they see reducing other person's workload – the social benefit – as an incentive. Consequently, as the social benefit increases, it gets more likely that a prosocial behaviour will be performed (Fig. 1.1B). However, others' benefits are discounted by the personal costs of the action to be performed. If the person is moving large furniture into a flat on the tenth floor of a building without an elevator, the personal costs (and also the other person's benefit in terms of workload reduction) are much larger. Reframing prosociality as goal-directed decision-making problems may be highly fruitful for providing testing between the different theories of prosociality outlined above (Contreras-Huerta, Pisauro, & Apps, 2020).

Thus, whether a prosocial behaviour is undertaken or not occurs through a valuation process, in which a person weighs up the self costs and the others' benefits. Such valuations guide decisions of whether to perform a prosocial act or not, such that the agent will choose the action with the highest value (Rangel et al., 2008). However, people vary in their sensitivity to self costs and other benefits, making the value of a prosocial action highly subjective. The same prosocial behaviour can be a product of idiosyncratic 'weightings' or 'sensitivities' to social benefits and personal costs. For instance, a person can donate money to a charity because either they are very sensitive to others' welfare, or rather they are very insensitive to their own profit. Thus, even though the prosocial action is the same, the mechanisms underlying it are different. Due to this subjective nature of the decision-making process, disentangling how people weight self-cost and other-benefit is crucial to understand prosocial behaviour.

How do we investigate people's sensitivity to costs and benefits? Goal-directed tasks typically require participants to make a choice between two options - one associated with higher benefits for higher costs, and another associated with less benefit but less costs (Contreras-Huerta et al., 2020). As such, if the subjective value of the higher benefit/cost option is lower than the alternative, then the latter is chosen. Often such tasks are analysed using computational models that formalise the valuation process and quantify the degree to which the benefit is discounted by the cost. Recent prosocial decision-making studies have started to use these kinds of experiments to unveil the mechanisms underlying prosocial behaviour (Crockett et al., 2014; Lockwood, Hamonet, et al., 2017). Crucially, these studies include *self* and *other* conditions - in half of the trials, the outcome of the decisions, i.e. the benefit, is delivered to participants themselves, while in the other half, to another unknown person. This methodological approach allows for disentangling parametrically the benefits from costs for self and other.

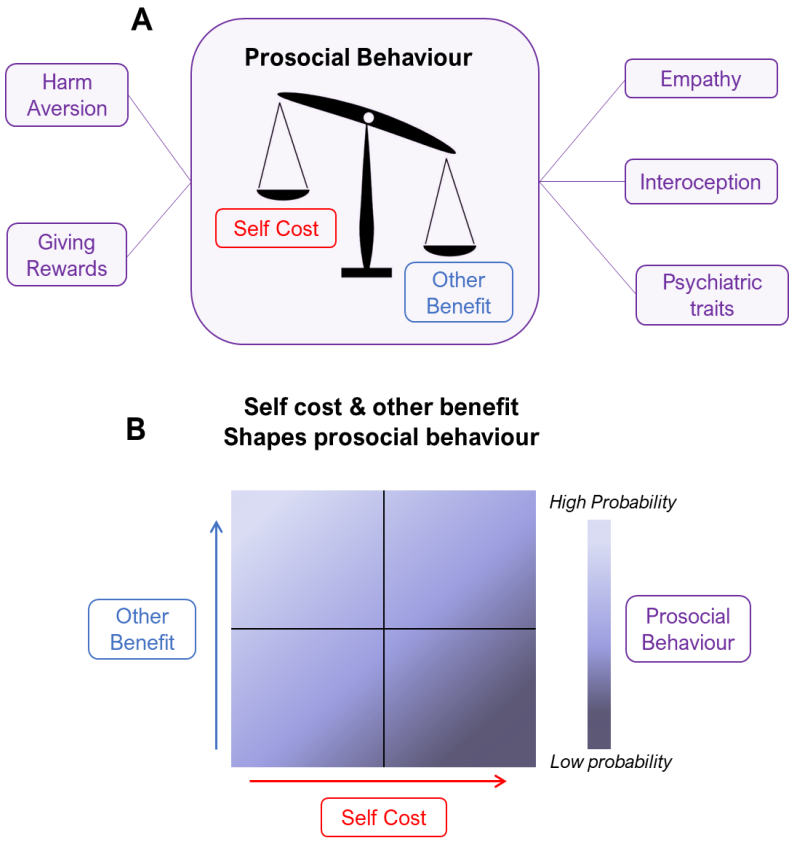


Figure 1.1. Schematic representation of prosocial motivation through cost-benefit evaluation. A. Prosocial motivation is the willingness to incur personal costs to benefit others, as a function of a cost-benefit evaluation. People who over-inflate the self-cost or under-weight the benefit to others will not value performing a prosocial act. This cost-benefit evaluation is influenced by the domain of the prosocial action, e.g. giving rewards or preventing others' harm (left), or characteristics of the agent, e.g. empathy capacity, interoceptive ability, psychiatric traits (right). B. Keeping other factors constant, people would choose to help when the other benefit is high, and the self cost low. Note that this illustration does not correspond to real data. *Adapted from Contreras-Huerta et al., 2020.*

Importantly, note that the definition of prosocial behaviour adopted in this thesis is restricted to a goal-directed perspective, which aligns with biological definitions of prosociality (de Waal, 2008; Kurzban et al., 2015) and provides a methodological program that improves some of the flaws of previous prosocial theories and paradigms mentioned above. However, this does not necessarily cover the whole spectrum of what has been called *prosocial behaviour* in the literature (see Batson et al., 2015; Padilla-Walker & Carlo, 2015; Schroeder & Graziano, 2015; for further discussion on the definitions of prosociality). For instance, some traditions have proposed prosociality as those actions that have the purpose to benefit one person or a group of people. This is a

broader perspective, which i) does not consider the costs associated with the potential action, and ii) does not involve *necessarily* a benefit for another person, as the crucial aspect here is the goal of the action. Crucially, the scope of this thesis is not giving a mechanistic account to all possible types of prosocial behaviours, nor giving a conceptualisation where all traditions studying this phenomenon can converge. Rather, this thesis claims that, certainly, and following biological and cognitive traditions, i) a wide range of prosocial behaviours require cost-benefit evaluations, ii) studying these evaluations can provide mechanistic explanations to the phenomenon, and iii) provides a methodological approach that has advantages compared with previous ones. Therefore, this is a behavioural and cognitive perspective of prosociality, which does not necessarily explain all the other levels of this complex concept.

In the next two sections I will outline two tasks that form the basis for much of the methodological approaches deployed in this thesis that follow this cost-benefit perspective to prosocial behaviour – examining prosocial effort, and harm aversion, based decisions.

1.3.1 Prosocial Effort

Many acts that occur during prosocial interactions are effortful. Typically, people are averse to exerting effort (Apps, Grima, Manohar, & Husain, 2015; Kool & Botvinick, 2018; Kurzban et al., 2013). If two acts are associated with the same reward, most species will choose the least demanding (Le Heron et al., 2018; Salamone & Correa, 2018). Although there are exceptions (Inzlicht, Shenhav, & Olivola, 2018), this aversion to effort has been robustly demonstrated for a range of physically and cognitively demanding tasks, and is a key factor in how people decide what activities to pursue (Apps et al., 2015; Chong et al., 2017; Kool & Botvinick, 2018). Thus, many prosocial behaviours with effort costs

associated are likely to be avoided if they require more effort than people are willing to undertake.

Recently, Lockwood et al. (2017) demonstrated that people will perform prosocial effortful acts to benefit others, but highlighted how influential effort costs are (Fig.1.2A). They used an effort-based decision-making paradigm that paralleled non-social effort-based tasks (Chong et al., 2017). Participants chose between 'working' for different levels of reward (2–10 credits), which required physical effort (30–70% of their own maximum grip strength on a handheld dynamometer), or 'resting', which required no effort but was associated with low reward (1 credit). On half of the trials participants would receive the reward themselves, but on the other half the rewards would be delivered to an anonymous other. This was a real economic decision, with participants required to put in a physical effort every time they chose to, in order to obtain a financial bonus for participating. Using this design, it is possible to dissociate how effort and reward sensitive people are by examining how willing they were to choose to work, as a function of the grip force and credits presented in the offers. Moreover, using this design it was also possible to control for differences in capacity — maximum grip strength — ensuring that variability in choice behaviour between people depended on differences in the willingness to put in effort to obtain rewards. Thus, reward and effort sensitivity could be examined for both self-benefitting and prosocial acts.

Lockwood, Hamonet, et al. (2017) found that at lower effort levels, people were almost equally willing to work to obtain rewards for themselves and the stranger, but at higher levels people became more prosocially apathetic, choosing to 'rest' more for other than self. Furthermore, they found that people overcame effort costs when the reward was high relative to low, but they did it in a lesser degree when the reward was for another person. This task therefore showed that both the self-cost and the other benefit matter for prosocial behaviour. Moreover, by comparing them with the self condition they could show a three-way interaction between beneficiary, effort level and reward level,

such that people weigh both the effort costs and rewarding benefits differently for self or other. Thus, neither costs nor benefits are the same when considering a prosocial versus self-benefitting act.

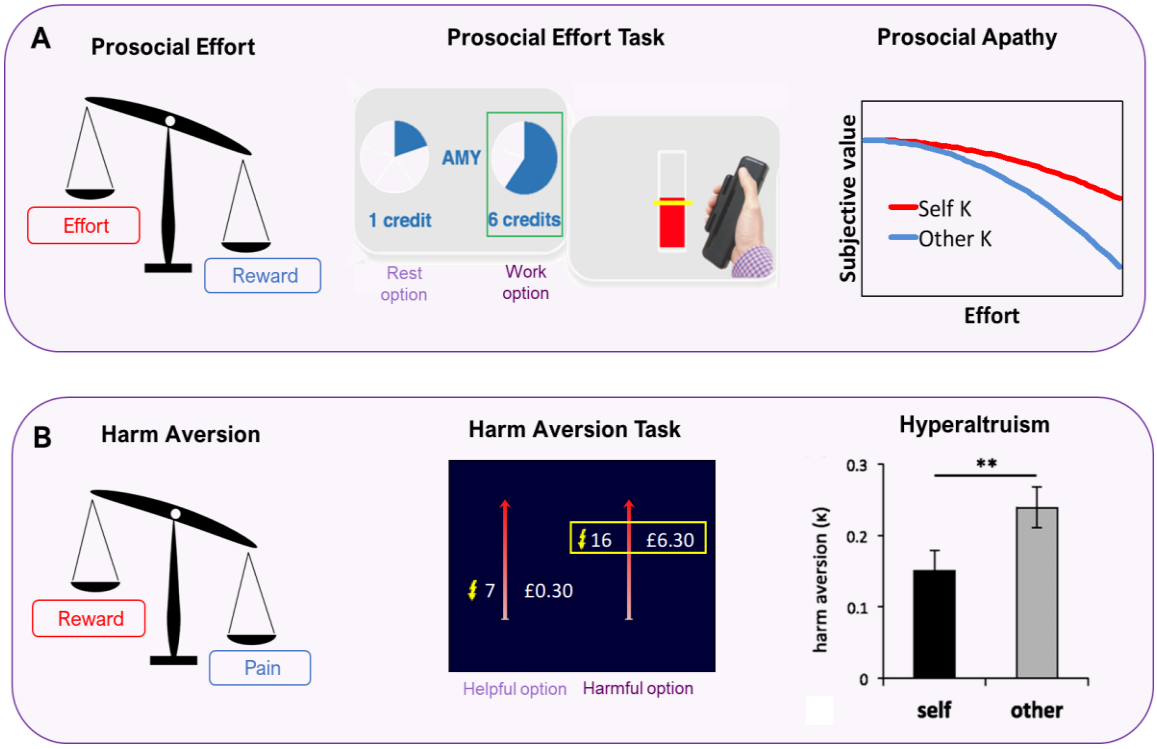


Figure 1.2. Two tasks measuring prosocial behaviour. A. *Left*. In prosocial effort, effort costs are weighing up against reward benefit for another person. *Middle*. In the prosocial effort task (Lockwood, Hamonet, et al., 2017), participants decide between a rest option associated with no effort (depicted by only one element in the pie chart) and low reward (number of credits), and a work option associated with higher levels of effort (depicted by more elements of the pie) and rewards. In half of the trials, reward is received by another person, while in the other half, it is received by participants themselves. After deciding, participants must exert the physical effort required. *Right*. In this task, participants show prosocial apathy, discounting the subjective value of rewards to a higher degree as effort increases (x-axis) for other (blue line) than self (red line). K corresponds to the discounting parameter calculated from an effort-discounting computational model. B. *Left*. In harm aversion, financial losses (cost) are weighing up against reducing others' harm (benefit) *Middle*. In the harm aversion task (Crockett et al., 2014, 2017), participants decide between a helpful option, associated with low financial gains for low harm (number of electric shocks), and a harmful option, associated with higher levels of profit and harm. In half of the trials, shocks are received by another person, while in the other half, they are received by participants themselves. *Right*. In this task, participants show hyperaltruism, with people being more averse to harm others for profit than themselves. κ is the harm aversion parameter, representing the degree in which profit is discounted by harm from a harm-discounting computational model. *Adapted from Lockwood, Hamonet, et al., 2017, and Crockett et al., 2014.*

Computational modelling revealed that this arises because people devalue rewards parabolically by physical effort (Chong et al., 2017; Hartmann, Hager, Tobler, & Kaiser,

2013; Lockwood, et al., 2017) - small differences in effort sensitivity to obtain a reward for someone else compared to oneself are amplified when the effort increases. Such aversion to prosocial efforts also extends to charitable donations, people donating more readily when it is easy to do so (Imas, 2014). Importantly, people display high variability in this task, which has been correlated with the age of the agent and levels of psychiatric traits such as psychopathy and apathy, suggesting that variability in prosocial effort depends on the characteristics of the agent (Lockwood, Hamonet, et al., 2017, Lockwood et al., 2021; Mosner et al., 2017).

Importantly, this task provides another measure of prosocial motivation. As people are required to perform the action after they have chosen to work, it is possible to measure how much energy is actually put in, and compare it with the effort level required. The discrepancy between the effort chosen and the effort actually exerted can reveal differences in effort-avoidance between a decision and its implementation, i.e. a degree of superficiality when energising the chosen behaviour. This could be relevant when explaining other types of discrepancy, such as moral hypocrisy, that might also be the result of skipping the costs of being prosocial. Taken together, using designs inspired by effort-based decision-making research revealed new insights into how people become prosocially apathetic when helping others out is highly effortful.

1.3.2 Harm Aversion

Moral theories have identified preventing to harm others as a powerful moral principle that shapes prosocial behaviour, with people typically finding harming others distasteful and distressing, being readily to avoid these actions (Cushman, 2013). From this perspective, it can be hypothesised that people might be more averse to harm others than themselves, given the moral nature of the former relative to the latter. This hypothesis was formally tested by Crockett et al. (2014), where participants traded-off

financial profit against a number of painful stimuli, i.e. electric shocks (Figure 1.2B). In this study, participants decided between two options – a harmful option, associated with high levels of profit but high number of shocks; and a helpful option, with low profit for low number of shocks. Crucially, in half of the trials, the electric shocks were delivered to participants themselves, while in the other half the shocks were received by an anonymous stranger. Using computational modelling, decisions were better described by a model that contained separate parameters for self and other conditions. The harm aversion parameter quantified the degree in which reward was discounted by harming self and other separately. Remarkably, people displayed higher harm aversion for others than for self in exchange of profit, i.e. 'hyperaltruism', effect that has been replicated in a series of experiments since this pioneering work (Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Crockett et al., 2015; Volz et al., 2017; Zhan et al., 2020).

However, although this effect has repeatedly been replicated at the group level, there is a high variability between people in their prosocial preferences in harm aversion, which might relate to the idiosyncratic cost-benefit evaluation of each participant. However, Crockett et al. (2014) did not disentangle whether the self cost or the other benefit in the decision was driving the hyperaltruism effect, as the harm aversion computational parameter quantifies the relationship between profit and harm. Following the action-based hypothesis in harm aversion, the profit earned through harming others might be devalued compared with when it is received by harming oneself. On the other hand, if hyperaltruism is an empathy-based prosocial behaviour, hyperaltruistic individuals would be more sensitive to the pain of others than themselves, which elicit more aversion to others' pain than their own. Finally, an alternative hypothesis is that a combination of both mechanisms is playing a role in hyperaltruism. In an fMRI study, Crockett and colleagues (2017) tested these hypotheses, examining whether individual differences in hyperaltruism correlated with individual differences in the neural signal for harm, profit, or both in the harm aversion task. They found distinctive neural circuits representing profit

and harm, which have been traditionally associated with self and vicarious reward and pain processing, respectively. However, only neural representation of profit predicted hyperaltruism, such that more hyperaltruistic people showed reduced responses to money earned by harming others compared with harming self. Intriguingly, using connectivity analysis, the authors found that this neural devaluation of immoral profit was functionally linked with neural representations of moral principles in the prefrontal cortex. This latter result suggests that moral norms encoded in the prefrontal cortex regulate neural responses to profit when earned immorally. These findings support the hypothesis that action-based mechanisms underlie hyperaltruism, in which ill-gotten profit is devalued.

1.3.3 Distinctive and common aspects of prosocial behaviour in different domains.

The tasks described above revealed notable differences in the pattern of behaviour depending on the context - participants are more willing to sacrifice profit to prevent harming others relative to themselves, i.e. hyperaltruism; but they are more reluctant to exert effort to reward others relative to themselves, i.e. prosocial apathy. Importantly, these studies share some common methodological implementations and theoretical assumptions. First, both tasks disentangle self from prosocially motivated actions. Secondly, both models assume that people based their decisions on the subjective value of the choices. Finally, both models explain behaviour in function of how much people discount rewards by aversive factors, represented by a parameter fitted per each participant and separately for *self* and *other*.

Theories on prosocial behaviour have suggested that harm aversion and prosocial effort might be uncorrelated (Carnes & Janoff-Bulman, 2012; Janoff-Bulman, Sheikh, & Hepp, 2009). In harm aversion, people need to inhibit a temptation through moral norms,

devaluing the rewarding activity. The same moral norm does not proscribe harming oneself for profit, and therefore hyperaltruism occurs. Prosocial effort does not have a moral norm as universal and consistent as harm aversion guiding behaviour. While profiting by harming others is intolerable in all contexts, not working for other people can be seen as an indifferent or immoral act depending on the context. Thus, in prosocial effort, people must activate their motivation to act through others' rewards, which they typically value to a lesser degree than rewards for self (Camerer, 2011). Volz et al (2017) revealed findings supporting this claim. They replicated the hyperaltruism effect in a modified version of the harm aversion task, but added an inversed condition – participants decided whether to get rewards for another person by getting painful electric shocks themselves. Results revealed that participants were less willing to accept harm to reward others compared with themselves, consistent with the pattern of behaviour shown in the prosocial effort task. Importantly, aversion to harming others for profit was not correlated with harming oneself for others' rewards. Thus, while harm aversion could be seen as a moral act, incurring costs to reward others might be seen as supererogatory, and therefore, it might have different mechanisms.

From a cost-benefit evaluation model, though, an important difference between prosocial effort and harm aversion models might underlie their observed patterns of behaviour. In the prosocial effort model, reward is the benefit that is discounted by effort cost, while in the harm aversion model, (losing) reward is the self cost, which is discounted by (preventing) harm, the other benefit. This difference in these models points out that costs and benefits between these contexts have opposite valences. In prosocial effort, the aversive action of exerting effort brings the positive outcome of reward, while in harm aversion, inhibition of a rewarding action avoids the negative outcome of harm. In this sense, the differences between these contexts emulate classic approach-avoidance motivations, i.e. behavioural activation and inhibition (Carver & Scheier, 2001; Gable, Reis, & Elliot, 2003; Gray, 1990). While people approach to rewards and try to follow

courses of actions that end in rewarding stimuli, they avoid pain and inhibit behaviours that bring harm. These basic motivations are the foundations of many goal-directed behaviours, and it can be a powerful framework applied to these two prosocial contexts. Crucially, both approach and avoidance motivations, even though independent in many aspects, share some common mechanisms under the cost-benefit framework. This therefore suggests that harm aversion and prosocial effort might also have common aspects explained by shared mechanisms.

It is unclear though what common mechanisms could underlie these two domains. Prosocial individuals across domains may be highly sensitive to the benefit received by another person irrespective of its valence, and regardless of whether, at the group level, people display different patterns of behaviours depending on the context. On the other hand, people who are not prosocial across domains might be too sensitive to the costs. It is also possible that people are sensitive only to the valence of outcomes and actions across beneficiaries and contexts. These hypotheses could be tested using tasks like the above-mentioned ones that disentangle cost vs benefits and self vs other motivations. Therefore, this cost-benefit model provides a potential common approach to prosocial behaviours. Measuring individual differences in sensitivities to self costs and other benefits, irrespective of their valence, can illuminate shared and distinctive mechanisms across different prosocial domains.

1.3.4 Dynamic costs in goal-directed behaviour: Opportunity costs.

Besides harm aversion and prosocial effort, towards the end of this thesis I will propose a novel framework within a prosocial cost-benefit model that considers a dynamic cost for the agent - opportunity cost. Opportunity cost corresponds to the missed benefit of an alternative, better decision than the chosen one. Cost-benefit decisions are therefore relative to the other alternative decisions available (Constantino & Daw, 2015).

The ideas behind opportunity costs stem from behavioural ecology and how animals made foraging decisions in the wild - organisms maximise rewards in the environment by making continuous opportunity cost-based decisions (Charnov, 1976). This differs from the tasks above, in that agents make decisions to obtain rewards in relation to the environment.

In an ecological niche, organisms must make decisions dynamically considering how the environment changes over time (Charnov, 1976; Mobbs, Trimmer, Blumstein, & Dayan, 2018). For example, many species must collect food from the environment, incurring significant costs to obtain it, e.g. travelling long distances. Once in a location with a certain amount of food available, the animal must decide when to leave that location, as the longer it spends collecting food there, the resources deplete over time and it might be missing the opportunity to collect rewards from another, richer location somewhere else in the environment. Thus, an animal facing this kind of decisional problem must contrast moment-to-moment the rewards being obtained in the current location with the opportunity cost of the environment.

Crucially, social animals such as bees, wolves and primates, including humans, not only forage to collect rewards for themselves but also to share it with the community (Amano, Ushiyama, Moriguchi, Fujita, & Higuchi, 2006; Hassell & Southwood, 1978; Kelly, 2013; Lemanski, Cook, Smith, & Pinter-Wollman, 2019; Metz, Vucetich, Smith, Stahler, & Peterson, 2011; Rita, Ranta, & Peuhkuri, 1997; Smith et al., 1983). Hunter-gatherer human societies, past and present, forage for resources which are consumed by all conspecific members. Therefore, cost-benefit evaluations in prosocial behaviour might have been shaped by ecological decision-making problems that we as species have faced from our evolutionary roots (Gabay & Apps, 2020). However, these models of dynamic ecological decisions have largely been ignored by theoretical and methodological approaches in prosocial behaviour research. Current social decision-making paradigms involve only instantaneous self and social outcomes, without

considering environmental contingencies and how they change the value of these outcomes. It is necessary, then, to reconsider cost-benefit prosocial evaluations from an ecological perspective, emulating decisional problems that our forager ancestors might have dealt with.

1.4 Neural underpinnings of individual differences in prosociality

In the previous section, I have proposed a model for prosocial behaviour based on cost-benefit evaluation, in which idiosyncratic sensitivities to self costs and other benefits can influence individual differences in prosocial motivation in multiple domains. What areas in the brain could be involved in signalling individual differences under this framework? From all the areas classically associated with prosocial behaviour, there are two regions, commonly co-activated across a number of cognitive, affective and social tasks that look like good candidates – the anterior cingulate cortex (ACC), and the anterior insula (AI). Crucially, AI and ACC have been involved in weighing up the costs and benefits of acting, ascribing a subjective value to actions (Droutman, Bechara, & Read, 2015; Naqvi & Bechara, 2009; Paulus, 2007; Rushworth & Behrens, 2008; Smith, Steinberg, & Chein, 2014; Wallis & Kennerley, 2011; Walton, Croxson, Behrens, Kennerley, & Rushworth, 2007; Walton, Rudebeck, et al., 2007). Critically, ACC and AI have also been involved in social processes that are fundamental for the present framework. Individual differences in their neural responses predict monetary donations (economic games), avoidance to harm others (moral principles), guilt feelings when acting immorally (moral hypocrisy), empathy for pain (empathy-altruism), reactivity to others' rewards and positive emotions (social rewards), and ability to perceive internal bodily changes (interoception; Boccia et al., 2017; Critchley et al., 2004; Fan, Duncan, de Greck, & Northoff, 2011; Greene et al., 2004; Hu et al., 2021; Lamm, Decety, & Singer, 2011; Li, Yu, Zhou, Kalenscher, & Zhou, 2020; Lockwood, 2016; Paulhus, Neumann, &

Hare, 2009; Rilling & Sanfey, 2011; Yu, Hu, Hu, & Zhou, 2014). ACC and AI responses might therefore be crucial for understanding the common and distinct mechanisms underlying individual differences in prosocial cost-benefit evaluations across domains.

In the following section, I will review anatomical aspects of the AI and ACC, with a special focus on the gyrus portion of the latter area, as it has been argued to be the subregion of the ACC specialised in processing socially relevant information (Apps, Rushworth, & Chang, 2016). Next, I will present evidence showing the involvement of ACC and AI in signalling negative and positive outcomes for others, and how this signal predicts prosocial behaviour. Finally, I will propose some functional distinction between ACC and AI in prosocial behaviour through interoceptive ability.

1.4.1 Anatomical aspects of ACC and AI

Broadly speaking, the cingulate cortex can be divided into three different subregions based on its cytoarchitecture – posterior, middle (MCC) and ACC (Palomero-Gallagher, Mohlberg, Zilles, & Vogt, 2008; Vogt, Nimchinsky, Vogt, & Hof, 1995). In the latter two, a sulcus (MCCs/ACCs) and a gyrus (MCCg/ACCg) portion can be also distinguished (Fig. 1.3). Within the MCC and ACC, there is an additional subdivision between areas 24, 32 and 25, which can be identified in humans, macaques and rodents (Neubert, Mars, Sallet, & Rushworth, 2015; Rushworth, Walton, Kennerley, & Bannerman, 2004; Vogt & Paxinos, 2014). The ACCs corresponds approximately to the anatomical area 24c, and is also often referred to as dorsal ACC. On the other hand, ACCg has been defined approximately in areas 24a/b.

Across the sulcal and gyral MCC and ACC there are common connections with multiple other brain regions, including within the prefrontal cortex (areas 10-14), dorsolateral prefrontal cortex (DLPFC), supplementary motor area (SMA), premotor and motor cortices, and dorsal AI (Lockwood, 2016; Picard & Strick, 1996; Showers, 1959;

Wang, Shima, Sawamura, & Tanji, 2001). However, the MCCg/ACCg, has connections with three functional distinctive brain networks that are somewhat different from the sulcal regions (Apps et al., 2016; Beckmann, Johansen-Berg, & Rushworth, 2009; Lockwood, 2016; Vogt, 2009) – (i) portions of the temporo-parietal junction (TPJ) and dorsomedial prefrontal cortex (DMPFC), both involved in mentalising processes; (ii) portions of amygdala, ventral AI, and ventromedial prefrontal cortex (VMPFC), all regions associated with socio-affective and value-based processing; and (iii) ventral premotor cortex and inferior-parietal lobule, both areas involved in the action-perception system. This connectivity profile suggests that ACCg has a functional strategic location to integrate different sources of social information, being proposed as having a key role in tracking socially relevant information in a cost and benefit evaluation framework (Apps et al., 2016).

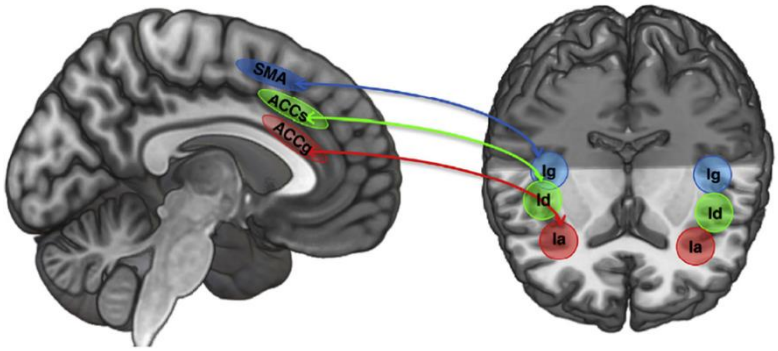


Figure 1.3. Location and connectivity patterns of anterior insula and anterior cingulate cortex. SMA = Supplementary Motor Area, ACCs/ACCg = Sulcus and gyrus portions of the anterior cingulate cortex respectively, lg = granular posterior insula, ld/la = dysgranular dorsal, and agranular ventral anterior insula respectively. *Adapted from Lockwood, 2016.*

The insular cortex is commonly divided into three cytoarchitectonic functional subregions based on the degree of granularity that the different insular portions show – the granular posterior cortex, the dysgranular dorsal anterior cortex, and the agranular ventral anterior cortex (Fig. 1.3; Chang, Yarkoni, Khaw, & Sanfey, 2013; Deen, Pitskel, & Pelphrey, 2011; Nomi, Schettini, Broce, Dick, & Uddin, 2018; Uddin, Kinnison, Pessoa, & Anderson, 2014). The posterior insula cortex (PI) has connections principally with primary

and secondary motor and somatosensory cortices, but also with MCC, SMA, VMPFC and temporal poles, being associated with sensorimotor processing (Deen et al., 2011; Lockwood, 2016; Nomi et al., 2018; Uddin et al., 2014; Uddin, Nomi, Hébert-Seropian, Ghaziri, & Boucher, 2017). It is also linked to interoception, as it receives afferent projections from the lamina I spinothalamocortical with internal sensory information from the body (Craig, 2009; Craig, 2002).

The dorsal portion of AI (dAI) has connections principally with ACCs, as well as medial prefrontal cortex (MPFC), DLPFC, superior parietal cortex, frontal inferior operculum, thalamus and ventral striatum, and it has been related to high cognitive function, attention, cognitive control and error processing (Deen et al., 2011; Lockwood, 2016; Uddin et al., 2014, 2017). Indeed, stronger connectivity between dAI and ACCs predicts better performance in cognitive tasks (Touroutoglou, Hollenbeck, Dickerson, & Feldman Barrett, 2012). Furthermore, dAI is connected and shares many pathways with PI, and it has also been proposed as a node for integrating different sources of information, including self-awareness and body ownership (Craig, 2002, 2009; Uddin et al., 2017). The ventral portion of AI (vAI), on the other hand, is connected with ACCg and multiple limbic regions, including amygdala and hippocampus, as well as orbitofrontal cortex, anterior temporal lobe and MPFC (Deen et al., 2011; Lockwood, 2016; Nomi et al., 2018; Uddin et al., 2014, 2017). Due to its connection, and its functional and structural closeness with limbic regions, vAI has been identified principally as processing socio-affective information. High connectivity between vAI and ACCg has been associated with stronger emotional experience (Touroutoglou et al., 2012). Thus, AI is believed to be an integrative paralimbic hub between cortical and limbic cortical zones.

1.4.2 ACC and AI signal others' outcomes across domains

ACC and AI have been involved in cost-benefit decision-making, including decisions on whether to exert effort for rewards, or avoiding a harmful outcome (Fauth-Bühler et

al., 2014; Klein-Flügge, Kennerley, Friston, & Bestmann, 2016; Koyama, Kato, & Mikami, 2000; Porreca & Navratilova, 2017; Xiao & Zhang, 2018; Yang et al., 2009). These regions are proposed to signal outcomes regardless of their valences, integrating affective and motivational, i.e. approach/avoidance, aspects of the outcome.

Previous studies have shown that ACC and AI also signal the anticipation and consumption of different kinds of rewards by other people (Basile, Schafroth, Karaskiewicz, Chang, & Murray, 2020; Mobbs et al., 2009; Morelli et al., 2015) and learning about contingencies that result in reward for others, especially the ACCg subregion (Burgos-Robles, Gothard, Monfils, Morozov, & Vicentic, 2019; Lockwood et al., 2015). Furthermore, these regions signal unexpected rewards received by others when their predictions are incorrect, and when another person makes a mistake in trial-error learning (Apps, Green, & Ramnani, 2013; Hill, Boorman, & Fried, 2016; Monfardini et al., 2013). Apps and Ramnani (2014) revealed that ACCg signalled specifically the net-value of others' rewards as a function of the effort costs that the other person would exert to obtain them, while ACCs signal varied across self and other net-values. Consistently with this social-specificity, ACCg also tracks the value of social advice and approval (Behrens, Hunt, Woolrich, & Rushworth, 2008; Jones et al., 2011). Likewise, in non-human primates, a significant proportion of neurons in the ACCg fires only in response to vicarious rewards (Chang et al., 2013), and the size of the grey matter volume in this area changes depending on monkeys' social network size (Sallet et al., 2011). Moreover, lesions in the ACCg, but not the ACCs, disrupt the specific processing of social relevant information but not other types of rewards such as food (Hadland, Rushworth, Gaffan, & Passingham, 2003; Rudebeck, Buckley, Walton, & Rushworth, 2006). Thus, the ACCg might be a key region signalling vicarious and social rewards.

AI and ACC have also seen consistently signalling others' harmful outcomes across very different tasks, including direct and indirect observation of harm, where harm is applied either by external devices (e.g. electric shocks, needles) or manifested through

facial expressions, and even if the harm is hypothetical (Contreras-Huerta et al., 2013; Danziger, Faillenot, & Peyron, 2009; Fan et al., 2011; Han et al., 2009; Jackson, Rainville, & Decety, 2006; Lamm et al., 2011; Lamm, Batson, & Decety, 2007; Lamm, Meltzoff, & Decety, 2010; Singer et al., 2004). Animal studies have supported the role of this network in vicarious harm, associated with learning and affective responses (Carrillo et al., 2019; Han et al., 2019; Jeon et al., 2010). Thus, ACC and AI signal vicarious rewards and harm, representing others' people outcomes regardless of their valence. These areas also respond to different outcome modalities, such as affective expressions (Fan et al., 2011; Lindquist, Satpute, Wager, Weber, & Barrett, 2016), and complex social outcomes (Corradi-Dell'Acqua, Civai, Rumiati, & Fink, 2013; Eisenberger, Lieberman, & Williams, 2003; Strobel et al., 2011), suggesting that ACC and AI are involved in representing the affective states of other people elicited by general positive and negative outcomes.

1.4.3 ACC and AI weight others' outcomes in prosocial behaviour

ACC and AI responses to others' outcomes can vary between people. For instance, ACCg vicarious reward signal covaries with individual differences in emotional contagion traits, such that people who have higher emotional contagion show more social specificity in ACCg responses to rewards (Lockwood et al., 2015, Fig. 1.4A). Furthermore, variability in ACCg responses to social cues predicts the influence of these cues on people's decisions (Behrens et al., 2008). Likewise, ACC and AI pain signal has been positively associated with individual differences in empathic traits, and negatively with clinical and subclinical psychiatric traits (Bird et al., 2010; Lockwood et al., 2013; Singer et al., 2004). Could these individual differences in ACC and AI responses to others' outcomes predict prosocial behaviour?

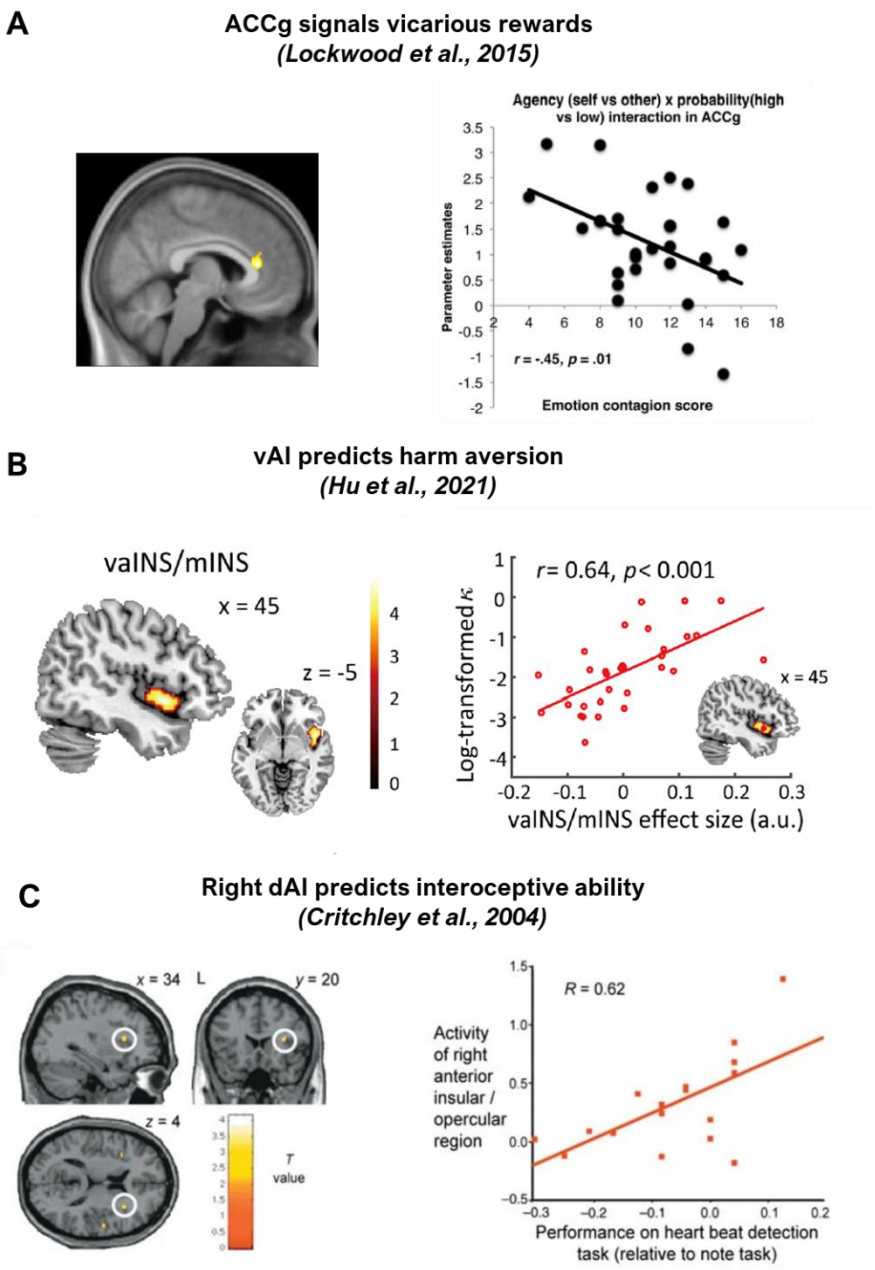


Figure 1.4. Individual differences in ACCg and AI signals predict empathy, prosocial behaviour and interoception. A. Individual differences in ACCg responses to other relative to self-reward prediction is associated with the emotional contagion sub-component of empathic traits. B. Individual differences in vAI responses to others' harm predict aversion to harm others for profit (represented by the parameter *K* in the plot). C. Individual differences in right dAI activity while performing an interoception accuracy task predicts interoceptive ability. ACCg = gyrus portion of anterior cingulate cortex, vAI/dAI = ventral and dorsal portions of the anterior insula respectively. The source study for each figure is stated in *Italic* in their title.

ACC and AI have been extensively associated with helping behaviour and charity donation (Bortolini et al., 2017; Cutler & Campbell-Meiklejohn, 2019; Gospic et al., 2013; Hare, Camerer, Knoepfle, & Rangel, 2010; Moll et al., 2006; Rameson, Morelli, &

Lieberman, 2012; Wang, Zhang, & Jia, 2019). Crucially, changes in ACC and AI activity to others' rewards can predict prosocial preferences. For instance, ACC is involved in processing prosocial preferences in reward-based learning in non-human primates i.e. greater tendency to give reward to another monkey relative to no one (Basile et al., 2020), and neurons in ACCg and the amygdala exhibit frequency-specific coordination that is enhanced for prosocial decisions but suppressed for antisocial decisions (Dal Monte, Chu, Fagan, & Chang, 2020). In humans, studies have shown that participants showing stronger AI responses to money earned by a charity are more prosocial and report more satisfaction when choosing to help (Fukuda et al., 2019; Harbaugh et al., 2007).

Previous studies have also shown that sensitivity to vicarious negative outcomes in these regions is linked to prosocial behaviour (Greening et al., 2014; Tusche, Bockler, Kanske, Trautwein, & Singer, 2016). One study revealed that stronger AI responses to ingroups getting electric shocks predicted helping behaviour towards them, effect that was associated with trait empathy (Hein et al., 2010). Intriguingly, extraordinary prosocial people (i.e. kidney donors) exhibit higher overlap between self and other pain representations in AI compared with controls (Brethel-Haurwitz et al., 2018). ACC and AI activity also predict prosocial behaviour in response to more abstract negative outcomes for others, such as immoral actions, injustice, unfairness, and social exclusion (Corradi-Dell'Acqua et al., 2013; David, Hu, Krüger, & Weber, 2017; Masten, Morelli, & Eisenberger, 2011; Strobel et al., 2011; Van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2009; White et al., 2017; Xiang, Lohrenz, & Montague, 2013), suggesting that these areas might signal social outcomes in a wide spectrum of modalities.

Crucially, AI can predict aversion to harm others (Hu et al., 2021, Fig. 1.4B) - people who were more averse to others' harm (aversive noise) in exchange of profit showed greater responses in vAI to others' pain. This effect also predicted higher scores in empathic traits, suggesting that neural response to others' outcomes might depend on

levels of empathy. Harm aversion related pain signals have also been found in rodents - rats avoid actions that result in harming others even at their own cost, effect that is mediated by ACC (Carrillo et al., 2019; Han et al., 2019; Hernandez-Lallement et al., 2020). Thus, responses to others' harm in these areas can predict harm aversion.

Taken together, individual differences in ACC and AI responses to others' outcomes, irrespective of their valences, can facilitate prosocial behaviour across domains. These areas seem to respond in an *other-regarding* manner, i.e. facilitating behaviours that increase the others' benefit. Since activity to others' outcomes in these areas is modulated by empathic traits, AI and ACC sensitivity to others' benefits might be relatively stable in an individual across domains.

1.4.4 Functional distinction between ACC and AI

Interoceptive ability is key for cost-benefit evaluations, as it signals current and future states of the body, which may be useful guiding actions depending on how the stimuli in the environment impact these states (Paulus, 2007; Paulus & Stewart, 2014; Seth, 2013). Given the link between interoception and socio-affective processes (Palmer & Tsakiris, 2018), this role of interoception may extend to prosocial behaviour.

Interoception might be crucial to disentangling the functional roles of ACC and AI in prosocial behaviour. Even though ACC and AI have been reported as coactivated in interoceptive tasks (Craig, 2009; Critchley & Garfinkel, 2017; Critchley et al., 2004), higher responses in dAI have been consistently found as predicting better interoceptive performance, an effect that has been linked to greater emotional experience and recognition (Fig. 1.4C, Craig, 2002, 2009; Critchley & Garfinkel, 2017; Critchley et al., 2004; Ernst, Northoff, Böker, Seifritz, & Grimm, 2013; Naqvi & Bechara, 2009; Terasawa et al., 2013, 2021; Zaki, Davis, & Ochsner, 2012). Importantly, vicarious responses to others' outcomes and emotions in AI can be modulated by interoceptive ability (Ernst et

al., 2013), whereas high levels of alexithymia, strongly linked to deficits in interoception (Murphy, Catmur, et al., 2018), have been associated with lower AI responses to vicarious pain and affective expressions (Bird et al., 2010; Silani et al., 2008). Alexithymia traits also predict less prosocial behaviour, which was related to lower AI activity to the distress of others (FeldmanHall et al., 2013). All this evidence suggests that, although ACC and AI roles are connected, the latter region might be processing vicarious outcomes anchored to the body.

The above is consistent with functional anatomical differences between and within the ACC and AI. While the ACC is associated with motivational-action aspects of decision-making, being involved in cognitive monitoring and emotional-arousal regulation, AI has been proposed as a hub integrating affective processing, salient stimuli evaluation, and homeostatic information (Craig, 2002; Dillon et al., 2008; Holroyd & Yeung, 2012; Lamm & Singer, 2010; Paulus, 2007; Paulus & Stewart, 2014; Rolls, 2016; Williams, Bush, Rauch, Cosgrove, & Eskandar, 2004). Importantly, within the AI, vAI is believed to be mainly involved in affective reactivity toward salient outcomes impacting self and others, and dAI (especially in the right hemisphere) has been more specifically linked to interoception, providing a bodily map for a wide range of mental processes (Craig, 2002, 2009; Critchley et al., 2004; Deen et al., 2011; Kober et al., 2008; Nomi et al., 2018; Uddin et al., 2014, 2017). Thus, AI, especially its right dorsal portion, might be key in processing representations of others' outcomes in prosocial behaviour considering interoceptive signals.

This functional distinction between vAI and dAI might predict different patterns of activity when these areas process self or other outcomes. Vicarious activity in vAI might overlap with first-hand responses, as suggested by empathy theories, following isomorphic representation of others' states (de Waal, 2008; Preston & de Waal, 2002). However, right dAI, being involved in the sense of body ownership and self-awareness, might be associated with distinctive representations for self and other outcomes. Indirect

evidence supports this hypothesis. First, extraordinarily prosocial individuals show higher overlap in left vAI responses to self and others' pain but not in right dAI (Brethel-Haurwitz et al., 2018). Furthermore, although patterns of activation within left AI between self and other negative outcomes are shared, the right AI does not follow this pattern, being specific to the target (Corradi-Dell'Acqua, Tusche, Vuilleumier, & Singer, 2016). Moreover, participants induced by a pain analgesia procedure showed reduced responses to self and vicarious pain in left but not right AI (Rütgen et al., 2015). Finally, higher activation to others' pain in vAI, but not dAI, was associated with more aversion to harm others, while dAI signalled self costs (Hu et al., 2021). These results suggest a functional distinction between AI subregions in prosocial behaviour, with vAI being associated with other-regarding mechanisms, while dAI anchored to self.

Taken together, the complimentary roles of ACC and AI in guiding prosocial actions does not mean that they perform the same computations, nor that subregions within these areas have the same functions. Interoceptive ability can shed light on the distinctive functions of ACC and AI, suggesting that AI, especially its right dorsal portion processes others' outcomes relying on bodily signals.

1.4.5 ACC and AI as a hub evaluating prosocial benefits and costs

Taken together these results indicate that ACC and AI process benefits to others, including avoiding harm and obtaining rewards, and show variability linked to traits and behaviours that are prosocial. This evidence all points to the ACC and AI, as well as other anatomically connected regions, as being crucial for making cost-benefit evaluations that guide decisions of whether to be prosocial. However, to date there is limited research examining this question, and even less that links it to individual differences in people's levels of prosocial motivation. In this thesis I will examine prosocial behaviour within a cost-benefit framework and examine its links to variability in ACC and AI function.

1.5 The present thesis

In this thesis I will use a cost-benefit framework to probe prosocial motivation. I will test specific hypotheses about how individual differences in prosocial motivation under this framework depend on sensitivities to personal cost and social benefits. Using this approach, I aim to illuminate cognitive and neural mechanisms underlying high prosocial dispositions across different prosocial domains. Namely, in Chapter 3, I examine the association between people's willingness to prosocially exert effort and their aversion to harm others, and the idiosyncratic cost-benefit processing that characterises people who are prosocial in both domains. Using fMRI, I will link effort-related behaviour to the function of the ACCg and AI when choosing whether to avoid harm to another. In Chapter 4, I identify the traits that are the most strongly associated with the shared variance between prosocial effort and harm aversion, among a pool of affective (e.g. empathy), psychiatric (e.g. personality disorders) and subclinical (e.g. apathy) traits. In Chapter 5, I examine whether people who are more hypocritical, i.e. judge others harshly for transgressions that they themselves have committed, are less willing to exert effort to benefit others, suggesting reduced prosocial motivation. Chapter 6 uses fMRI to investigate whether interoception is associated with sensitivities to either the social benefit or the personal cost in prosocial behaviour, and whether ACC and AI have distinctive or/and common roles in this relationship. Finally, in Chapter 7, dynamic costs are introduced into this cost-benefit model, deconstructing rewards into different components (rewards being received vs rewards in the environment), examining whether only some aspects of rewards shape prosociality. This unified conceptual framework allows me to focus on individual differences in prosocial behaviour across different prosocial domains, shedding light on the psychological and neural features of highly prosocial individuals.

Chapter 2

General Methods

2.1 Summary

In this thesis I research prosocial behaviour through five different studies with the aim of shedding light on the neural, psychological and computational mechanisms underlying cost-benefit prosocial decisions. Even though each particular study aims to understand a specific aspect of prosociality, similar methods were used across experiments. In this chapter, I will outline those common aspects, which include general procedures, manipulation of the social partners, two tasks that measure different aspects of prosocial behaviour, and analysis of prosocial decisions through computational modelling and mixed effects statistical models.

2.2 General Procedure

2.2.1 Participants

For the lab-based studies (Chapters 3, 5-7), participants were recruited from the University of Oxford and local residents of Oxford, U.K. Experiments occurred at University of Oxford facilities - behavioural sessions were conducted at the Department of Experimental Psychology, while fMRI sessions took place at the Wellcome Centre for Integrative Neuroimaging (WIN). To ensure a homogeneous sample and remove potential confounds, participants with a history of neurological or neuropsychiatric disorders, psychoactive medication or drug use, pregnancy, and aged outside the 18-40 range were excluded from participation. Furthermore, people with more than two years of study in

psychology were also excluded, avoiding the possibility that knowledge of the field could affect their decisions. Finally, participants who had previously participated in studies involving social interaction or electric shocks were also excluded due to concerns that prior experience with similar experimental settings would influence psychological and brain processes tested in the studies.

All participants in all studies of this thesis gave written informed consent and were compensated for their time. All studies were approved by Medical Sciences Interdivisional Research Ethics Committee, University of Oxford.

2.2.2 Apparatus

All prosocial and fMRI tasks reported in Chapters 3, 5-7 were programmed and presented to participants using Psychtoolbox 3 implemented in Matlab 2012a. All online resources across chapters were programmed and presented to participants using Qualtrics.com©, including the online prosocial tasks used in Chapter 4, all the scales included in the different studies, and the debriefing questions at the end of the experimental sessions.

Data handling was performed in Matlab 2017a. Behavioural analyses were performed in SPSS 26 and R 4.0.0 across the different studies (see the different chapters' Methods sections for details). Computational parameters were estimated individually for each participant using nonlinear optimisation implemented in Matlab 2017a for maximum likelihood estimation. Brain data were analysed using SPM12 and Python (see Chapters 3 and 6 for details).

Prosocial behavioural tasks and scales were completed by participants in a desktop PC. For the fMRI tasks, stimuli were projected from a laptop computer. For the prosocial effort tasks used in Chapters 3, 5 and 6, force was recorded using a hand-held TSD121B-MRI in a BIOPAC MP-160 system. For the harm aversion task, electric shocks were

delivered using a Digitimer DS7A electric stimulator. For the interoception task used in Chapter 6, white noise was played through Philips SHP2000 Over-Ear Corded Audio Headphones in the internal condition, and through the speakers of a Dell Inspiron 14 3000 laptop computer for the external condition. In this same task, participants blew a Mini Wright Standard Peak Flow Meter, and used a standard sleep blindfold. For each participant in this task, disposable mouth pieces were used.

2.2.3 Social manipulation in lab-based studies: Role Assignment Procedure

In order to probe prosocial behaviour, participants made decisions in two kinds of trials in the different tasks that they completed across studies: *self* trials, where the beneficiary/recipient of the decision's outcome is the decision-maker, i.e. participants themselves, the Deciders (see below); and *other* trials, where the beneficiary/recipient of the outcome is another unknown person, the Receiver (see below). With these two kinds of trials, decisions can be disentangled between self-benefit actions, and prosocial actions that benefit others. This is crucial as people vary greatly in how they make decisions across social and non-social settings. Thus, comparing behaviour in *other* and *self* trials can determine the variance that is unique to prosocial acts. Therefore, in order to derive reliable conclusions about prosocial behaviour from these studies, it is critical that participants truly believe that the outcomes of their decisions have a real consequence for another person.

I followed a rigorous protocol to manipulate the social aspect of the tasks performed by the participants. This protocol has been proven successful in making people believe and trust in the social consequences of their actions in *other* trials (e.g. Crockett, et al, 2014, 2017; Lockwood, Hamonet, et al., 2017, Figure 2.1). This social manipulation protocol is framed as a "role assignment" procedure. Participants are told that two participants take part of the experiment, and that there are two roles involved which

determine the characteristics of the tasks that each participant performs. These roles correspond to the Decider on one hand, and the Receiver on the other. The Decider performs a task in which they sometimes make decisions that could benefit themselves, and other times that could benefit the Receiver. The Receiver performs a completely different task that does not benefit anyone. These roles are assigned randomly by selecting one of two balls from a box. However, and unbeknown to the participants, in reality all participants are assigned the role of the Decider, as the Receiver is a confederate. Once all the data are collected for each study, participants are contacted by email and debriefed about the deception used in the social manipulation.

The series of events to complete the protocol are the following. In the preparation stage, participants are inside the experimental room with the experimenter who explains the role assignment procedure. Then, the random assignment occurs, when a second experimenter arrives at the experimental room with the confederate standing on opposite sides of the participant in a semi-opened door, with the participant and confederate obscured from each other view (Fig. 2.1A). Both participants and confederates are instructed not to talk out loud, and they are handed a yellow rubber glove before selecting the ball from the box, preventing features that could reveal their identities from being seen. Once anonymity has been ensured, both participants and confederates are told to place their gloved hand in front of the door and to wave to one another in order to make sure that they are aware of the other person's presence. Then they proceed to take a ball from the box held by one of the experimenters (Fig. 2.1B). The order in which they take the ball is decided by tossing a coin. Once the random assignment is completed, the second experimenter and the confederate leave, and participants are informed that they will play the role of the Decider (Fig. 2.1C). Participants are informed that their choices in the tasks are anonymous, that they will not meet or interact with the Receiver, and that Receivers will perform a different unrelated task in parallel. These instructions aim to minimise concerns about reputation, social desirability and reciprocity in their decision-

making, all variables that can influence social behaviour (Chung & Monroe, 2003; Rilling & Sanfey, 2011).

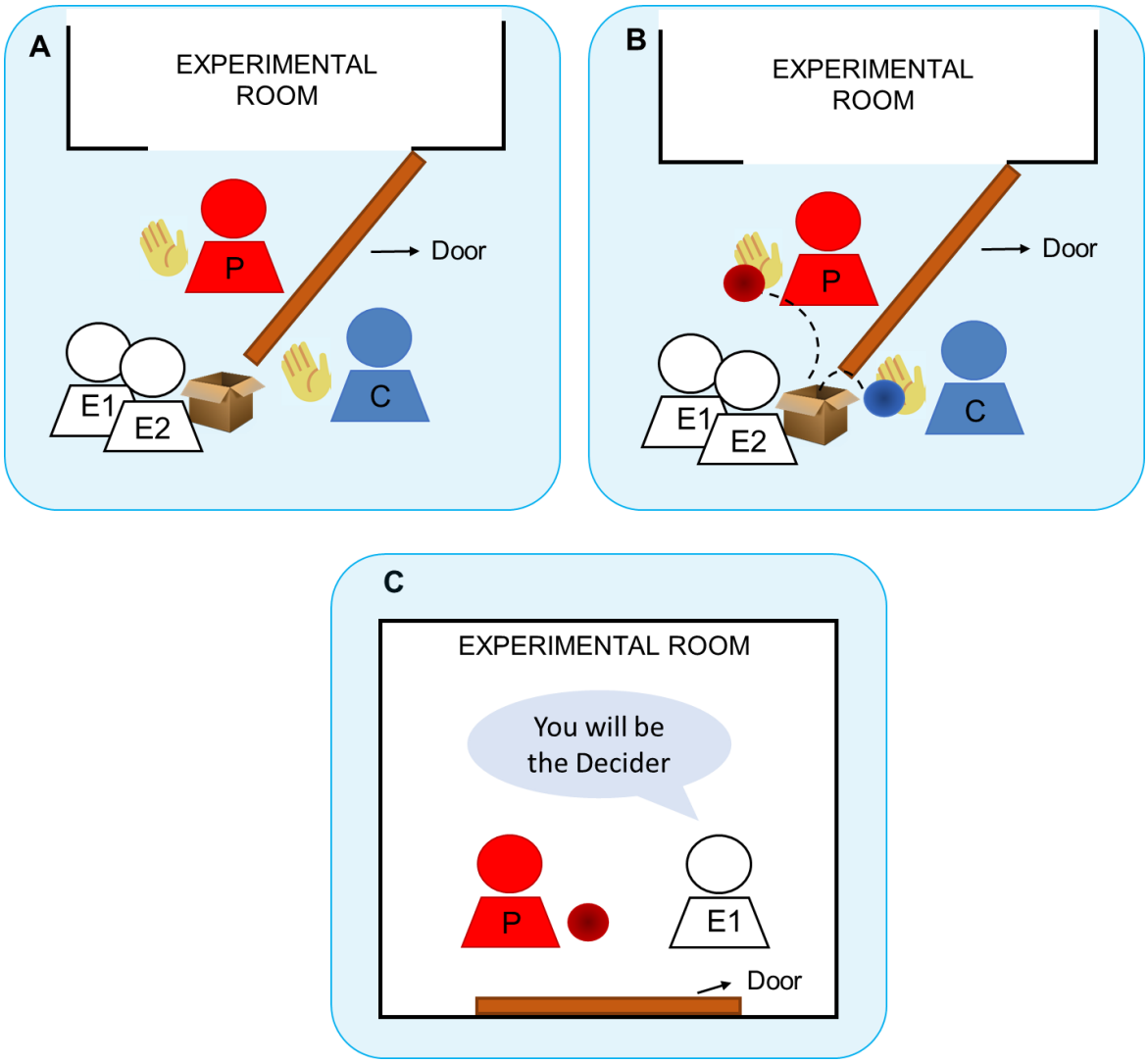


Figure 2.1 Social manipulation in lab-based studies. A. *Random assignment – waving hands.* With the door opened, Participant P and the confederate C stand on the opposite sides of the door, while Experimenters E1 and E2 instruct them to be silent, and waive their hands, which are covered by a yellow rubber glove, to one another. B. *Random assignment – selecting balls.* P and C take a ball from a box hold by one of the experimenters. The order in which P and C take the ball is determined by tossing a coin. C. *Role information*. Once the balls have been selected, C and E2 leave while P and E1 enter the experimental room. E1 informs P that they have been assigned the role of the Decider

To ensure that participants believed the social manipulation and the existence of the Receiver, participants are asked debriefing questions at the end of each study. These questions aim to provoke participants to indicate their beliefs about the deception

throughout the experiment. Crucially, these questions are subtle, and none of them directly ask the reliability of the procedure, avoiding inducing concerns about the veracity of the manipulation. If a subject reports being aware of the deception during the experiment, they are excluded from subsequent analyses.

2.2.3.1 Second social manipulation in multistage, multitasks experiments.

Chapters 3, 5 and 6 are multistage, multitask studies that involved two sessions: an fMRI one, where participants completed the required task while in the scanner, and a behavioural one, where they performed the task in a desktop computer. In these studies, the social manipulation described above always occurred in the first fMRI session, before getting into the scanner. However, in the second behavioural session participants completed the prosocial effort task (see below), which also involves outcomes for another person, and therefore, requires social manipulation. Thus, a second social manipulation was completed in the behavioural session, similar to the fMRI session one but with some important differences.

At the beginning of the prosocial effort task, participants receive instructions and perform its first stages (Fig. 2.2A-B; see next section for details). Next, participants are introduced to the main decision task and its social aspects. Here, participants are told that they will be paired with one of the receivers of the previous fMRI session, but not necessarily the same one they had in their scanning session. A second experimenter then comes to the room, and both the confederate/receiver and the participant stay in opposite sides of a semi-open door, waiving their hands to each other, similar to the random assignment procedure in the fMRI session (Fig 2.1A). After this, both the second experimenter and the receiver leave, and participants proceed to perform the prosocial effort task.

Thus, in the second social manipulation the random role assignment through selecting balls (Fig. 2.1B) was removed. Having been given the role of decider a second time in a row could have be seen as unlikely by participants (Kahneman & Tversky, 1972), and might make them question the credibility of the procedure. Furthermore, the 'waiving hands' step was kept in this second social manipulation to make sure that participants are aware of the receiver presence, making the social aspect of the decision-making task more believable.

2.3 Measuring two manifestations of prosocial motivation

In Chapters 3 to 6, prosociality is probed through decisions in two different tasks that measure two broad dimensions of prosocial behaviour – harm aversion and prosocial effort. Harm aversion is a moral, normative dimension that proscribes harming others, with people typically being more averse to harming others than themselves (Crockett et al., 2014; Fehr & Fischbacher, 2003; Smith, 2010). Prosocial effort is an instrumental aspect of prosocial behaviour, where people must be sufficiently incentivised by others' rewards in order to overcome effort costs (Contreras-Huerta et al., 2020; Lockwood, Hamonet, et al., 2017). Even though these tasks measure two different aspects of prosocial behaviour, both require participants to weigh up costs and benefits to make decisions according to their preferences. In the following, I describe in detail the lab-based versions of these tasks used in Chapters 3, 5-6.

2.3.1 Prosocial Effort Task

In the prosocial effort task (Lockwood, Hamonet, et al., 2017), participants have to decide whether to exert force to obtain real monetary rewards for themselves and for another unknown person, probing the willingness of putting in effort to benefit others

compared to self. This task can be divided in three different stages: (i) calibration stage, (ii) practice stage, and (iii) main decision task.

2.3.1.1 Calibration stage

Effort levels in the prosocial effort task are operationalised as different degrees of grip force according to each participant's maximum voluntary contraction (MVC). Thus, in the calibration stage the MVC is calculated for each participant. At the beginning of the task, participants are told to grip a handheld dynamometer three times within a five second window per trial (Fig. 2.2A). In the first trial of the calibration, participants are instructed to squeeze the handle with as much force as possible. The force exerted by the participant is represented on the screen by a rectangle bar getting filled. In the second trial, participants are instructed to exert force that equals a specific goal illustrated by a yellow line in the bar to be filled. The goal in this second trial is calculated extracting the MVC from trial 1 multiplied by 1.1. Finally, for the third trial participants are given the same instruction as for trial 2. However, the goal for this last trial is calculated as the highest value of force across trials 1 and 2, multiplied by 1.05. The MVC across any of the 3 trials is used to create different effort levels in the rest of the experiment.

This calibration stage has two aims. First, the effort levels must not depend on variations of each individual's strength, being instead relative to each participant's MVC which represents a subject-specific effort threshold. Second, the effort levels extracted from the MVC must be achievable for all participants. This is important as otherwise effort avoidance could be interpreted as risk aversion (Kahneman & Tversky, 1979). Thus, the calibration trials ensure that rewards in the main decision task are being discounted by effort and not by risk of not achieving the goal.

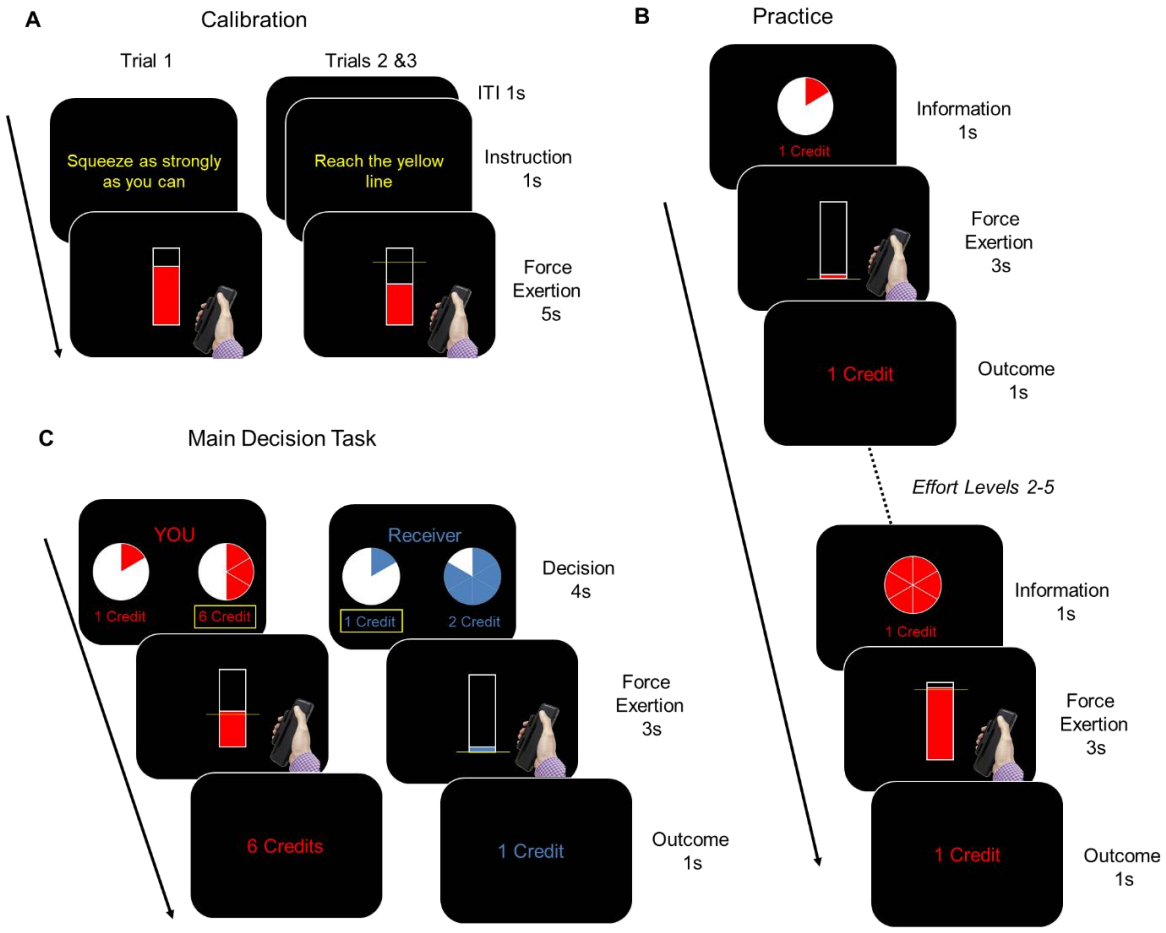


Figure 2.2. The Prosocial Effort Task. A. *Calibration stage*. Participants complete 3 trials where they squeeze a handle with their dominant hand for 5 seconds, to calculate their maximum voluntary contraction (MVC). In trial 1 (left), participants are instructed to squeeze as strongly as they can. Force exerted is represented by a bar getting filled on the screen. For trials 2 and 3 (right), participants are instructed to reach the yellow line in the bar with their force. The yellow line is fixed according to participants' performance in the previous trials. MVC corresponds to the maximum force across the 3 trials. B. *Practice stage*. Participants complete 12 practice trials where they experience each effort level (0, 30, 40, 50, 60 and 70% of their MVC) in an increasing order twice. Effort levels are represented by segments in a pie chart, such that the lowest level of effort corresponds to one element in the pie chart, while the highest is illustrated by six elements of the pie. After being informed about the level of effort to perform in each trial, participants exert the required force squeezing the handle for at least 1 second out of 3 seconds. C. *Main Decision Task*. Participants decide between a 'rest' option (0% of their MVC), associated to 1 credit, and a 'work' offer, with variable higher effort (30-70% of their MVC) and higher reward (2-10 credits) associated to it. Once participants make their decision, they must exert the required force for at least 1 out of 3 seconds. They do not earn rewards if they fail to do so. Importantly, participants themselves receive the reward in half of the trials (*self* trials, in red), while in the other half the reward is for the Receiver (*other* trials, in blue). This task then measures the willingness to exert effort to obtain rewards for self, separately from the willingness to exert effort to benefit another.

2.3.1.2 Practice stage

Once MVC has been calculated, participants completed a practice stage where they get familiarised with the force requirements for each effort level and their depiction on the screen (Fig. 2.2B). Participants experience each effort level in increasing order twice, completing 12 trials across the practice stage. The five effort levels were defined as 30, 40, 50, 60 and 70% of participants' own MVC, plus a rest baseline corresponding to a 0% of the MVC.

For each trial, participants first see an information screen for one second depicting the effort level to perform and the amount of reward in exchange. Effort is represented by segments of a pie chart, such that one element means rest, while from two to six elements represent the five increasing levels of effort, respectively. Rewards are represented by different number of credits. Participants are told at the beginning of the task that these credits are converted into real money at the end of the study multiplying them by a factor only known by the researcher. However, during the practice trials only one credit was on offer. This was to ensure that participants performed each effort level accurately (Manohar et al., 2015; Pessiglione et al., 2018). After the information screen, participants must exert the force required indicated by a yellow line in the bar to be filled. The force they are exerting is shown to them in real-time by the bar being filled in red proportional to their grip. To achieve the goal and get the credit on offer, participants must equal the required level of effort for at least one second within a three second window. If they do so, participants get a screen for one second indicating that one credit was earned. Otherwise, another screen appears instead showing that zero credits were earned. Thus, this practice stage ensures participants are familiar with the force required to obtain rewards.

2.3.1.3 Main Decision task

Participants then are ready to trade-off rewards against effort in the main decision task (Fig. 2.2C). In this task, participants are presented with two options: a baseline 'rest' option, associated with no effort required (i.e. 0% of their MVC) for one credit, and a 'work' offer, where a higher amount of rewards is offered in exchange of one of the five different levels of effort. There are five levels of rewards for this offer, namely 2, 4, 6, 8 and 10 credits. Crucially, on half of the trials participants make choices and exert effort to benefit themselves, i.e. the Decider (*self* trials), while on the other half they decide and exert force to benefit another unknown person, i.e. the Receiver (*other* trials). In this setting, participants are told that the money accumulated in *self* and *other* trials will be delivered to themselves and the Receiver at the end of the experiment, respectively. Participants are told that their identities are confidential, and that the Receiver is not aware of a bonus earned by another person's work, ensuring there were no reciprocity or reputational effects on behaviour.

Every trial starts with a decision screen, where the rest option and the work offer are presented. Similar to the practice stage, levels of effort are represented by segments in a pie chart and number of credits are explicitly stated in text. Participants choose one of the options by pressing a button, which highlights it in yellow. Each effort/reward combination for the work offer is repeated three times per beneficiary condition (i.e. 75 trials for *self* and 75 for *other*), and the effort and reward levels are varied independently over the trials and pseudo-randomised across participants. Once four seconds has passed, the force exertion screen is displayed, where participants must perform the required effort level following the same protocol completed in the practice stage. If participants do not make any choice within the four seconds window of the decision stage, they rest for three seconds (0% MVC) but they earn zero credits for that trial, encouraging people to make active choices. At the end of each trial, the number of credits earned is displayed for one second. All trials are the same duration regardless of what choice is made, avoiding any

possibility of choices being made due to temporal discounting effects (Critchfield & Kollins, 2001).

2.3.2 Harm Aversion Task

The harm aversion task (Crockett et al., 2014, 2015, 2017) probes harm aversion in a precise manner, capturing the variability in which people conform to this norm. This task measures how much people are willing to pay in order to avoid units of pain inflicted to themselves and to a stranger. Thus, this task assesses how monetary reward is devalued by pain separately for self and other, allowing the moral component of harm aversion to be disentangled from a self-centred valuation of pain. This task contains two stages: (i) a pain thresholding procedure, and (ii) the main decision task.

2.3.2.1 Pain thresholding procedure.

Degrees of pain are operationalised in this task as number of electric shocks, which are consistently considered as a painful stimulus across people. At the beginning of the session, and before the main decision task itself, a standard procedure is performed in order to determine each participant's pain threshold for electric shocks (Fig. 2.3A). This procedure has two aims. First, it standardises the pain units to be traded-off in the decision task across participants. Thus, the electric shocks created for the main task are adjusted according to each participant's own pain threshold, being mildly painful but tolerable. Second, this procedure allows participants to experience this unit of pain prior to making decisions in the harm aversion task, getting familiarised with the outcomes of their choices.

In this procedure (Crockett et al., 2014), two electrodes are placed on the back of the participant's left wrist. The first part of the thresholding procedure consists in a series of electric shocks that participants rate according to their pain experience on an 11-point

scale, from 0 (“I don’t feel the stimulus in my skin”) to 10 (“I don’t want a second stimulus of that kind”). This procedure starts with a low-voltage electric shock (0.1 mA), followed by a series of shocks that vary in their intensity in a small milliamp ratio of 3:1. This first stage of the thresholding stops when the participants rate one of the shocks as 10. This procedure is repeated three times per participant, and the shock rated as 10 in the third block is recorded as the threshold.

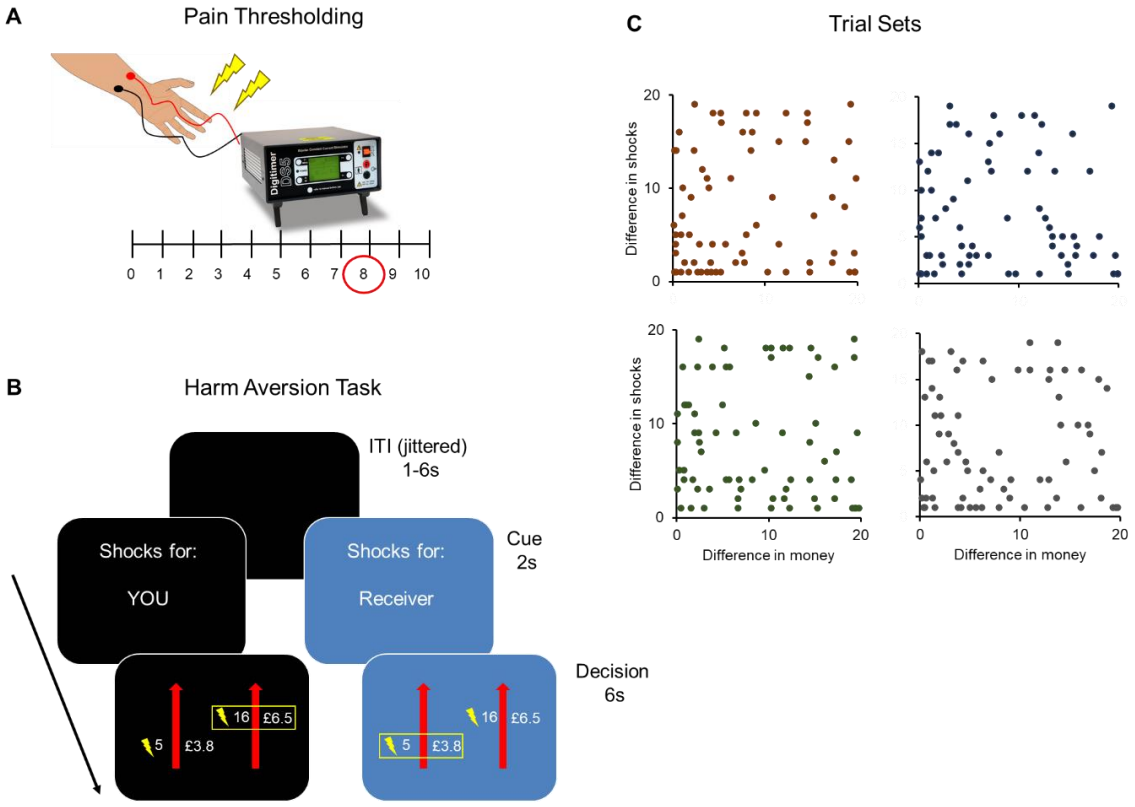


Figure 2.3. Harm Aversion Task protocol. A. *Pain thresholding procedure*. Prior to the decision-making component of the task, participants complete a pain thresholding procedure to (i) get familiarised with the pain units used in the harm aversion task and (ii) to determine the physical intensity of the shocks that participants will trade-off against money in the decision task. The intensity of the shocks is matched to a subjective rating of 8 in a 0-10 pain scale across participants. B. *Harm Aversion Task*. Participants must decide between a ‘helpful’ option, associated with a lower amount of money and shocks, and a ‘harmful’ option, with higher money and number of shocks. In half of the trials, shocks are for participants themselves (*self* trials, black background), while in the other half shocks are for the Receiver (*other* trials, blue background). C. *Trial sets*. Four trial sets with different combinations of money and shocks were used for the harm aversion task and counterbalanced across participants. These trial sequences ensure that the amounts of profit and the amount of shocks offered to participants are uncorrelated. Scatter plots show how differences in money and shocks between the ‘helpful’ and ‘harmful’ options do not correlate across trial sets ($|r| < 0.08$, $p > 0.5$). Each dot is a trial, while each plot is one trial set.

To ensure that participants are familiar with the experience of pain, they subsequently complete three blocks of seven shocks each. Shocks are randomly delivered ranging from 40-100% (with an interval of 10%) of the participants' maximum threshold, with participants rating each of them after receiving it. These shocks are fitted into a sigmoid function to estimate the current-to-rating response curve. Thus, a shock intensity matched with a level 8 of participants' pain experience (estimated from the derived function) is fixed as pain unit to be traded-off in the decision task. Participants then are explicitly told that this level will be the outcome of their decisions in the main task.

2.3.2.2 Main decision task: Harm Aversion.

In this task, participants trade-off money against electric shocks (Fig. 2.3B). Crucially, in half of the trials the electric shocks are delivered to participants themselves, i.e. the Decider (*self* trial), while in the other half the shocks are received by a stranger, i.e. the Receiver (*other* trials). The money, however, is always rewarded to the Decider across trials. Nevertheless, neither money nor shocks are delivered during the task, ensuring choice independence and avoiding habituation to the painful shocks. Participants are told that one trial (either *self* or *other*) will be randomly selected from the whole protocol and implemented once the task is finished. Thus, if a *self* trial is selected, shocks and money corresponding to participants' choice in that trial are delivered to them at the end of the session. However, if an *other* trial is selected, participants receive the money and they are told that shocks will be delivered to the Receiver at the end of their session.

On each trial of the harm aversion task, participants have to choose within a six seconds window between two options: a harmful option of more shocks and more money, and a helpful option that results in fewer shocks and less money. Once they make their decision, that option is highlighted in yellow. If they fail to make a choice within those six seconds, the missed trial is repeated at the end of the session. Transitions between the

recipient of the electric shocks are cued with an instruction screen lasting two seconds. The order of *self* and *other* trials is pseudo-randomised such that no more than four consecutive trials are for the same recipient. An interval jitter is included after each trial lasting one to six seconds.

The sets of trials used in this task were optimised to give the most efficient estimates of harm aversion parameters (κ, see Computational Modelling section) from participants' decisions. 72 trials were created following the criteria given in Crockett et al. (2014). This trial optimisation is introduced to satisfy two constraints. First, trials must be optimised to give the most efficient estimates of potential participants' harm aversion parameters κ_{self} and κ_{other}. Second, the relative amounts of profit and pain that would result from participants' choices must be decorrelated across trials. To fulfil these constraints, a set of 72 trials were created, each containing a pair of choices that matched the indifference point of a specific κ value. Thus, a set of evenly distributed κ values were used, ranging from 0.05 to 0.95. For each κ value, 10,000 random pairs of positive shock movements (i.e. difference between the shock options, Δs) and positive money movements (i.e. difference between the money options, Δm) were created. These optimised pairs [Δs, Δm] were transformed into choices containing different amounts of shocks and money for the helpful and the harmful option as follows: the shocks in the helpful option was a positive integer between 0 and 20, randomly drawn from a uniform discrete distribution with the constraint that $0 < \text{shocks} + \Delta s < 20$. Similarly, money for the helpful option was a positive number between 0 and 20, randomly drawn from a uniform discrete distribution, rounded to the nearest 10th and constrained such that $0 < \text{money} + \Delta m < 20$. For the harmful option, shocks and money were then set by adding Δs and Δm to the amount of shocks and money in the helpful option, respectively. Following this process, a set of 72 trials were generated, such that money and shocks in the helpful options were lower in amount than the harmful options.

Next, the procedure above was repeated 10,000 times. For each iteration, correlations between Δm and Δs were calculated, and the trial sets with the lowest correlations were selected. Four different sets of 72 trials were selected in this way ($|r| < 0.08$, $p > 0.5$), counterbalanced across participants (Fig. 2.3C). For each of these sets, four catch trials were added where the harmful option had lower amount of money compared with the helpful option, resulting in 76 trials per set. Half of the trials in each set were randomly selected to fix the harmful option in the right side of the screen, while the other half had the harmful option in the left side. Finally, the 76 trials of each set were duplicated to have identical trials in *self* and *other* conditions. Thus, participants completed 152 trials distributed in two runs of 76 trials each (half *self* and half *other* per run).

2.4 Analysis of behavioural data

Across different studies of this thesis, the primary outcomes are the decisions people make in the two tasks outlined above. These decisions were typically analysed using two approaches. First, I used computational models in order to precisely formalise the decision processes underlying participants' behaviour. In these models, a specific relationship between the variables involved in the decision is assumed and tested on participants' behaviour (e.g. reward discounted parabolically by effort). Secondly, I used mixed effects statistical models to analyse raw task behaviour without assuming a formal relationship between the variables involved in the decision. Using this tool, the influence of each separated variable can be tested (e.g. sensitivity to reward vs effort). In this section, I describe in detail these two different approaches.

2.4.1 Computational Modelling

Computational models have a number of advantages compared with model-free statistical analyses (Busemeyer & Diederich, 2010, 2014; Crockett, 2016). First of all, it gives the opportunity to discover latent variables not apparent in behavioural decision data alone. This requires a higher degree of abstraction – the formalisation of the relationship between elements that guide decisions, i.e. how rewards are discounted by effort and pain in the prosocial effort and harm aversion tasks, respectively. Secondly, it provides more predictive power, with precise quantitative predictions based on underlying theoretical assumptions. Thus, through the quantification of different parameters that signal the weight each participant places on latent psychological phenomena, we can falsify cognitive hypotheses more strictly. Finally, this approach addresses a more mechanistic integration of the different elements of a decision, not only merely describing the effects but also giving a formal definition of the function involved. This mechanistic approach can also explain more accurately how the brain operates when making a choice. For instance, neurons in frontal areas can encode multiplexed representations of variables involved in a decision, creating an abstract neuronal currency, i.e. neurons represent value as a function of effort and reward (Kennerley, Dahmubed, Lara, & Wallis, 2009). This is particularly important for building bridges between brain and behaviour, with a common theoretical formal framework across cognitive and neural mechanisms.

Here, I will describe the two models that were fitted to participants' decisions in three of the five empirical chapters of this thesis (Chapters 3-5). These models precisely quantify the influence of effort on the devaluation of money in the prosocial effort task, and the influence of pain on the devaluation of money in the harm aversion task. The way in which these models predict decisions is by estimating the underlying subjective values from the relationship between objective features of the options offered, i.e. amount of reward and amount of effort/pain for each option. It is assumed that these valuations guide decisions of whether to choose an option or not, such that the action with the

highest subjective value is chosen over other available actions.

2.4.1.1 Prosocial effort task

Participants' choices in the prosocial effort task were fitted using a model that describes the degree to which rewards are parabolically discounted by effort (Chong et al., 2017; Hartmann et al., 2013; Lockwood, Hamonet, et al., 2017). It is largely assumed that effort discounts rewards, and organisms seek to minimise effort when they can (Hull, 1943). Research in non-social effort-based decisions has mostly proposed models in which rewards are discounted by effort either linearly (i.e. constant discount across effort levels), hyperbolically (i.e. convex, lower levels of effort impacting more subject's decisions than higher levels) or parabolically (i.e. concave, higher levels of effort impacting more decisions than lower). Even though there are different reasons to theoretically propose each discount model, studies comparing the performance between them have revealed that parabolic models fit better to effort-based decisions when the cost is physically demanding (Chong et al., 2017; Hartmann et al., 2013; Klein-Flügge, Kennerley, Saraiva, Penny, & Bestmann, 2015). Indeed, these models have been compared in the context of the prosocial effort task, supporting the idea that rewards are in fact parabolically discounted by effort for both *self* and *other* (Lockwood, Hamonet, et al, 2017, Lockwood et al., 2021). In the latter studies, models containing parameters across *self* and *other*, or separate parameters for *self* and *other*, were also compared. Results showed that a parabolic model with two different discount parameters for *self* and *other* explained most of the variance in participants' choices.

Thus, following the parabolic model used in Lockwood, Hamonet, et al., (2017), devaluation of reward by effort for *self* and *other* can be represented by the following formula for trial *i* and participant *j*:

$$SV_{ij} = R_i - \lambda_j {E_i}^2$$

$$\lambda = \begin{cases} \lambda_{self} \; if \; self \; trial \\ \lambda_{other} \; if \; other \; trial \end{cases}$$

Where *SV* is the subjective value of the work offer with a reward *R* (i.e. number of credits) and effort *E* (i.e. effort level). The degree by which reward is discounted by effort is weighted by the discount parameter *λ* and is different for *self* and *other* trials. The parameter λ is estimated across all trials and thus reflects how participants discount all the effort and reward combinations for *self* and *other.* Hence, when λ is 0, the participant does not discount reward by effort. However, when λ increases, the reward is increasingly discounted by effort.

2.4.1.2 Harm aversion task

For the harm aversion task, decisions were fitted into a model that explains choices by differences in subjective values between harmful and helpful options. This model has been validated extensively in previous studies, showing that it outperforms a range of alternative models (Crockett et al., 2014, 2015, 2017). The model describes the difference in value between the options of trial *i* for participant *j* as follows:

$$\Delta V_{ij} = \left(1 - \kappa_j\right)\Delta m_i - \kappa_j \Delta s_i$$

$$\kappa = \begin{cases} \kappa_{self} \; if \; self \; trial \\ \kappa_{other} \; if \; other \; trial \end{cases}$$

Where *ΔV* is the difference between the subjective value of the harmful and helpful option, and *Δm* and *Δs* are objective differences in money and shocks between harmful and helpful options respectively. In this model, Δm and Δs are weighted by a free parameter *κ*, which represents harm aversion. When κ is 0, any profit will be accepted regardless the amount of shocks associated to it. When κ approaches to 1, participants

become more harm averse, sacrificing an increasing amount of profit to avoid additional shocks. Depending on who is receiving the shocks, κ_{other} captures the subjective cost of harming the receiver (*other* trials), while κ_{self} explains the cost of harming oneself (*self* trials).

2.4.1.3 Transforming value into choice probability: the softmax algorithm.

Both models described above assume that participants' decisions are driven by value, i.e. participants assign value to potential choices, choosing the option with the highest one. Therefore, converting the value space into a probability distribution is crucial to understand how people make their decisions. One approach that is commonly used in cognitive science is the softmax algorithm (Sutton & Barto, 1981), which transforms choice value into probabilities. The softmax function converts the value of a choice into the probability of having chosen given its value, in comparison with the value of other potential actions. Even though it is assumed that people will choose the option with the highest value, the softmax function allows some randomness (or noise) in the decision process. Consequently, the softmax function can be expressed as the following:

$$P(a)_{j(i)} = \frac{\exp\left(\frac{V_{j(i)}}{\beta_j}\right)}{\sum_i \exp\left(\frac{V_{j(i)}}{\beta_j}\right)}$$

Where the probability *P* of the chosen action *a* in the trial *i* for participant *j* is a function of the exponent of value *V* of that chosen action over the sum of the exponent of the values of all potential choices. Numerator (action chosen) and denominator (the total of available choices) are weighted by a stochasticity parameter β, estimating the noise or temperature of participant's choices, which scales their sensitivity to value. Thus, β defines the linearity of the slope. When β approaches 0, the slope becomes more linear, meaning more noise in choice probability. While β gets higher, the sigmoid approximates a step function, making choice preferences deterministic. From this function it can be

deduced that an action becomes more likely as its value approaches to the value of all the potential choices available, following the assumption of that the option with the highest value is chosen.

Thus, for both the prosocial effort and the harm aversion tasks, subjective value was converted into probabilities using a softmax function as the one described above. For the prosocial effort task, the softmax converts the value of the work offer SV such that:

$$P(work)_{j(i)} = \frac{\exp\left(\beta_j \, SV_{j(i)}\right)}{\exp\left(\beta_j \, 1\right) + \exp\left(\beta_j \, SV_{j(i)}\right)}$$

Where *P* is the probability to work of a participant *j* in a trial *i*. When a choice of rest, $1\text{-}P_{j(i)}$ is used to calculate the probability of not choosing to work. On the other hand, for the harm aversion task the softmax function is used to transform the value of the harmful option relative to the helpful option, ΔV, such that:

$$P(harm)_{j(i)} = \frac{1}{1 + exp^{-\beta_j \, \Delta V_{j(i)}}}$$

Both prosocial effort and harm aversion models have free parameters that weight the information involved in the decision. The discount parameter λ and the harm aversion parameter κ need to be estimated for *self* and *other* for each participant, together with their temperature β in the softmax. Maximum likelihood allows for estimation of the values of the parameters that best fit participants' behaviour from a reasonable range that contains all the possible combinations of the free parameters. Thus, the probabilities of the choices given by the softmax algorithm increase, such that:

$$L = \sum_i \ln P(a)_{j(i)}$$

Where the likelihood of the set of parameters *L* is calculated from the natural log of the probability of the action *a* in trial *i* by participant *j* given by the model. When the model is perfect, L = 0 as $P(a)_{j(i)}$ is 1. Then, if P decreases, the likelihood that the free parameters explain the underlying decisional process decreases, and L becomes

negative. Then, to choose the best values that fit into model decisions:

$$\theta' = \arg\max \theta \ (L)$$

This function identifies the set of parameters θ that gives a log-likelihood L closest to zero, giving the parameter with the highest probabilities to reflect participants' decisions. Thus, all parameters used in both the harm aversion task, i.e. κ_{self}, κ_{other}, β_{self}, and β_{other} (Crockett et al., 2014, 2015, 2017) and the prosocial effort task, i.e. λ_{self}, λ_{other} and β (Lockwood, Hamonet, et al., 2017, Lockwood et al., 2021), were estimated individually for each participant using nonlinear optimisation implemented in Matlab (MathWorks) for maximum likelihood estimation.

2.4.2 Mixed Effects Models

Even though computational modelling is a very powerful tool to analyse behaviour, statistical techniques applied to raw behavioural data are also fundamental in order to infer conclusions about prosociality. Computational models assume a specific interaction between objective aspects of a decision (e.g. reward is discounted linearly by pain and parabolically by effort in the harm aversion and the prosocial effort tasks, respectively), and the computational parameters obtained from these models reflect these psychological mechanisms. However, disentangling how different variables affect prosocial decisions (e.g. effort, reward), without assuming a specific interaction between them, can allow for different questions to be addressed. For instance, with this approach, specific hypotheses can be tested on whether a specific trait modulates how people weight effort, reward or both when they make a prosocial decision. Furthermore, it can test whether sensitivity to the benefit, costs, or both, predominantly guide some specific actions. Therefore, these analyses can be helpful to understand how people process different pieces of information when acting prosocially.

I used mixed models instead of most standard statistical techniques such as repeated measures ANOVA because they provide important advantages (Field & Wright, 2011; Harrison et al., 2018; Singmann & Kellen, 2019). First, this approach allows one to account for subject-specific variance, solving potential violations of the independence principle (see below). This makes the results more accurate and decreases the inflation of Type I errors. Secondly, mixed models can account for binary distributions such as decisions between two options. Third, as mixed models consider the idiosyncratic variability of each subject, it can address problems associated with missing data and different number of data points per subject. Therefore, where possible, in this thesis I have applied the principles described below. Exceptions for this are noted, in particular in Chapter 4, which I conducted and published before I began implementing mixed-effects models.

In the following, I will explain briefly how mixed-effects models work. This section does not intend to be an exhaustive description of this technique (see instead Field & Wright, 2011; Harrison et al., 2018; Singmann & Kellen, 2019). Rather, here I explain the main ideas underlying this statistical tool.

2.4.2.1 Random-intercepts model: solving data independence violation

The best way to understand mixed effects modelling is with a specific example. Let us focus on the force exertion data from the prosocial effort task explained above. In each trial, participants choose between a rest option and a work offer associated to some reward in exchange of some effort, exerting the force required every time they choose the latter (Fig. 2.2C). In this example, each participant will have a number of data points that will depend on their decisions to either work or rest - i.e. someone who is effort averse and rests a lot will have far fewer trials where they exert force at the higher effort levels, than a motivated person who works on the majority of trials. For the sake of simplicity, let us imagine a simple model where the research question is how force exertion varies as a

function of the reward offered. Thus, in an ordinary linear regression, the (simplified) model would look like the following:

$$F_i = \beta_0 + \beta_1 r_i$$

Where *F* is the force exerted for trial *i* with a reward *r*. Then, the β parameter for the variable of interest can be calculated, and this will be the same for all trials and for all participants. This type of parameters is known as *fixed effects*, as they are constant. Fixed effects are the effects commonly used in ordinary regression models and they are estimated at the level of the population, either as a main or as an interaction effect (Fig. 2.4A).

Even though this model can give us important insights about the force exerted by participants, there is an inherent problem in this data set that makes this approach not ideal. It is reasonable to assume that the behaviour of a participant in one trial will be correlated with their behaviour in other trials, as the force exerted in one trial depends on each participant's pattern of decisions. Thus, this kind of data violates the statistical assumption of data independence, i.e. the probability that the value of a data point is independent from the value of the other data points. Independence violation can lead to an increase in Type I error.

Mixed-effects models (from now on MMs) are a generalisation of common regression models that deal with the independence violation, capturing the dependence of each cluster of data points through *random effect* parameters, as opposed to the fixed effects. This improves the estimation of the effects and the statistical power without inflating false positives. Random effects cluster random or stochastic variability from different sources. Thus, in the case of force exertion, this grouping variable could correspond to participants, i.e. all force data points are grouped separately for each subject. The way in which MMs do this is allowing the intercept to vary according to each participant. Thus, the previous linear equation becomes:

$$F_i = \beta_{0j[i]} + \beta_1 r_i$$

Where the intercept β_0 is calculated for each participant *j*, while the slope for reward, represented by β_1, remains constant. This model is typically called a *random-intercept model* (Fig. 2.4B). While in ordinary linear models the regression line is applied to all participants, being less accurate and with high residual errors, in a random-intercepts model the variance associated with each participant's idiosyncratic pattern in force exertion is accounted for, i.e. the deviation from the fixed intercept. These deviations follow a normal distribution with a variance estimated by the model. Thus, MMs use random intercepts that account for the correlated data points that exist in the data coming from a common cluster, e.g. participants, something that standard regressions do not account for.

2.4.2.2 Random-slopes model: Model complexity in MMs

Even though the previous model mitigates the dependency problem associated to the force exertion data, it does not account for all the potential dependencies that underlie this design. The previous random-intercept model assumes that the differences between reward levels are similarly perceived across participants. But this is not necessarily the case. It is reasonable to assume that there might be differences across participants in how sensitive they are to reward units. For instance, a person could be mainly indifferent if a reward increases one unit, while another could be much more willing to exert high levels of force for only one unit more. Thus, if we want to acknowledge these differences in reward sensitivity across participants, we should also allow random slopes. In principle, one could fit a model where we are only interested in controlling for reward sensitivity, without assuming clusters across participant i.e. different intercepts. Such a "pure" *random-slopes model* can be expressed as the following:

$$F_i = \beta_0 + \beta_{1j[i]} r_i$$

Where the intercept is a constant, fixed effect for all subjects, but the slopes vary across participants (Fig. 2.4C). Even though there must be experimental designs and research questions where a model like that is relevant, in the force exertion example this is unlikely. If people vary in how sensitive they are to different magnitudes of rewards (random-slopes), they probably also vary in how much force they exert on average (random-intercepts). Thus, this more complex model could be expressed as the following:

$$F_i = \beta_{0j[i]} + \beta_{1j[i]} r_i$$

Where both intercepts and slopes are adjusted for every participant (Fig. 2.4D). Commonly, random slopes are used together with random intercepts because both the starting point and the degree of the slope vary across participants. It is important to note that random effects do not alter the interpretation of the fixed affects. To answer the question whether reward influences the amount of force exerted in a trial, we need to look at the fixed effect of reward on force exertion, i.e. the effects at the population level. In this example, the random effect is included to control for potential variability across participants and deal with independency violation. Thus, in a MM typically both fixed and random effects are included, allowing the model to estimate parameters for single individuals, but within a gaussian structure that considers the whole data set. Thus, even though MMs acknowledge variability across participants, they also account for these individuals to be similar in some respect.

In this simple example, going for a full MM is easy and does not require neither theoretical complexity nor computational/statistical power. However, this is different when a more elaborate design is considered. For the force exerted example, we could think in more complex questions related not only to reward influencing force exertion, but also the effort level, and the beneficiary of the outcome, together with their interaction. Thus, in terms of model complexity, in principle, we could estimate the variance of the random factor (i.e. participants, random-intercepts), and also the variance for each within-subject

main and interaction effects (i.e. reward x effort x beneficiary, random-slopes). Introducing a maximal model like this allows one to account for all potential dependencies and variances associated to our data, protecting our results against Type I errors (Barr, Levy, Scheepers, & Tily, 2013; Judd, Westfall, & Kenny, 2012; Schielzeth & Forstmeier, 2008). Therefore, some authors have recommended that the maximal model always should be fitted where it converges and where the data is not overfitted (Barr et al., 2013).

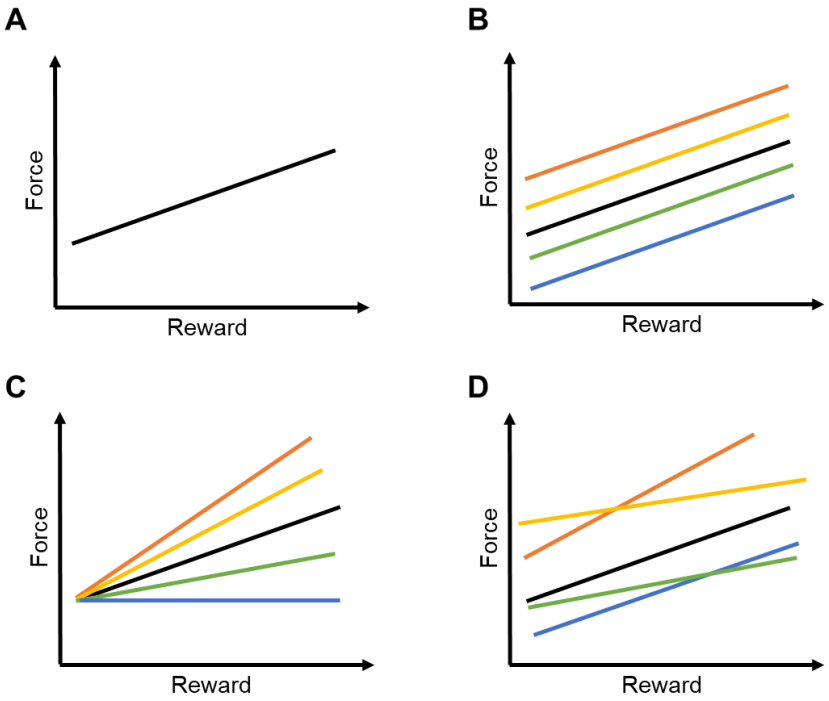


Figure 2.4 Different Linear Models Structures. A. *Fixed-effect model*. Simple regression model where the regression line is fixed across all participants. It is assumed in this example that reward has the same effect on force exertions for all subjects, such that more reward incentivised more force. B. *Random-intercept model*. Mixed effects model where the intercept is randomised while the slope remains constant. In the example, the black line corresponds to the fixed (overall) effect of reward on force exertion, while colour lines correspond to four different subjects. Thus, lines are adjusted for each participant's motivation pattern – i.e. some participants exert more force for the same reward than others. Note however that the lines follow the same slope across participants. C. *Random-slopes model.* Mixed effects model where the slopes are randomised while the intercept remain constant. In the example, participants exert the same force for low reward, but weight differently the other levels of reward, with people having steeper slopes than others. D. *Random-intercepts and random-slopes model.* Mixed effects model that combines the models described in B and C, i.e. both intercepts and slopes are randomised. Participants exert force differently for similar rewards, and at the same time they weight differently the variation in reward units. *Note:* Black line corresponds to fixed effect and colour lines to random effects. The plots are hypothetical representations of the role of reward on force exertion in the prosocial effort task, and they do not illustrate real data.

However, the literature is far from being universal in the support of this recommendation. In fact, model complexity has been a debated topic in the field for years, with no straightforward procedure to establish how complex a MM must be. There are multiple problems associated with the maximal model approach – among others, maximal modelling can overfit the data, going against the parsimonious principle (i.e. Occam's razor); it can significantly increase Type II error rate, as the inclusion of random effects decreases statistical power; and it might not make sense from a theoretical point of view (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; Vandekerckhove, Matzke, & Wagenmakers, 2014). In fact, the maximal model rule has shown to be over conservative beyond the nominal Type I rate, and a model that facilitates the highest statistical power within a typical α criterion could be more appropriate (Matuschek et al., 2017; Stroup, 2012).

Different solutions have been proposed for model complexity. First, from a data-driven perspective, model comparison between more complex and simpler models using goodness-of-fit tests like Akaike information criterion (AIC) procedure can be used to decide which model fits better the data (Seedorff, Oleson, & McMurray, 2019; Vandekerckhove et al., 2014). Secondly, models can be reduced excluding random effects that are highly correlated between each other (correlations of ± 1 or close, which suggests model overfitting), that do not show much variance within the model, or/and that have only a small number of levels, e.g. < 5 (Bates, Kliegl, Vasishth, & Baayen, 2015; Bolker, 2015; Singmann & Kellen, 2019; Westfall, Kenny, & Judd, 2014). Third, from a theoretical point of view, it has been advised that random slopes should not be used where there are not clear hypotheses due to Type II inflation, accepting at the same time that using only random-intercepts could inflate Type I error (Harrison et al., 2018). From this perspective, only a powerful theoretical background, clear experimental designs, and prior analyses plans can guide a proper model structure.

In this thesis, MMs were used mainly to test for effects of specific variables on prosocial decisions and actions. Specifically, in Chapter 3, I test whether the potential relationship between prosocial effort and harm aversion are triggered either by sensitivity to self costs, to other benefits, or both, illuminating the mechanisms underlying prosocial decisions across domains. In Chapters 5 and 6, I test whether certain traits (i.e. moral hypocrisy and interoception, respectively) modulate either sensitivity to effort, to reward or both on prosocial effort. Given that these hypotheses distinguish sensitivity to self costs and other benefits, MMs were favoured respective to computational models as they allow to disentangle those effects separately. Finally, they are also used in Chapter 7, in which I develop a new prosocial foraging, opportunity-cost paradigm and test for specific effects of different sources of reward information on patch-leaving decisions. Importantly, I follow the different suggestions conveyed above in building model complexity, with a focus on a more theoretical approach rather than a maximal model guideline, but also using data driven recommendations when required. The rationale of the complexity of these models are indicated in each specific chapter.

Chapter 3

Common mechanisms underlie prosocial behaviour in different domains

3.1 Summary

Prosocial behaviours are actions that benefit others at a personal cost. For instance, obtaining rewards for another person can require exerting effort, i.e. prosocial effort; and to avoid harming others we might lose out some of the benefits we can obtain, i.e. harm aversion. Individual differences in the sensitivity to others' outcomes are linked to people's willingness to incur costs to benefit others and also to activity in the ACC, AI, and TPJ. However, it is unknown whether sensitivity to outcomes in one prosocial domain can predict prosocial behaviour in another domain with outcomes of an opposite valence, i.e. whether sensitivity to others' harm is linked to sensitivity to others' rewards. Here, I examined this question in a study where participants completed two prosocial decision-making tasks – an fMRI harm aversion task, where participants traded-off profit against electric shocks delivered to either themselves or another person; and a prosocial effort task, measuring participants' willingness to exert physical effort to reward either themselves or another person. Behaviourally, results revealed that people who were more averse to harm others were more willing to exert effort to obtain rewards for others. Crucially, individual differences in ACC, AI and TPJ activity signalling others' harm in the harm aversion task predicted sensitivity to others' rewards in the prosocial effort task. These findings suggest that people who are prosocial in different contexts might be driven by others' outcomes regardless their valences, suggesting common mechanisms across domains.

3.2 Introduction

Prosocial behaviour, i.e. costly actions that benefit others, is a key component of healthy social relationships, ensuring social cohesion and group bonding (de Waal, 2008; Fehr & Fischbacher, 2003). Yet, an agent's motivation to act prosocially varies depending on the characteristics of the others' benefit and the costly action associated to them. For example, a person might be prosocial when reducing others' harm, but might be reluctant to perform effortful actions to reward others. Previous studies suggest that prosocial behaviours in these different contexts might have distinctive psychological and neural mechanisms underlying them, affecting the extant in which they manifest (Carnes & Janoff-Bulman, 2012; Contreras-Huerta et al., 2020; Decety & Cowell, 2017; Janoff-Bulman et al., 2009). However, there are people who are consistently prosocial across contexts, regardless of whether these behaviours involve preventing others' harm or working for others' rewards. But, what are the mechanisms that drive prosocial behaviour across different contexts?

People learn from early years the principle of 'do no harm' through socialisation (Decety & Cowell, 2017; House, 2018; Tremblay, 2000). This is a strong moral norm that can be found across multiple cultures (Decety & Cowell, 2017; Gert, 2004; Turiel, 2002). This norm proscribes harming others, and people can incur great costs to prevent others' suffering (Marsh et al., 2014; Rand & Epstein, 2014). Studies have found that actions that could result in harm, even hypothetically, are experienced as distasteful and aversive, with people being typically reluctant to perform them (Cushman et al., 2012; Greene et al., 2004; Lockwood et al., 2020; Miller & Cushman, 2013). Consistently, a series of studies (Crockett et al., 2014, 2015, 2017), have shown that individuals are 'hyperaltruistic' in a harm aversion context, being willing to sacrifice more personal profit to avoid painful shocks delivered to others compared to themselves.

In contrast, actively working to obtain rewards for others does not follow the guidance of a proscriptive moral norm. Effort is aversive and people typically avoid it, needing increasing amount of rewards as effort gets higher to perform an action (Kool & Botvinick, 2018; Kurzban et al., 2013; Le Heron et al., 2018). Thus, it is expected that when the action benefits another person, people are even more averse to effort compared with when the benefit is for oneself, with only those who are sufficiently incentivised by others' rewards being willing to overcome such a cost. Lockwood and colleagues (Lockwood, Hamonet, et al., 2017) indeed showed that people on average are 'prosocially apathetic', being more motivated to work to get rewards for themselves than for others. Thus, when preventing harm, people tend to be more willing to incur costs to benefit others than in a prosocial effort context. Indeed, in a study where people showed hyperaltruism, participants were also less willing to harm themselves to benefit others relative to self, and this effect was uncorrelated with aversion to harm others (Volz et al., 2017). Thus, the valence of others' outcome/benefit in prosocial behaviour shapes the motivation of agents to act.

When getting rewards for others by incurring aversive effort, people must be highly sensitive to others' benefits to overcome the costs. Since people typically value more their own than others' rewards (Camerer, 2011), they are more willing to incur costs to benefit themselves than others. In contrast, the mere action of profiting while harming others can be aversive, and therefore, avoided. Profiting by harming oneself does not have that aversive connotation, and people are more willing to accept rewards even if that results in an aversive stimulus. Therefore, the moral norm proscribing harming others might devalue the otherwise rewarding experience of obtaining profit (Stellar & Willer, 2014). The latter hypothesis was tested by a recent fMRI study where neural responses to the profit received by harming self and others were dissociated from the neural signal to the harmful outcome (Crockett et al., 2017). Thus, the authors tested whether neural sensitivity to profit, harm or both predicted hyperaltruism. Results revealed that people

who exhibited less striatal activity to profit earned by harming others relative to self were more hyperaltruistic, suggesting that moral transgressions disrupt neural representations of action value. How could harm aversion be associated with prosocial effort, when the latter seems to be focus on others' outcomes/benefit rather than the value of the action?

Importantly, this fMRI study also found that self and others' harm was signalled by the ACC, AI and TPJ, but their activation patterns did not predict hyperaltruism. However, previous studies have shown that individual differences in the responses of ACC, AI and TPJ to vicarious pain and negative outcomes predict prosocial behaviour (Greening et al., 2014; Hein et al., 2010; Hu et al., 2021; Masten et al., 2011; Tusche et al., 2016). Importantly, these areas do not only signal vicarious pain, but also track anticipated or received rewards by others (Apps & Ramnani, 2014; Basile et al., 2020; Braams et al., 2014; Chang et al., 2013; Lockwood et al., 2015; Lockwood, 2016; Mobbs et al., 2009; Morelli et al., 2015), and individual differences in their signal have been associated with prosocial behaviour (Cutler & Campbell-Meiklejohn, 2019; Fukuda et al., 2019; Harbaugh et al., 2007; Hare et al., 2010; Hutcherson, Bushong, & Rangel, 2015; Park et al., 2017), including at the cost of putting in effort (Bortolini et al., 2017; Le Bouc & Pessiglione, 2013). Thus, individual differences in the sensitivity to others' outcomes regardless of their valence is signalled by ACC, AI and TPJ activity, which might facilitate prosocial behaviour in different domains.

Consistently, ACC and AI have been identified as a core system in empathy, being suggested as a neural circuit processing salient outcomes received by self and others that elicit affective and motivational states (Corradi-Dell'Acqua et al., 2016; Fan et al., 2011; Lamm et al., 2011; Lockwood, 2016). TPJ, on the other hand, has been proposed as a more socio-cognitive counterpart, involved in signalling inferences of affective and mental states of other people (Decety & Grèzes, 2006; Döhnel et al., 2012; Mahy, Moses, & Pfeifer, 2014; Saxe, 2010; Saxe & Powell, 2006; Shamay-Tsoory, 2011). Thus, these areas might process others' outcomes as a potential benefit for another person

regardless of their valence, which may drive prosocial behaviour. From this perspective, it might be possible that people who tend to be prosocial across different contexts are more sensitive to others' outcomes in different domains, signalled by ACC, AI and TPJ. This *other-regarding* hypothesis proposes that variability in neural responses to others' harm during harm aversion decisions might predict being incentivised by others' rewards in prosocial effort. Since harm aversion has been reported to be driven mainly by moral principles, this signal does not predict hyperaltruism (Crockett et al., 2017). However, in a context where the sensitivity to others' benefits is crucial, such as prosocial effort, it could predict prosocial behaviour.

However, a competing hypothesis predicts different relationships between ACC, AI and TPJ activity with prosocial effort. Some authors have proposed that the role of these areas in prosocial behaviour is tracking conflict between self-interest and moral principles (Chang, Smith, Dufwenberg, & Sanfey, 2011; FeldmanHall, Mobbs, & Dalgleish, 2014; Greene et al., 2004; Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012; Obeso, Moisa, Ruff, & Dreher, 2018). From this *moral-conflict* hypothesis, higher activity to profit earned by harming others vs self signals the conflict of receiving a reward by immoral means, which could predict prosocial preferences when exerting effort for others. Furthermore, other two alternative hypotheses are possible considering the different combinations between profit/harm and reward/effort in each domain respectively. A *valence-dependant* hypothesis, predicting neural responses to others' harm (aversive stimulus) being associated with sensitivity to effort aversion, and a *self-centred* hypothesis, predicting that the self-costs across prosocial domains will be related instead, such that individual differences in signal to profit when harming others would predict effort sensitivity when working for others.

Here, I tested whether harm aversion and prosocial effort share variance, such that people who are more averse to harm others are more willing to exert effort to reward others. Crucially, I hypothesise that people who are prosocial in these two contexts are

more sensitive to others' benefits regardless of the valence of these outcomes in each task. Participants completed an fMRI version of the harm aversion task and a behavioural version of the prosocial effort task. I measured BOLD signal changes to profit and harm separately during harm aversion decisions for *self* and *other* within ACC, AI and TPJ, testing for their correlations with individual differences in behavioural sensitivity to effort and reward in effort-based decisions for *self* and *other* separately. Thus, in this chapter, I aimed to illuminate the common mechanisms that underlie prosocial behaviour across domains.

3.3 Methods

3.3.1 Participants

Eighty healthy volunteers aged 18-38 years were recruited from the University of Oxford and from local residents of Oxford. The study was approved by the University of Oxford ethics committee. Participants with a history of neurological or neuropsychiatric disorders, psychoactive medication or drug use and pregnancy were excluded from the study. Furthermore, participants who had previously participated in studies involving social interaction and who had studied at least one year of psychology were also excluded due to concerns that prior experience could influence psychological and brain processes in the study. All participants gave written informed consent and were compensated for their time with an average of £100 for their participation (£15 per hour for the fMRI session and £8 per hour the behavioural session) in the different stages of the study, plus some bonuses earned though their decisions in the tasks.

In the current study, 66 participants (age M = 22.6, SD = 3.7, 38 females) were included from the initial sample of 80 participants. Four participants failed to attend the behavioural session. Two participants expressed doubts about the social manipulation and the existence of the receiver and therefore were excluded from analysis. Seven

participants had excessive head motion in the scanner and/or technical problems with the harm aversion task. Finally, one participant's neuroimaging data was not registered due to technical issues of the scanner.

3.3.2 Procedure

Participants completed a multi-stage, multi-task study investigating social behaviour and cognition. Within this study, participants completed the harm aversion task while in the scanner, and at least one week later, participants completed the prosocial effort task in a behavioural session. Behavioural pilot data demonstrated that the order in which these tasks are completed does not alter prosocial decisions (for details see Chapter 4).

3.3.2.1 fMRI session: Harm aversion task

At the beginning of the session, participants completed a pain thresholding procedure and a role assignment protocol where they were randomly assigned the role of the Decider and another unknown person, who was actually a confederate, was assigned the role of the Receiver (see Chapter 2 for details). They were instructed that their decisions had consequences for both themselves and receivers, and that they were completely confidential and anonymous. They were also informed that receivers would complete different tasks compared with deciders, which did not involve outcomes for other people.

After practicing six trials, participants completed the version of the harm aversion task extensively described in Chapter 2 (Fig. 3.1A) while in the scanner. In this task, participants traded-off units of pain, represented by a number of electric shocks, against profit, in the form of money. In each trial, participants made a series of binary choices where one option, the *harmful* option, contained more electric shocks in exchange of higher amounts of money, while the other option, the *helpful* option, was associated with fewer shocks and less money. The money was always received by the participants

themselves, i.e. the Deciders, but the shocks were allocated to an unknown person on half of the trials, the Receiver (i.e. *other* condition), and to the participant in the other half of the trials (i.e. *self* condition). Importantly, no shocks nor money were delivered inside the scanner. Instead, one of the trials was randomly selected and informed to the participant outside the scanner. In case that the selected trial was a *self* one, participants received the corresponding electric shocks. In case that it was an *other* trial, participants were told that the shocks was going to be delivered to the receiver in another room. However, no shocks were actually given as the receiver was a confederate. Finally, the money associated with the trial was added to the participant payment and given to them at the end of the behavioural session.

3.3.2.2 Behavioural session: Prosocial effort task

At least one week after the fMRI session, participants attended a behavioural session where they completed a set of tasks, including the prosocial effort task. At the end of the session, participants answered final debriefing questions about the experimental setup and received reimbursement for their participation plus the financial bonuses earned throughout the study. Participants completed the version of the prosocial effort task described in detail in Chapter 2 (Fig. 3.1B), where different magnitudes of financial rewards are traded-off against different levels of effort. Money in this task was represented by a number of credits, while effort was operationalised as grip force. In every trial, participants decided between two options: a rest baseline, related to no effort and low reward (1 credit); and a work offer, associated with a higher amount of effort that varied across 30-70% of each participant's maximum voluntary contraction (MVC), in exchange of higher profit, i.e. 2-10 credits. Crucially, in half of the trials the reward was for participants themselves (*self* trials), while in the other half rewards were for an unknown person, the receiver (*other* trials). If participants chose the work option, they were required to perform the specific effort level squeezing a handle with their dominant hand

for at least one second in a three seconds time-window to obtain the reward in offer, getting zero credit for that trial if they failed to do so. If they rather chose to rest, they just waited for three seconds without making any effort and obtained the one credit baseline in exchange.

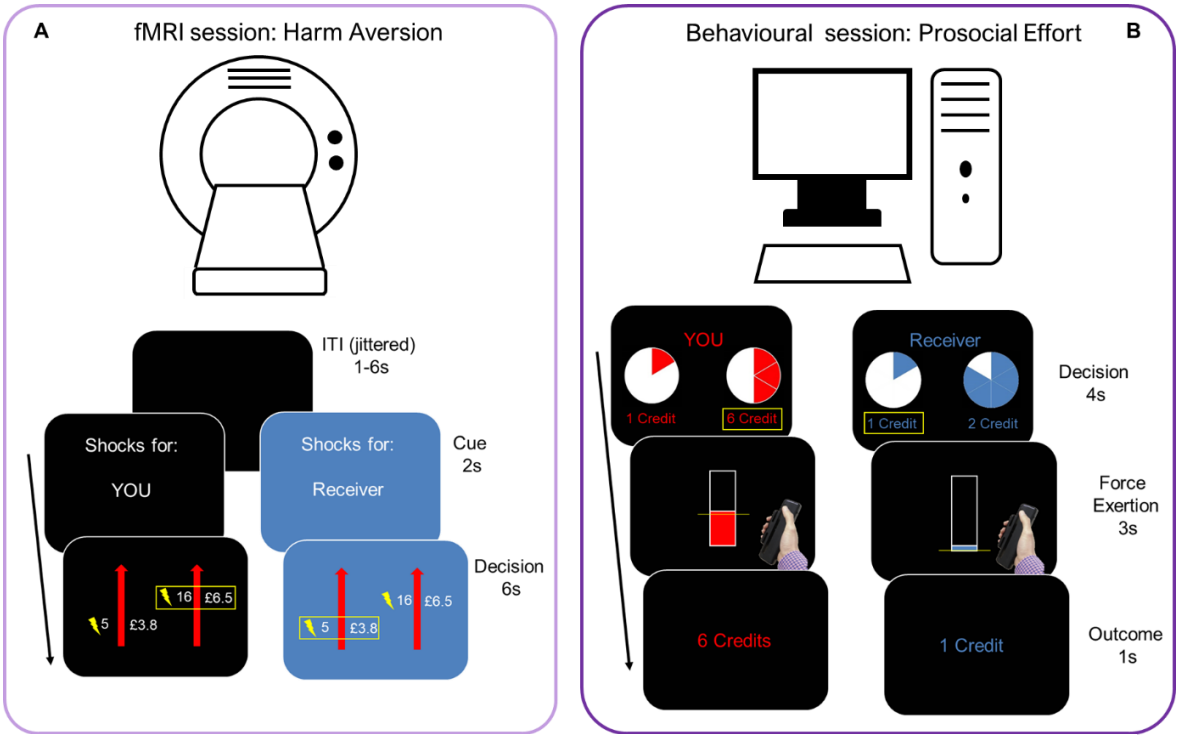


Figure 3.1 General Procedure. *A. fMRI session - Harm Aversion Task*. Participants completed the harm aversion task while in the scanner, where they decided between a 'helpful' option, associated with a lower amount of money and shocks, and a 'harmful' option, with more money for more of shocks. In half of the trials, shocks were for participants themselves (*self* trials, black background), while in the other half, shocks were for the Receiver (*other* trials, blue background). *B. Behavioural session – Prosocial Effort*. ≥ one week, participants undertook a behavioural session where they made prosocial effort decisions, choosing between a 'rest' option (0% of their maximum voluntary contraction - MVC), associated to 1 credit, and a 'work' offer, with variable higher effort (30-70% of their MVC) and higher reward (2-10 credits) associated to it. After their decision, participants had to exert the required force for at least 1 out of 3 seconds. They do not earn rewards if they fail to do so. Importantly, participants themselves received the reward in half of the trials (*self* trials, in red), while in the other half the reward was for the Receiver (*other* trials, in blue).

3.3.3 Analysis of behavioural data

Computational models were fitted to choices in the harm aversion and the prosocial effort tasks, which precisely quantify the degree in which pain devalues profit and the degree in which money is devalued by effort, respectively (see Chapter 2 for details). For

the harm aversion task, decisions were analysed using a model that calculates the subjective values between harmful and helpful options (Crockett et al., 2014, 2015, 2017). Thus, difference in value between the options can be described as follows:

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

$$\kappa = \begin{cases} \kappa_{self} \; if \; self \; trial \\ \kappa_{other} \; if \; other \; trial \end{cases}$$

Where ΔV is the difference between the subjective value of the harmful and helpful option, and Δm and Δs are objective differences in money and shocks between harmful and helpful options respectively. In this model, κ is a free parameter that represents how harm averse participants are. When κ is 0, participants will accept any offer regardless the pain in exchange. As κ increases and approaches to 1, participants get harm averse, paying more money to avoid additional shocks. When shocks are delivered to the other unknown person, κ_{other} captures the subjective cost of harming the receiver (*other* trials), while κ_{self} represents the cost of participants harming themselves (*self* trials).

For the prosocial effort task, effort-based decisions were fitted into a model (Lockwood, Hamonet, et al., 2017, Lockwood et al., 2021) that measures the degree in which effort discounts rewards parabolically, such that:

$$SV = R - \lambda E^2$$

$$\lambda = \begin{cases} \lambda_{self} \; if \; self \; trial \\ \lambda_{other} \; if \; other \; trial \end{cases}$$

Where SV is the subjective value of the work offer with a magnitude of reward R and effort level E. The free parameter λ weights the discount effect of effort on reward, such that when λ is 0, the participant does not discount reward by effort. As λ approaches to 1.5, effort discounts reward in a higher degree. λ was set with a maximum of 1.5 as any value beyond this constraint does not change the predicted decision in the task, i.e. the value of resting (i.e. 1 credit for 0 effort) is higher than the values for the work offer.

Finally, the parameter λ represents differently the discount of reward depending on whether it is getting by *self* or *other*.

A softmax function was used in both tasks to transform trial-by-trial differences in value into choice probabilities. All parameters described above were estimated individually for each participant using nonlinear optimization implemented in Matlab (MathWorks) for maximum likelihood estimation.

3.3.3.1 Statistical Analysis

Non-parametric tests were used to compare computational parameters within and between tasks as these measures were not normally distributed. Computational models were complemented with statistical analysis on raw decision data for each task to separate the effects of different sources of information on choices. Computational parameters represent a specific relationship between costs and benefits in the decision process (e.g. reward being discounted parabolically by effort), integrating both aspects of choice. However, using mixed-effect models (MMs), costs and benefits can be disentangled, being able to test whether the former, the latter, or both, are guiding *self* and *other* decisions. Thus, a MM for the prosocial effort task was built, such that:

$$DW_i = \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 R_i B_i + \beta_5 E_i B_i + \beta_6 E_i R_i + \beta_7 E_i R_i B_i$$

Where decision to work *DW* in the trial *i* is predicted by the fixed effects of reward *R*, effort *E*, beneficiary *B*, and their interactions. Random-intercepts were clustered in subjects *j*, and random slopes on effort and reward were included as it is expected that participants will vary in their sensitivities to this information. Model comparison using AIC between an only random-intercept MM and one with random-slopes confirmed that the latter improves model fitting (random-slopes model = 5830.0, random-intercept only model = 7037.2).

Another MM was built for the harm aversion task, such that:

$$DH_i = \beta_{0j[i]} + \beta_{1j[i]}\Delta s_i + \beta_{2j[i]}\Delta m_i + \beta_3 Re_i + \beta_4 \Delta s_i Re_i + \beta_5 \Delta m_i Re_i + \beta_6 \Delta s_i \Delta m_i + \beta_7 \Delta s_i \Delta m_i Re_i$$

Where decision to help *DH* in the trial *i* is predicted by the objective difference in shocks Δs and money Δm between options in i, and the recipient of the shocks *Re* for that trial i, together with their interactions. Like the prosocial effort MM, this model had random-intercepts clustered by subjects, and random-slopes for Δm and Δs, as sensitivities to these sources of information are expected to vary across participants. Looking at the AIC values, model comparison confirmed that adding random-slopes improved model fitting (random-slopes model = 6069.4, random-intercept only model = 7182.0). The outcome variable for all models, i.e. DW and DH, was a binary, factor variable in a logistic MM. These analyses were implemented in R using the *glmer* function.

3.3.4 fMRI acquisition and preprocessing

Functional MRI scanning was performed on a 3-Tesla Siemens Prisma scanner at the Wellcome Centre for Integrative Neuroimaging (WIN) at The University of Oxford. Functional images were obtained with multiband T2*-weighted echo-planar imaging (EPI) sequence. The EPI images were acquired in an ascending manner, at an oblique angle (approximately 30˚) to the AC-PC plane to minimize the signal dropout in the orbitofrontal areas. The following acquisition parameters were used: 72 slices in interleaved ascending order; matrix size: 108 ×108; voxel size: 2 × 2 × 2 mm^3 with 1 mm gap; echo time (TE) =30 ms; repetition time (TR) = 1,570 ms; flip angle = 70°; field of view (FOV) = 216 * 216 mm^2. The structural image was taken using a magnetization prepared rapid gradient echo (MPRAGE) sequence with 192 slices; TR=1,900 ms; TE = 3.97 ms; field of view = 192 × 192 mm^2; voxel size = 1 × 1 × 1 mm^3 resolution. We also acquired a field map (short TE =

4.92 ms; long TE = 7.38 ms; TR = 482.0 ms; resolution = 2 × 2 × 2 mm^3; FOV = 219 × 219 mm^2) to correct distortions in the functional images.

MRI data were preprocessed and analysed using SPM12[1]. Functional images were realigned and unwarped with reference to the fieldmap and co-registered to the participant's own structural image, correcting for distortions caused by susceptibility-induced field inhomogeneities. The EPI images were then realigned and coregistered to the subject's own anatomical image. The structural image was processed using a unified segmentation procedure combining segmentation, bias correction, and spatial normalisation to the Montreal Neurological Institute (MNI) template; the same normalisation parameters were then used to normalise the EPI images. Finally, images were spatially smoothed with an SPM default Gaussian kernel (8-mm full-width at half-maximum).

3.3.5 fMRI analysis

A general linear model (GLM) was built to identify regions in which the BOLD signal parametrically varied at the onset to the amount of money and the number of shocks that would follow after choosing the harmful option relative to the helpful option for *self* and *other* trials, time-locked to when the options were presented on the screen. In this model, BOLD signals were regressed against four critical regressors containing the onsets of (i) *self* trials where the left option was selected, (ii) *self* trials where the right option was selected, (iii) *other* trials where the left option was selected, and (iv) *other* trials where the right option was selected. The duration of these four regressors was set to the participant's reaction time on that trial (i.e., the interval between the presentation of options and button-press). For each of these four regressors, four parametric modulators were included: the amount of money and shocks for the harmful option and the helpful

[1] www.fil.ion.ucl.ac.uk/spm

option, regardless of what participants decided. These four parametric modulators competed for variance during the estimation, rather than being serially orthogonalized as is standard in SPM. Regressors of no interest were included corresponding to onsets of button presses, cue for transitions between conditions, and missing trials, as well as six nuisance regressors to control for head motion.

Using the model described above, I extracted the statistical maps for *self* and *other* conditions from anatomical masks of regions of interest (ROI) using *MarsBar* toolbox in SPM. I based my analyses on bilateral ACCg, bilateral dorsal and ventral portions of anterior insula (dAI and vAI respectively) and bilateral posterior and anterior portions of temporo-parietal junction (TPJp and TPJa respectively), making a total of five regions of interest. I used anatomical masks instead of peak voxels because of two reasons. First, as mentioned in Chapter 1, different portions of an area can be linked to different processes, and have different connectivity fingerprints (Apps, Rushworth, & Chang, 2016; Neubert, Mars, Sallet, & Rushworth, 2015; Vogt & Paxinos, 2014), making them incomparable. Second, in this thesis I am focusing on individual differences in hypothesis-driven ROIs, considering their involvement reported in previous work. Thus, it is possible that these regions might not be significant at the group level for effects across all participants, but they show enough variability in their pattern of activation to predict behaviour in another task instead. Therefore, these masks allow me to be anatomically precise in my predictions, be accurate in the structural characteristics of the hypothesised regions and subregions, and account for the individual variability in their pattern of activations.

For the AI portions, I used the parcellations created by Deen et al., (2011), which are derived from a voxel-wise k-means clustering approach applied to resting-state and divide the insular cortex in three different complexes for each hemisphere: posterior insula, dAI and vAI. This subdivision of the insula has been replicated in other work (Chang et al., 2013; Nomi et al., 2016) and used in other studies (Nomi et al., 2018; Uddin et al., 2014).

The 6 insula ROIs were acquired in 2 mm MNI space directly from the author website[2]. I fused left and right parcels to create bilateral AI masks using *imcalc* function in SPM12.

For the ACCg region, I used thresholded masks taken from the resting-state connectivity-based parcellation by Neubert et al. (2015). This parcellation divides the frontal cortex into 21 different regions. The parcels were acquired in 2mm MNI space directly from the author website[3]. I created the ACCg mask duplicating the left area 24ab of this parcellation onto the right hemisphere, creating a bilateral mask (the atlas was created in participants with a paracingulate sulcus only in the left hemisphere). From this bilateral 24ab, I also removed those voxels located in the posterior portion of the mask, to capture more closely what the literature has identified as ACCg (Apps et al., 2016; Apps & Sallet, 2017). I used only the gyrus portion of the ACC in this study as it is one of the subregion of the ACC that has been more widely associated with social processing, including vicarious pain and rewards (Apps et al., 2016, see Chapter 1). Finally, for the TPJ area I used the masks described in Mars et al. (2012), obtained from the same website[3], where they used tractography-parcellation and resting-state connectivity to divide TPJ in posterior and anterior portions for the right hemisphere. To create the left hemisphere masks, the original parcels were mirrored and then combined with the right one to create bilateral TPJ masks.

Thus, from these five regions (vAI, dAI, ACCg, TPJa and TPJp), I extracted the statistical maps for ∆m and ∆s in *self* and *other* trials and correlated them with the behavioural measures of prosocial effort, controlling for multiple comparison.

[2] https://bendeen.com/data/
[3] http://www.rbmars.dds.nl/CBPatlases.htm

3.3.5.1 Statistical analyses for the association between neural responses and behavioural measures.

To test whether parametric maps correlated with prosocial motivation, betas obtained from the GLM described above across the five ROIs were correlated with sensitivity to effort and reward in *self* and *other* trials, extracted from two MMs where decisions to work were predicted by effort, rewards and their interaction. Importantly, each MM considered only *self* and *other* trials respectively. These two models were clustered with random-intercepts for each participant, and had random-slopes for reward and effort, such that:

$$DW_{self} = \beta_{0j[self]} + \beta_{1j[self]}R_i + \beta_{2j[self]}E_i + \beta_{3[self]}E_iR_i$$

$$DW_{other} = \beta_{0j[other]} + \beta_{1j[other]}R_i + \beta_{2j[other]}E_i + \beta_{3[other]}E_iR_i$$

Where *DW* is specific for *self* and *other* trials separately. Thus, neural responses in the five ROIs for money and shocks in the harm aversion task were correlated with sensitivities (betas estimates) for effort and reward in the prosocial effort task in *other* and *self* trials. Pearson correlations were used to test for association between neural responses and behavioural measures. Visual inspection of the brain data revealed that there were some evident outliers in the statistical maps, which were located many standard deviations from the mean in the distribution. Thus, for each correlation between ROI response and behavioural measures, participants who had values below and above three standard deviations were not considered. All contrasts of the ROI beta values followed a normal distribution (Kolmogorov–Smirnov test). Likewise, all behavioural betas in the prosocial effort task were normally distributed. Results obtained from these correlations were corrected for multiple comparison across the five ROIs using false discovery rate (FDR, Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001).

To test whether correlations with fMRI responses were consistent with potential behavioural correlations between harm aversion and prosocial effort, an extra analysis

was performed. The behavioural effort and reward beta values for *self* and *other* in the prosocial effort task were correlated with behavioural sensitivities to money and harm separately for *self* and *other* in harm aversion task. To extract sensitivities for money and harm, two MMs were built with decisions to help in the harm aversion task being predicted by differences between money, shocks and their interaction in each trial. Each MM considered only *self* and *other* trials respectively. These two models were clustered with random-intercepts for each participant, and had random-slopes for money and shocks, such that:

$$DH_{self} = \beta_{0j[self]} + \beta_{1j[self]}\Delta m_i + \beta_{2j[self]}\Delta s_i + \beta_{3[self]}\Delta m_i \Delta s_i$$

$$DH_{other} = \beta_{0j[other]} + \beta_{1j[other]}\Delta m_i + \beta_{2j[other]}\Delta s_i + \beta_{3[other]}\Delta m_i \Delta s_i$$

Where *DH* is specific for *self* and *other* trials separately. Thus, betas for effort and reward on one hand, and for money and shocks for the other, were tested for associations using Pearson correlations.

3.4 Results

3.4.1 Hyperaltruism and prosocial apathy effects are replicated.

First, I tested whether behavioural results for each task, robustly reported in previous studies, were replicated in the current data set – hyperaltruism, i.e. people being more averse to harm others than themselves for profit (Crockett et al., 2014, 2015, 2017); and prosocial apathy, i.e. people being less willing to put in effort to benefit others compared with themselves (Lockwood, Hamonet, et al., 2017, Lockwood et al., 2021). I took two approaches to test these effects: MMs and computational modelling. For the former, decisions in both tasks were modelled using the different sources of information in each trial as predictors. Thus, for the harm aversion task, trial-by-trial decisions to help were

predicted by differences in money between options (Δm), differences in shocks between options (Δs*)*, the recipient of the shocks, and their interactions. Using this model, results revealed main effects of Δm (b = -4.22, SEM = 0.4, z = -10.67, p < 0.001) and Δs (b = 1.61, SEM = 0.18, z = 9.17, p < 0.001), having opposite effects on decisions. Thus, higher differences in money between options (i.e. the profit obtained by harming is higher) predicted less help, while a higher difference in shocks (i.e. choosing to harm implies higher levels of pain) predicted more help. Importantly, there was a main effect of recipient (b = 1.32, SEM = 0.09, z = 15.44, p < 0.001), such that participants were more willing to help when the recipient of the shocks was the receiver, replicating in this way the hyperaltruism effect (Fig. 3.2A). Extending these results, a Δm x recipient interaction was found (b = 0.62, SEM = 0.09, z = 6.86, p < 0.001), indicating that, even though as Δm gets higher people help to a lesser degree, this effect is more pronounced in *self* than *other* trials, with people being less likely to harm for high profit if the shocks are received by the receiver (Fig. 3.2B). No other significant effects were found.

With regards to the prosocial effort task, decisions to work were fitted into a MM where the predictors were effort levels, reward magnitude and beneficiary, together with their interactions. Results revealed main effects of effort (b = -2.52, SEM = 0.21, z = -11.82, p < 0.001), reward (b = 2.22, SEM = 0.18, z = 12.1, p < 0.001) and beneficiary (b = -3.45, SEM = 0.11, z = -32.13, p < 0.001). More importantly, significant interactions between effort x beneficiary (b = 0.26, SEM = 0.1, z = 2.61, p < 0.01) and between reward x beneficiary (b = -.093, SEM = 0.1, z = -9.62, p < 0.001) were also found. Participants were more reluctant to work when the effort level was high, and when the magnitude of the reward was low, and this effect was amplified when the beneficiary was the receiver (Fig. 3.2C and 3.2D), replicating the prosocial apathy effect. No other significant effects were found.

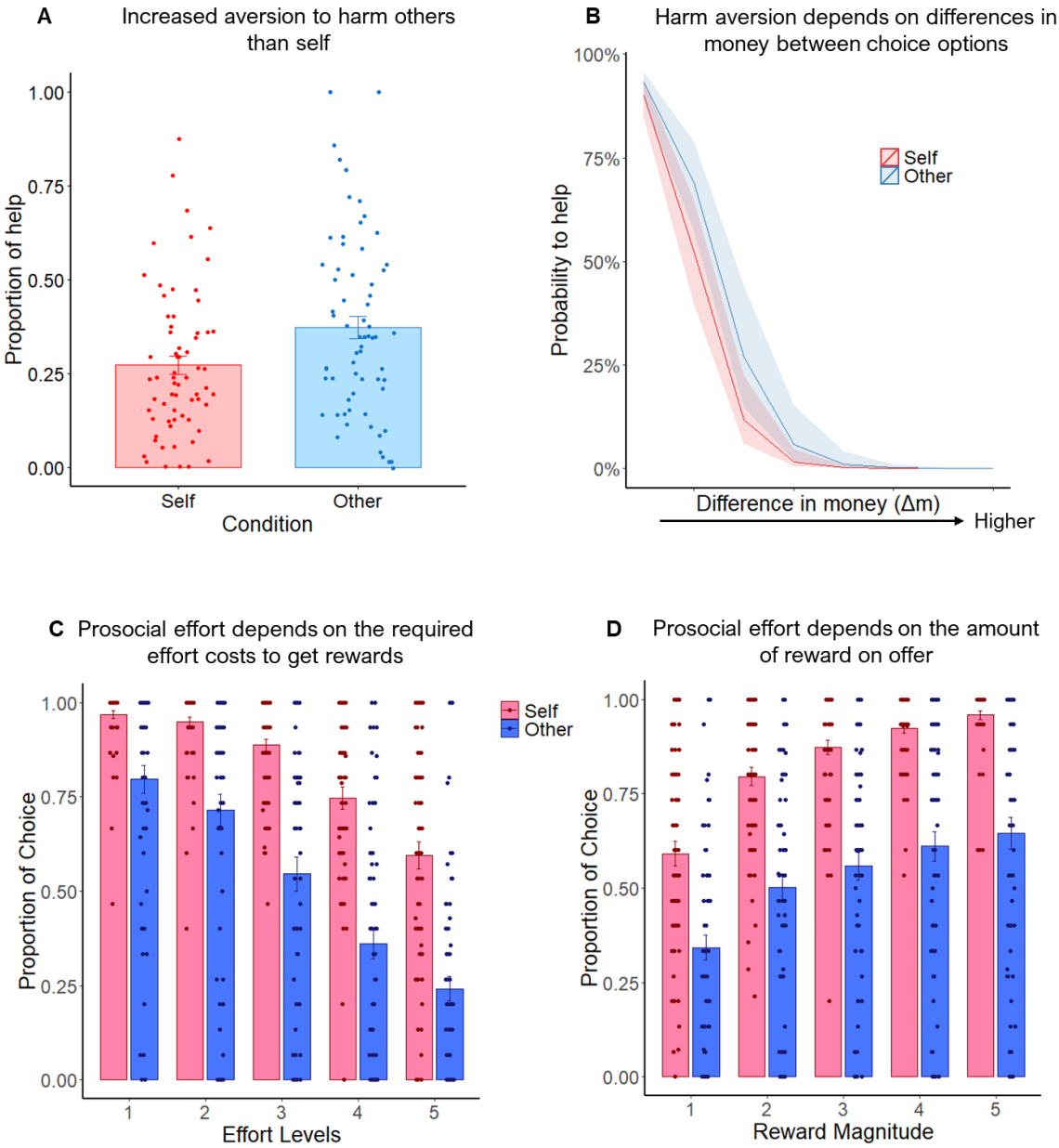


Figure 3.2. Hyperaltruism and prosocial apathy effects are replicated. A. Participants were more averse to harm others than themselves when trading-off electric shocks against profit. The y-axis indicates the proportion of helpful (less harm for less profit) relative to harmful (more harm for more profit) choices in the harm aversion task. B. Statistical model prediction for the probability to choose the helpful relative to the harmful option (y-axis) in function of the difference in money between the options (x-axis). The model predicts that participants will be less averse to harm as the difference in money of the harmful relative to the helpful option is higher, an effect that is augmented in *self* compared with *other* trials. C. In the prosocial effort task, participants were more reluctant to work at higher effort levels (x-axis), an effect that was augmented when the beneficiary was another unknown person relative to self. Y-axis depicts proportions of choosing the work offer (higher effort for higher reward) relative to a rest baseline option (no effort for small reward). D. Participants were more willing to work when the reward in offer was higher (x-axis), an effect that was amplified when the beneficiary was participants themselves compared with others. Y-axis depicts proportions of choosing the work offer relative to a rest baseline option. In all plots, dots depict each participant, and error bars represent SEM.

Afterwards, I tested whether these effects in trial-by-trial decisions were captured by the parameters obtained from the computational models for the harm aversion and the prosocial effort tasks. First, I compared how averse participants were to harming others relative to self. A Wilcoxon signed-rank test revealed that participants were significantly more averse to harm others (M = 0.38, SEM = 0.03) than themselves (M = 0.27, SEM = 0.03; $\kappa_{other} > \kappa_{self}$, z = -3.45, p < 0.001, Fig. 3.3A), with the majority of participants showing this trend (65.2%, Fig. 3.3A). This supported the hyperaltruism effect revealed by the MM. With regards to the prosocial effort task, a similar analysis revealed that participants discounted rewards by effort in a higher degree when the beneficiary was another person (M = 0.28, SEM = 0.04) than themselves (M = 0.06, SEM = 0.007, $\lambda_{other} > \lambda_{self}$, z = -6.94, p < 0.001; Fig. 3.3B). In contrast with the harm aversion task, the vast majority of the participants displayed a self-focused profile, with 95.5% of them being more averse to exerting effort for the other person (Fig. 3.3B). Therefore, these results revealed that the same individuals showed a very different behavioural pattern depending on the task-context, i.e. in harm aversion, people are hyperaltruistc, while when they need to be put in effort to work, they are prosocially apathetic.

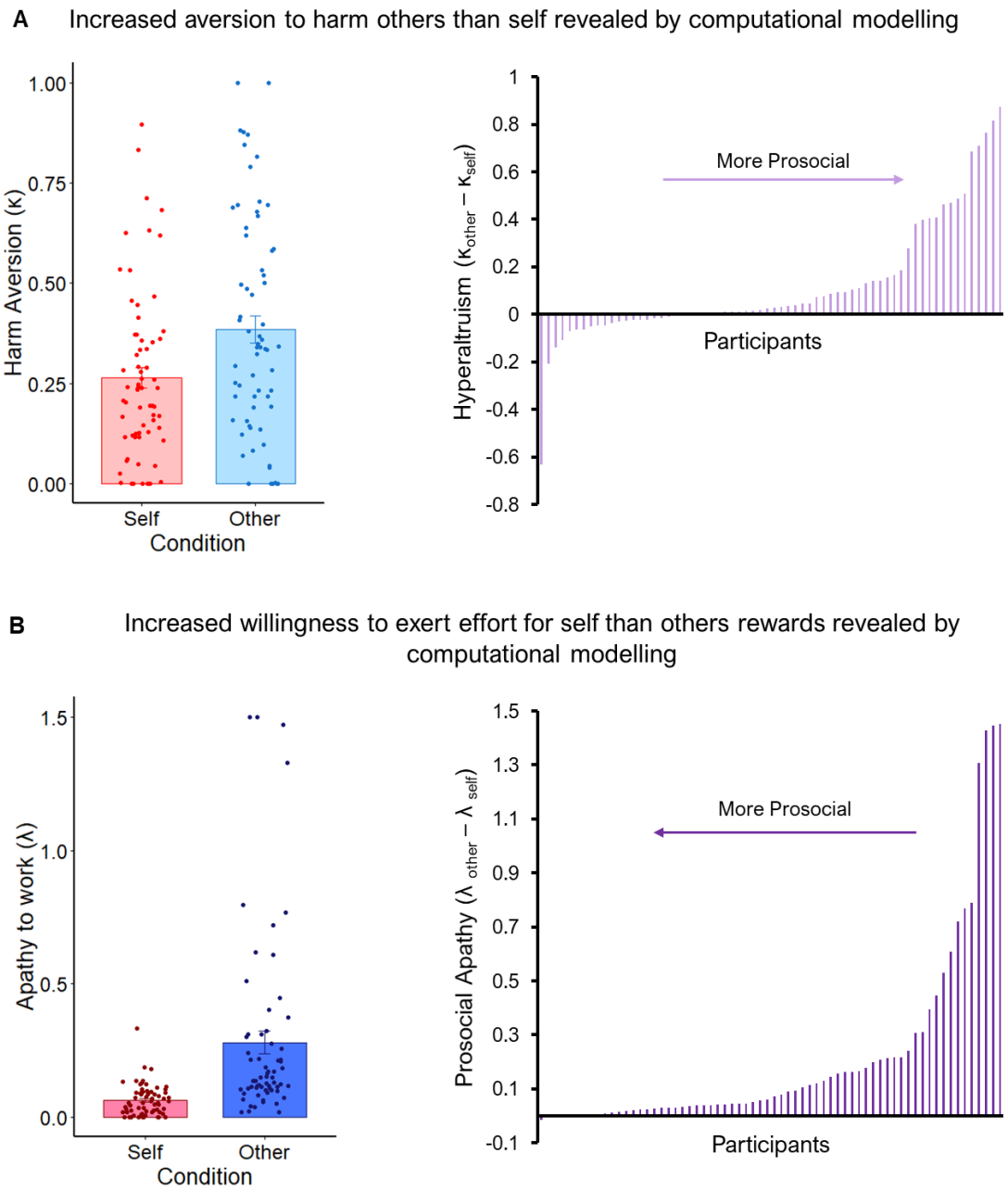


Figure 3.3. Hyperaltruism and Prosocial apathy effects are supported by computational modelling. A. *Left panel.* Participants are more averse to harm others compared with themselves as measured by computational parameters ($\kappa_{other} > \kappa_{self}$, $p < 0.001$). Y-axis depicts the value of the harm aversion parameter κ, which represents the degree in which profit is discounted by harm. *Right panel.* The hyperaltruism effect ($\kappa_{other} > \kappa_{self}$, y-axis) is seen in 62.5% of the current sample. The x-axis represents each individual ordered from lowest to highest values of hyperaltruism. B. *Left panel*. Participants are more apathetic to exert effort for others compared with themselves as measured by computational parameters ($\lambda_{other} > \lambda_{self}$, $p < 0.001$). Y-axis depicts the value of the discount parameter λ, which represents the degree in which reward is discounted by effort. *Right panel*. The prosocial apathy effect ($\lambda_{other} > \lambda_{self}$, y-axis) is seen in 95.5% of the current sample. The x-axis represents each individual ordered from lowest to highest values of prosocial apathy.

3.4.2 Prosocial behaviour is correlated across domains

Next, I tested whether prosocial behaviours in both tasks were correlated, such that people who were more prosocial in one task were also more prosocial in the other. In this analysis, I reverse-coded the prosocial apathy ($\lambda_{other} - \lambda_{self}$) effect and called this variable *prosocial effort*, such that positive values for the parameters in both the harm aversion and prosocial effort tasks meant higher prosocial preferences. Spearman-rank correlation analysis revealed that prosocial effort and hyperaltruism ($\kappa_{other} - \kappa_{self}$) were significantly correlated, with people who were more willing to exert effort for others being also more averse to harming others, in both cases relative to self ($\rho = 0.33$, $p < 0.008$, Fig. 3.4A). Next, I tested whether this correlation was triggered mainly by an effect of decisions in the *other* or in the *self* trials of each task. I correlated hyperaltruism with λ_{self} and λ_{other} on the one hand, and prosocial effort with κ_{self} and κ_{other} on the other. This analysis revealed that hyperaltruism was significantly correlated with λ_{other} (Spearman $\rho = -0.28$, $p < 0.03$) but not with λ_{self} (Spearman $\rho = 0.21$, $p > 0.09$). These correlations were significantly different from each other (Fisher-Z transformation, $z = -2.81$, $p < 0.005$). Likewise, prosocial effort showed a significant correlation with κ_{other} (Spearman $\rho = 0.28$, $p < 0.03$) but not with κ_{self} (Spearman $\rho = -0.02$, $p > 0.8$). Even though these latter correlations were not significantly different from each other, they followed the same trend as the above-mentioned hyperaltruism correlations (Fisher-Z transformation, $z = 1.73$, $p < 0.08$). Thus, prosocial behaviour was correlated between tasks mainly due to variability in participants' choices in the trials where the outcome was delivered to another person.

Thus, people who were prosocial in the harm aversion task tended to also be more likely to exert effort for others. But, is it the self cost, the other benefit, or a combination of both that trigger this shared variance across tasks? To answer this question, I built two MMs for *self* and *other* trials for each task. For the prosocial effort task, decisions to work were predicted by reward, effort and their interaction separately for *self* and *other*. On the

other hand, for the harm aversion task, decisions to help were predicted by differences in money, shocks and their interaction separately for *self* and *other*. I extracted random-slopes for reward and effort from the prosocial effort models, and money and shocks from the harm aversion models, to obtain each participant's sensitivity to costs and outcomes/benefits for each task, respectively. Next, I tested for correlations between tasks sensitivities in the *other* condition, in the *self* condition, and in their difference (i.e. *self* - *other*).

For the *other* condition, sensitivity to others' harm in the harm aversion task was significantly correlated with sensitivity to others' rewards in the prosocial effort task ($r = 0.3$, $p < 0.02$, Fig. 3.4B), such that as people were more averse to accept money for higher number of shocks, they were more willing to work as greater rewards were on offer. There were no other significant correlations. On the other hand, for the *self* condition, sensitivity to self harm was significantly associated with sensitivity to effort costs ($r = 0.31$, $p < 0.02$, Fig. 3.4C), such that people who were more averse to receive painful shocks for money were also more averse to exert effort for rewards. There were no other significant correlations found. Thus, while in a social context others' outcomes/benefits motivated behaviour across domains, self-benefitting decisions were driven mainly by aversion to negative-valence actions. Finally, differences between *other* and *self* betas for effort and reward respectively in the prosocial effort task, and for money and shocks respectively in the harm aversion task, were tested for correlations. Results revealed that differences in sensitivity to *other* relative to *self* profit in the harm aversion task was significantly correlated with differences in sensitivity to *self* and *other* rewards in the prosocial effort task ($r = 0.25$, $p < 0.05$, Fig. 3.4D). Thus, as people weighted more equally profit earned by harming others than themselves, they valued in higher degree their own than others' rewards when deciding to exert effort. No other significant correlations were found.

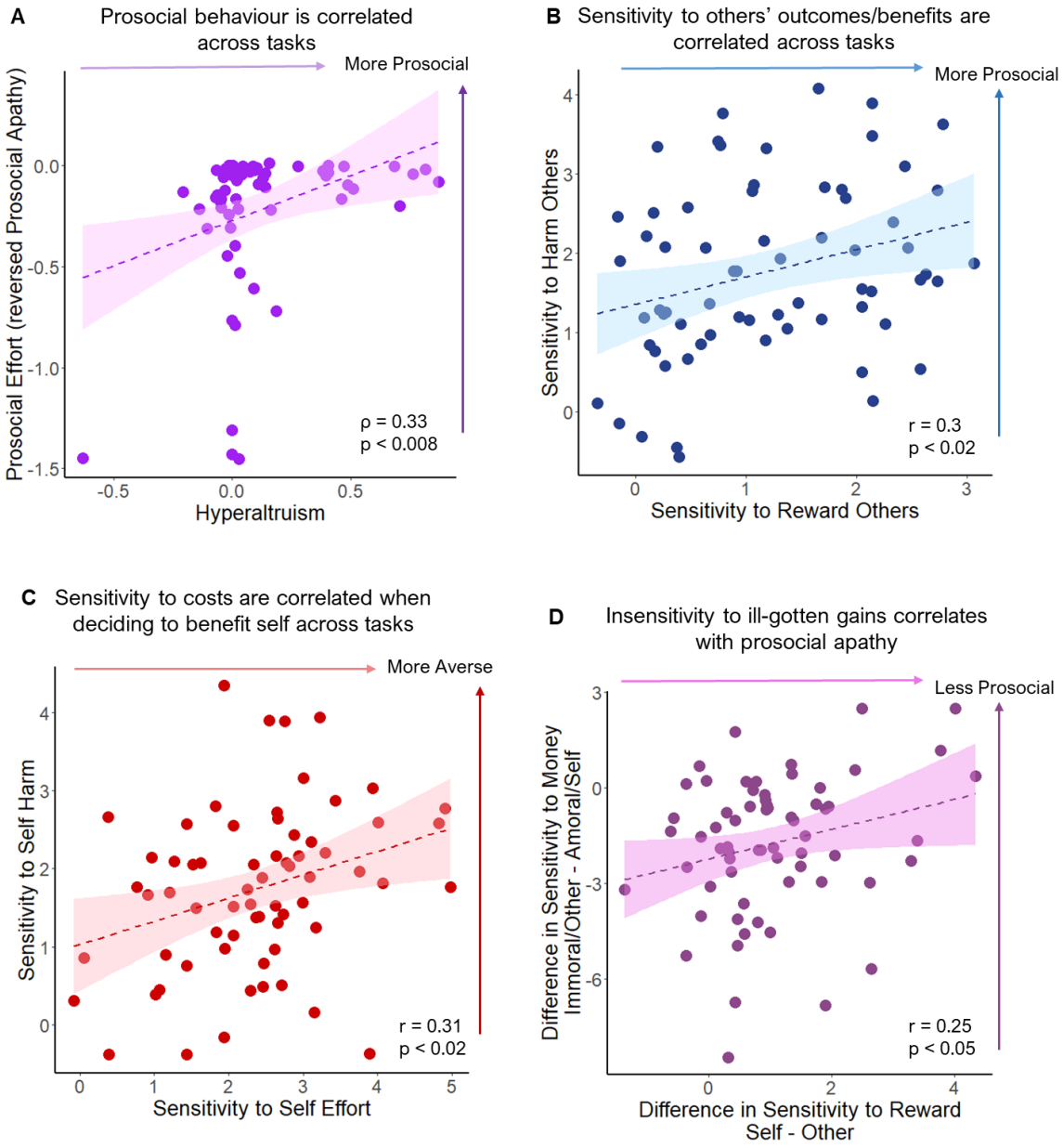


Figure 3.4. Prosocial behaviour in harm aversion and prosocial effort tasks share variance. A. Participants who were more averse to harm others than themselves in the harm aversion task (κ_{other} - κ_{self}, x-axis) were also more willing to work for others relative to self in the prosocial effort task (λ_{other} - λ_{self}, y-axis). X-axis and y-axis are the prosocial parameters extracted from computational models for each task. B. For the *other* condition in both tasks, sensitivity to others' benefits were correlated, such that people who were more sensitive to others' rewards in the prosocial effort task (x-axis) were more sensitive to others' harm in the harm aversion task (y-axis). 'Sensitivity' corresponds to the beta parameters extracted for each participant from mixed-effect models where prosocial decisions for each task were predicted by self cost, other benefit, and their interaction. C. For the *self* condition in both tasks, self costs were correlated, such that people who were more sensitive to effort (x-axis) were more sensitive to harm when benefitting self. 'Sensitivity' corresponds to each participant's beta parameter extracted from models, where self-benefitting decisions in each task were predicted by cost, and benefits, and their interaction. D. For differences between *self* and *other* conditions in both tasks, sensitivity to profit and rewards in both tasks were correlated, such that people who were more sensitive to self relative to other rewards in the prosocial effort task (x-axis), were more similarly sensitive to money earned by harming others than self in the harm aversion task (y-axis). X-axis depicts self betas minus other betas for reward in prosocial effort, while y-axis corresponds other betas minus self betas for money in harm aversion.

3.4.3 fMRI results

What mechanisms in the brain lead to these two types of prosocial behaviours being related to each other? I disentangled signals to harm and profit for *self* and *other* in the harm aversion task, testing for correlations with prosocial effort actions. Thus, I examined whether prosocial effort could be explained by neural responses to anticipated others' harm or by neural sensitivity to profit earned from harming others. I used a GLM in five specific regions in the ACC, AI and TPJ to identify how BOLD signals varied parametrically to the objective amounts of money and shocks (Δm and Δs respectively) that would result if the harmful relative to the helpful option is chosen, in *self* and *other* trials separately. One hypothesis predicts that activity to others' harm will be related to sensitivity to reward in the effort task – *other-regarding hypothesis*, which was observed in people's behaviour (see above). A second hypothesis predicts that, relative to self, neural responses to profit earned by harming others will be associated with sensitivity to reward in the prosocial effort task – *moral-conflict hypothesis.* A third hypothesis predicts that signal to others' harm will correlate with sensitivity to effortful actions when working for others – *valence-related hypothesis*. Finally, a fourth hypothesis predicts that signal to profit by harming others will correlate with effort sensitivity when benefiting others – *self-centred hypothesis.* Statistical maps signalling profit and shocks in these regions were then extracted and correlated with behaviour in the prosocial effort task. The behavioural indexes used for prosocial effort were the beta values for effort and reward extracted separately for *self* and *other*, and their difference (*self - other*).

In line with the other-regarding hypothesis, neural responses to anticipated others' harm (Δs_{other}) in vAI ($r = -0.42$, $p < 0.003$ FDR corrected), ACCg ($r = -0.3$, $p < 0.03$ FDR corrected) and TPJp ($r = -0.3$, $p < 0.03$ FDR corrected) were correlated with sensitivity to others' relative to self rewards in the prosocial effort task (Fig. 3.5). Therefore, as these regions showed more positive parametric responses to others' harm, participants were

more incentivised by others' rewards compared to self rewards, indicating reduced prosocial apathy. This correlation was significantly different from the correlation between self harm (Δs_{self}) signal and sensitivity to other vs self rewards in vAI (Fisher z-transformation, $z = 2.75$, $p < 0.007$) and ACCg ($z = 2.78$, $p < 0.006$), but not in TPJp ($z = 1.44$, $p > 0.1$). In fact, there were not significant correlations between Δs_{self} signal and prosocial effort indicators. Furthermore, I examined whether the effects of the statistical maps for others' harm were specific for sensitivity to *other* vs *self* rewards, testing whether this correlation was significantly different from sensitivity to effort in *other* relative to *self* trials. Indeed, these correlations were significantly different from each other in vAI ($z = 3.31$, $p < 0.001$) and ACCg ($z = 2.49$, $p < 0.02$), but not in TPJp, although it had the same trend as the other areas ($z = 1.82$, $p < 0.07$).

Finally, even though individual differences are the focus of this study, I performed a group-level analysis testing whether responses to others' harm in vAI, ACCg and TPJp were significantly different from zero. These analyses revealed that ACCg ($t = -2.2$, $p < 0.04$, $M = -0.05$, $SEM = 0.02$), and TPJp ($t = -2.7$, $p < 0.009$, $M = -0.09$, $SEM = 0.03$) responses to others' harm were significantly different from zero at the group level, but not in the vAI. For completeness, I also performed a whole brain analysis to identify those clusters that significantly signalled harm and profit for *self* and *other*. Tables showing these uncorrected results ($p < 0.001$, $k > 30$) can be found in Appendix A.

Taken together, neural responses to others' harm in ACCg, vAI and TPJp, but not in dAI nor TPJa, are associated with sensitivity to others' vs self rewards in prosocial effort, suggesting that individual differences in how these areas respond to others' outcomes can predict prosocial behaviour in another prosocial domain. Importantly, these results also suggest differences between subregions in their roles in prosocial behaviour.

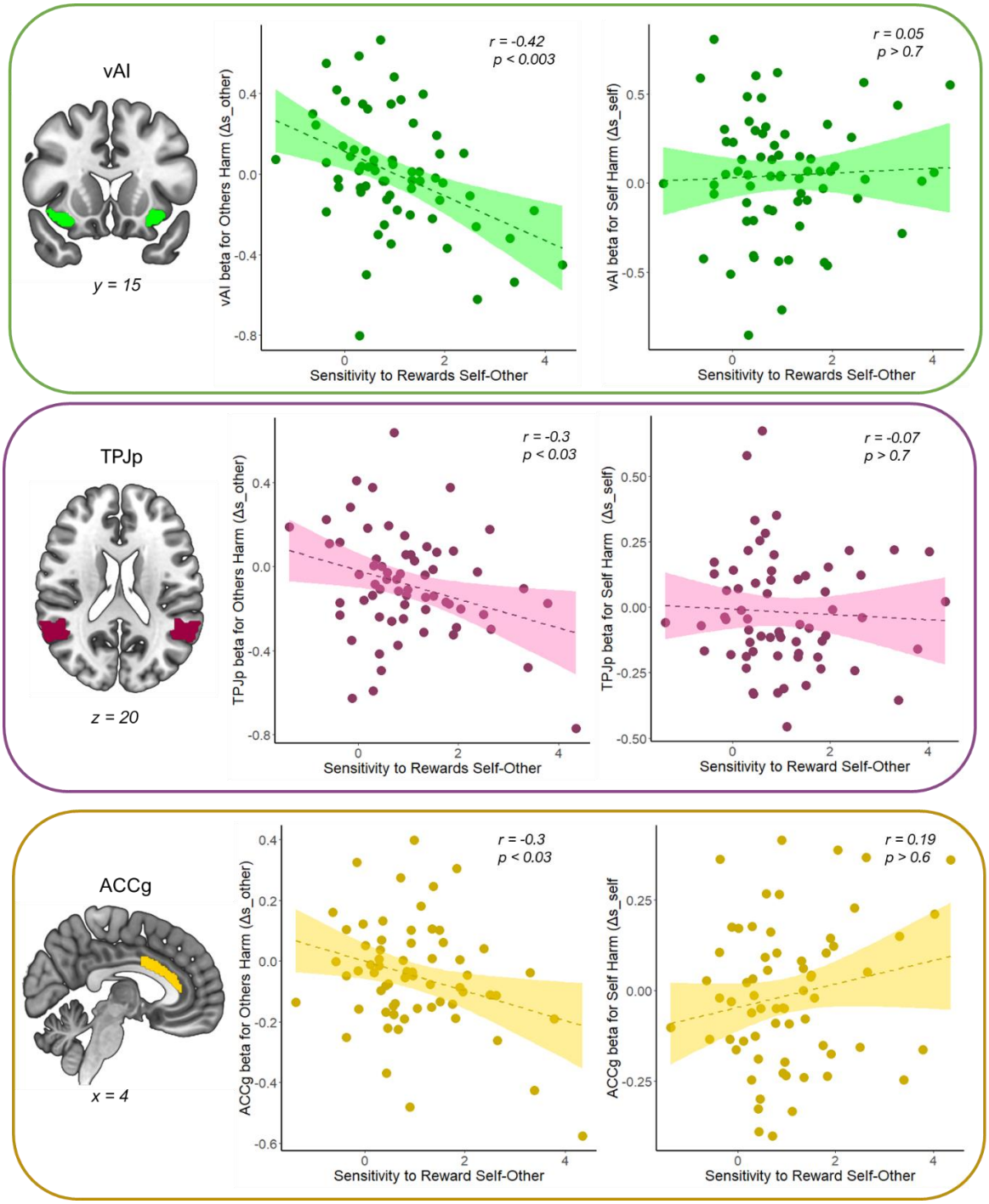


Figure 3.5. Correlations between parametric maps for anticipated harm and sensitivity to rewards for others relative to self in the prosocial effort task. The ventral portion of anterior insula (vAI, top panel), the posterior portion of temporo-parietal junction (TPJp, middle panel), and the gyrus portion of anterior cingulate cortex (ACCg, bottom panel) responses to others' harm (plots in the left) were correlated with sensitivity to self vs other rewards in the prosocial effort task. Lower beta values from statistical maps in these areas (y-axis) were associated with higher differences in behavioural reward betas for self vs others predicting prosocial decisions (x-axis). In contrast responses to self harm in these areas (plots in the right) were not correlated with differences in reward sensitivity between self and other in the prosocial effort task.

Next, I tested whether responses to anticipated harm and profit for *others* versus *self* trials ($\Delta s_{other} > \Delta s_{self}$ and $\Delta m_{other} > \Delta m_{self}$ respectively) were correlated with prosocial effort indexes. No significant correlations survived multiple comparison for $\Delta s_{other} > \Delta s_{self}$ contrast. However, there was a significant correlation between $\Delta m_{other} > \Delta m_{self}$ signal in the TPJa ($r = 0.44$, $p < 0.002$ FDR corrected, Fig. 3.6). Stronger responses to profit obtained by harming others versus self were associated with a greater incentivisation for obtaining self rewards rather than others' rewards in the prosocial effort task. This effect was specific for the sensitivity to rewards, as effort sensitivity for self vs others did not correlate with $\Delta m_{other} > \Delta m_{self}$, and this correlation was significantly different from the one with sensitivity to rewards ($z = 2.61$, $p < 0.01$). At the group level, $\Delta m_{other} > \Delta m_{self}$ responses in TPJa was not significantly different from zero. Thus, the signal to profit in TPJa was significant only in relationship with sensitivity to rewards in prosocial effort. Thus, people who showed higher differences in TPJa responses to money earned by immoral compared with amoral means, were more incentivised to work for others' relative to self rewards, consistent with a moral-conflict role of this area

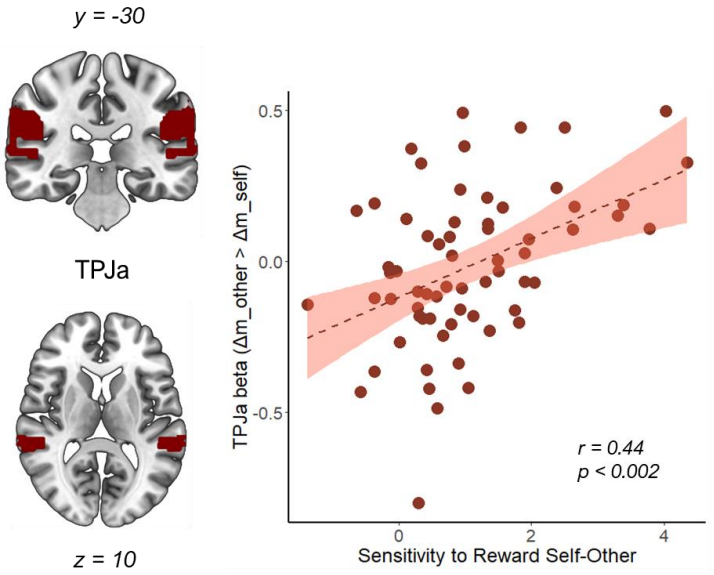


Figure 3.6. Anterior temporo-parietal junction (TPJa) responses to profit obtained by harming others vs self ($\Delta m_{other} > \Delta m_{self}$) is associated with prosocial effort. *Left.* TPJa mask used to extract statistical maps for $\Delta m_{other} > \Delta m_{self}$ contrast. *Right*. Significant correlation between betas extracted from $\Delta m_{other} > \Delta m_{self}$ in TPJa (y-axis) and reward sensitivity (behavioural betas, x-axis) for others vs self in the prosocial effort task. People with higher responses to profit obtained by harming others relative to self valued more their own rewards compared with others', being less willing to exert effort to help others

3.5 Discussion

Prosocial behaviour is manifested in different forms, with people typically being more willing to incur personal costs when avoiding harming others, than when effortfully getting rewards for others (Crockett et al., 2014, 2015, 2017; Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021). Could individual differences in these prosocial manifestations reveal some common mechanisms underlying prosocial behaviour across domains? Here, I demonstrated that neural and behavioural sensitivity to others' harm was associated with being prosocially incentivised by others' rewards, suggesting that representing others' outcomes/benefits, regardless of their positive or negative valences, underlie this association. Thus, other-regarding mechanisms might drive prosociality across domains. Furthermore, neural responses to profit in TPJa obtained by harming others vs self was also associated with prosocial effort. This indicates that sensitivity to moral conflict, encoded by TPJa, can predict prosocial behaviour in a different context. Thus, variability in TPJ, ACC and AI signals during harm aversion decisions can predict, one week later, how willing a person is to overcome effort to get rewards for others, suggesting that common mechanisms exist across different domains of prosociality.

Higher behavioural sensitivity and activity to others' harm in vAI, ACCg and TPJp during harm aversion decisions were associated with being more incentivised by others rewards in working prosocially. These findings support the other-regarding hypothesis as a common mechanism underlying different prosocial behaviours. ACCg, vAI, and TPJp signals might encode the benefit and not the valence of the outcome to be received by others. Harm is a negative outcome, and prosocial actions aim to reduce it when suffered by others. On the other hand, reward is a positive outcome, and prosocial behaviours aim to increase it in others. Thus, these areas seem to represent the beneficial value of the outcomes irrespective of their modality, guiding behaviours to help (Apps et al., 2016; Corradi-Dell'Acqua et al., 2016; Fan et al., 2011; Hare et al., 2010; Hein et al., 2010; Hutcherson et al., 2015; Lamm et al., 2011; Masten et al., 2011; Saxe, 2010).

Furthermore, these results highlight differences between different subregions of AI and TPJ. While the other-regarding relationship between harm aversion and prosocial effort was found in the vAI and TPJp, it was not present in dAI and TPJa signals. Typically, the latter two subregions have been involved in attentional, working memory, and cognitive processes, while vAI and TPJp have been proposed as their socio-affective counterparts, being identified as crucial for empathic representations of others' emotions (Chang et al., 2013; Corbetta, Patel, & Shulman, 2008; Deen et al., 2011; Mars et al., 2012; Nomi et al., 2018; Patel, Sestieri, & Corbetta, 2019; Patel et al., 2015; Uddin et al., 2014, 2017). Likewise, variability of vAI to others' outcomes can elicit prosocial behaviour mediated by empathy, while dAI seems to be anchored to self processes (Hu et al., 2021). Dorsal AI is a key region for awareness and body perception, and it might not be recruited for evaluating social outcomes in relation to prosocial behaviour (Craig, 2009; Deen et al., 2011; Nomi et al., 2018). Thus, vAI and TPJp, together with ACCg, which seems to be the specific portion of the ACC processing relevant motivational stimuli in social context (Apps & Ramnani, 2014; Apps et al., 2016; Behrens et al., 2008; Chang et al., 2013; Hernandez-Lallement et al., 2020; Lockwood, 2016), constitute a hub specialised in representing others' affective states and evaluating others' anticipated outcomes, which guide decision-making and motivation.

Stressing the differences between subregions in prosocial behaviour, TPJa appears as having a moral conflict role in predicting prosociality. In contrast with the other-regarding involvement of the vAI, ACCg and TPJp in prosocial motivation, TPJa might be signalling the tension between self and other interests. People whose TPJa activity did not distinguish between obtaining profit by harming others (immoral action) from harming oneself (amoral action), were willing to work in a higher degree for their own rewards compared with others'. This finding shed light on the debate regarding TPJ function in social behaviour. On the one hand, TPJ has been proposed as underlying other-regarding mechanisms through representation of others' affect, as previous studies have

shown that higher activity in this area predicts prosocial decisions, such as monetary donations (Hare et al., 2010; Hutcherson et al., 2015; Park et al., 2017; Strombach et al., 2015). According to the current findings, this role lies on TPJp. On the other hand, TPJ has been also suggested as signalling conflict in prosocial decisions, as enhanced activity has been observed when people face more difficult prosocial choices (FeldmanHall et al., 2014; Morishima et al., 2012), and its function is causally necessary to signal moral conflict between personal profit and moral costs (Obeso et al., 2018). According to the current results, this latter role seems to be more predominantly associated with TPJa. These findings reveal that individual differences in the activity of these two areas are associated with two different aspects of prosocial behaviour, which are consistent with their distinction as socio-affective (TPJp) and cognitive-attentional (TPJa) functional portions of TPJ (Mars et al., 2012).

Even though prosocial effort and harm aversion were correlated, behavioural results at the group level suggest that the motivation underlying behaviour in each task is different. In harm aversion, people were more averse to harm others than themselves for profit, and this effect was triggered by sensitivity to profit – people were more sensitive to money when harming oneself than others. This is consistent with previous neural results suggesting that profit earned by immoral means is devalued (Crockett et al., 2017). Thus, in harm aversion, the focus is on the costly action. On the other hand, in the prosocial effort task people were more willing to work for self than other rewards, effect triggered by sensitivity to effort and rewards. As prosocial effort does not have proscriptive morality associated, people must be sufficiently incentivised by others' reward to overcome the effort. Thus, despite differences at the group level between domains, at the individual level neural and behavioural sensitivity to others' harm predicted being more incentivised by others' rewards in the prosocial effort task. These results suggest that variability in responses to others' benefits is powerful in predicting prosocial behaviour – even though TPJ, AI, and ACC other-regarding signals do not seem to influence harm aversion

behaviour, its variability predicts prosocial behaviour *off-line*, in a task where one of the focuses is on others' benefits. Thus, sensitivity to others' benefits might be a relatively stable trait, guiding behaviour when the context allows it. From this perspective, empathic traits, which modulates sensitivity to others' outcomes/benefits (FeldmanHall et al., 2015; Hein et al., 2010; Hu et al., 2021; Lockwood et al., 2015, 2016), could eventually predict both prosocial effort and harm aversion. This hypothesis will be tested in the next chapter.

One limitation of the current study is that the established relationship between harm aversion and prosocial effort is rather indirect. Future studies should illuminate neural *online* signal during prosocial effort decisions, testing whether sensitivity to others' rewards in these areas drive prosocial behaviour, as the current results suggest. It might be also interesting to examine how sensitivity to others' benefit and self costs interact when both have negative valences, e.g. exerting effort to prevent harm or vice versa. It is unclear when the valence is changed in such a manner whether people will conform to the moral norm of preventing others harm or will continue to prioritise the self effort costs. Furthermore, future research could elucidate whether this shared variance between prosocial domains could be extended to prosocial actions other than harm aversion and prosocial effort (e.g. charity donations), expanding our knowledge in the common mechanisms underpinning prosociality.

In this chapter, I found that individual differences in neural responses to others' harm in TPJp, vAI and ACCg were associated with sensitivity to others' rewards in effort-based prosocial decisions, indicating that these regions might be a hub which represents others' benefits and facilitates prosocial behaviour across contexts. Furthermore, signal for profit obtained by immoral vs amoral acts in TPJa was also correlated with prosocial effort, implying a self-cost/other-benefit conflict role of this area, consistent with previous studies. Thus, these results suggest shared mechanisms between two domains of prosocial behaviour.

Chapter 4

Transdiagnostic markers of affective sensitivity are associated with prosocial behaviour in different domains

Publication: Contreras-Huerta, L. S., et al. (2020). *Emotion*.

4.1 Summary

Psychiatric disorders have been suggested to be related to prosocial disturbances, which may underlie many of their social impairments. However, broader affective traits, present in different degrees in both psychiatric and healthy populations, have also been linked to variability in prosociality. Therefore, it is unclear to what extent prosocial variability is explained by specific psychiatric disorders relative to broad affective traits. Using a computational, transdiagnostic approach in two online studies, I found that participants who reported being more affectively reactive across a broad cluster of traits, including empathy and apathy, manifested greater frequencies of prosocial actions in two different contexts: they reported being more averse to harming others for profit, and they were more willing to exert effort to benefit others. These findings help illuminate the profile of prosociality across psychiatric conditions as well as the architecture of prosocial behaviour in healthy individuals.

4.2 Introduction

Prosocial behaviours lie at the core of healthy social relationships, facilitating cooperation within communities (de Waal, 2008; Fehr & Camerer, 2007; Fehr & Fischbacher, 2003). Nevertheless, the frequencies of these behaviours and the extent to which they are manifested can vary from person to person, and in particular seem to be disturbed across a variety of psychiatric disorders (Bartz & Hollander, 2006; Bora, Yucel, & Allen, 2009; Robson et al., 2019). A hallmark of several psychiatric disorders is impairments in affective processing and interpersonal functioning. For instance, criteria for diagnosing depressive, anxiety and personality disorder (PD) include affective disturbances such as sadness/irritability, anxiety/fear, and impairment in a range of emotional responses that have a significant impact on social functioning (American Psychiatric Association, 2013). Thus, changes in prosocial behaviours may be one manifestation of a general impairment in interpersonal functioning in these disorders. In fact, previous studies have associated depression, anxiety and PDs with poor capability in creating and maintaining social relationships, reduced cooperation and reciprocity, and violence towards others and oneself (Blair, 2007; Glenn, Raine, & Schug, 2009; Howard, 2015; King-Casas et al., 2008; Koenigs, Kruepke, & Newman, 2010; Mier et al., 2013; Rilling et al., 2007; Robson et al., 2019; Unoka et al., 2009; Yu, Geddes, & Fazel, 2012). Therefore, affective disturbances in psychiatric disorders impact negatively patients' social lives, with low prosocial disposition putatively being one of those manifestations.

At the same time, broad affective traits that are not associated with specific psychiatric categories have also been related to prosocial behaviour in healthy individuals. For example, people who are less empathetic towards others are less inclined to behave prosocially in a variety of different scenarios (Batson, 2008; Batson, Eklund, Chermok, Hoyt, & Ortiz, 2007; Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016; FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; Hein, Silani, Preuschoff, Batson, & Singer, 2010; Lockwood, 2016; Morelli, Rameson, & Lieberman, 2014). In addition,

apathy and alexithymia have been associated with reduced motivation to help others (FeldmanHall, Dalgleish, & Mobbs, 2013; Lockwood, et al., 2013, 2017; Patil & Silani, 2014a, 2014b). Importantly, even though these traits have been linked to prosocial behaviour in healthy participants, they are prevalent across multiple psychiatric disorders. In particular, trait-level apathy and alexithymia have been linked to depressive, anxiety, and PDs among other psychiatric conditions (Ang et al., 2018; Chow, 2000; De Berardis et al., 2017; Hendryx et al., 1991; Husain & Roiser, 2018; Lander, Lutz-Zois, Rye, & Goodnight, 2012; Le Heron et al., 2018; Marin, 1990; New et al., 2012; Starkstein & Leentjens, 2008; Strauss & Cohen, 2017; Taylor & Bagby, 2004; Thobois et al., 2017; Van Reekum, Stuss, & Ostrander, 2005), while reduced empathy has been associated with PD, apathy and alexithymia (American Psychiatric Association, 2013; Bird & Viding, 2014; Herpertz & Bertsch, 2014; Lander et al., 2012; Lockwood, Ang, et al., 2017; Lockwood et al., 2013; New et al., 2012; Valdespino, Antezana, Ghane, & Richey, 2017). Therefore, these broader, affective traits may not only modulate social behaviour in healthy participants, but also play a role in the social dysfunctions and affective impairments across multiple psychiatric disorders and symptoms.

This situation relates to the more general problem of diagnostic categorisation in psychiatry, where ‘normal’ is separated arbitrarily from ‘abnormal’, without considering disorders as extreme cases of a mental health continuum (Braun, 2018; McGorry & Nelson, 2016; Newton-Howes & Clark, 2015). In order to address this challenge, the Research Domain Criteria (RDoC) initiative was developed to identify biologically plausible, transdiagnostic signatures of different psychiatric disorders (Insel et al., 2010). Although aspects of social functioning are included in RDoC, prosocial behaviour has largely been overlooked, which is surprising given its relevance for healthy social functioning. Therefore, the precise relationship between prosocial deficits and psychopathology has yet to be established, in particular the specificity of such deficits to distinct psychiatric syndromes versus broader affective traits present across healthy and

psychiatric populations. A transdiagnostic approach may help identify specific phenotypes within a dimensional affective and psychiatric domain that underlie deficits in prosociality, addressing the high comorbidity that exists across different disorders, and the high multicollinearity between different traits.

In this chapter, I adopt a dimensional approach to investigate the relationship between individual differences in prosocial behaviour and both specific psychiatric traits and broader affective traits. I aim to elucidate, using Canonical Correlation Analysis (CCA), whether specific psychiatric categories can explain prosocial behaviour even in a spectrum where broader affective traits are considered. CCA determines the combination of variables (i.e. psychiatric/affective traits) that are predictive of high levels of another combination of variables (e.g. different types of prosocial behaviour; Sherry & Henson, 2005; Wang et al., 2020). Importantly, CCA can also identify which variables are key predictors, which are collinear with the other variables, and which may be predictive only because they supress other variables (Nimon, Henson, & Gates, 2010). Here, I use CCA to identify which traits are more strongly linked to prosocial behaviours when multicollinearity and suppression effects are controlled for.

Importantly, most existing studies have measured unidimensional aspects of prosociality in relationship with traits and psychiatric disorders, typically related to monetary allocations to strangers or charities (Declerck, Boone, & Emonds, 2013; Koenigs, Kruepke, Zeier, & Newman, 2012; Rilling, King-Casas, & Sanfey, 2008; Robson et al., 2019). However, different dimensions of prosocial behaviour can be associated with different decisional patterns, such as hyperaltruism in a harm aversion context (Crockett et al., 2014, 2015, 2017) or prosocial apathy in an effort-based scenario (Lockwood, Hamonet et al., 2017; Lockwood et al., 2021). Crucially, in the previous chapter I found that prosocial behaviours in these two contexts were correlated, suggesting common prosocial mechanisms. Here, I aimed to replicate this effect, using

this common variance to understand how psychiatric/affective traits relate to prosocial behaviour.

To address these questions, I adapted for an online setting the harm aversion and the prosocial effort tasks described in Chapter 2. Using computational modelling of choices in both tasks, I measured participants' preferences for costly actions that benefited either others or themselves. The aim of study 1 (n = 113) was to examine whether the adapted versions of these tasks could reproduce the same behaviours seen in the lab. In study 2 (n = 212), I aimed to replicate study 1 within a larger sample. Finally, I collapsed both samples to maximise power to investigate, using CCA, which constellation of traits were predictive of prosocial behaviour. I focused on specific psychiatric traits, including PD (general, psychopathy and borderline, the latter two being the subtypes that have been more robustly linked to impaired social cognition and behaviour), depression and anxiety; and broad affective traits, including different dimensions of apathy, alexithymia, and empathy.

4.3 Methods

4.3.1 Participants

Data were collected online using Amazon's Mechanical Turk (AMT), and all procedures were administered through Qualtrics.com©. Participants received a base rate payment of $7.85, plus a bonus based on their performance on the prosocial effort task (M = $0.9, SD = $0.13). All subjects were from the USA and were over 18 years old. Exclusion criteria are detailed below. Participants provided their consent online and their identities were unknown.

For study 1, the sample size was based on previous studies that have revealed hyperaltruism and prosocial apathy effects (Crockett et al., 2014, 2015, 2017; Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021), aiming to replicate these lab-based effects

in adapted version of the tasks collected in online samples. This number was further increased to accommodate the potential noise introduced by adapting these lab-based tasks into online platforms. Thus, in study 1, 113 participants successfully completed the whole protocol (50 females, Age M = 35.7, SD = 11.6).

Study 2 aimed (i) to double the sample size of study 1 in order to replicate its results and have more statistical power to correlate behaviours in the two tasks, and (ii) have enough power to perform the CCA once both studies samples are collapsed, in line with recommended sample sizes (Hair, Black, Babin, & Anderson, 2010; Leach & Henson, 2014). Thus, in study 2, 212 participants successfully completed the whole protocol (101 females, Age M = 37.7, SD = 10.6). Then, the collapsed sample was of 325 participants included in the CCA.

4.3.2 Procedure

Participants were taken to a link on the Qualtrics website which contained versions of both tasks and self-report measures. Participants performed the self-report measures first, followed by either the prosocial effort task or the harm aversion task. For study 1, participants always completed the tasks in the same order: harm aversion task first, and prosocial effort task second. For study 2, the orders were counterbalanced: harm aversion first (n = 111), and prosocial effort first (n = 101). Within the self-report section, the different scales were administered in a fully randomised order. Each task followed the same sequence: detailed instructions about the task, followed by practice trials, comprehension questions (see below) and then the main blocks of the experiment.

4.3.3 Exclusion criteria and quality assurance

First, I identified potential MTurk bots and survey-farmers assigning a score from 0 to 8 to each participant based on their location and qualitative responses, where higher

scores mean more likelihood of poor data quality (Prims & Motyl, 2018). Qualitative responses of participants with a score higher than 0 were closely examined to check quality. All participants showed low scores (i.e. ≤ 3) and their responses did not suggest presence of bots or survey-farmers.

To ensure the quality of data collected from the online platform, I followed recommendations from previous studies using AMT and established exclusion criteria to improve the quality of the data. As such, participants had to successfully answer comprehension questions about the instructions, attention-check trials, and a cut-off for missed trials (Crump, McDonnell, & Gureckis, 2013). Thus, after a practice set, participants had to answer three comprehension questions about the instructions of the tasks. If the participant answered any of these questions incorrectly, they received an additional summary of the instructions. Then they were required to answer the same comprehension questions, which if any were failed again the study was terminated. I also applied additional exclusion criteria based on task performance. For the harm aversion task, four attention-check trials were included. In these trials, one of the two options was obviously more attractive compared to the other (higher reward in exchange of less amount of electric shocks). If any of these four attention-check trials was responded to incorrectly, participants were excluded from the final analyses (study 1, n = 2; study 2, n = 19). For the prosocial effort task participants were excluded if they missed (i.e. decision not made within six seconds time window) more than 10% of the trials (study 1, n = 4; study 2, n = 10). Both tasks had different exclusion criteria due to their characteristics. Thus, I did not include attention-check trials in the prosocial effort task as each decision in this task was followed by a direct action (e.g. work or rest) associated to real outcomes (money). On the other hand, I did not include missed-trial criteria in the harm aversion task because, unlike the prosocial effort task, responses on the task were not time-restricted. Finally, participants were only allowed to participate if they were using a laptop with either mouse or trackpad, or a desktop computer.

4.3.4 Behavioural Tasks

I created an online version of the harm aversion task, where participants had to hypothetically imagine that money was being traded-off against painful shocks in a laboratory setting, but neither the money nor shocks were actually delivered. For the prosocial effort task, I created a version in which effort was operationalised as ticking boxes on the screen in a fixed time window. Different from their laboratory counterparts, the social aspects of both online tasks, as well as the bonuses that participants could earn in the prosocial effort task, were fully transparent.

4.3.4.1 Hypothetical harm aversion task

Participants were instructed to imagine a hypothetical experiment in which they were randomly assigned to the role of Decider and another unknown person was assigned the role of Receiver. As the Decider, they would trade-off money against electric shocks that could be delivered either to themselves (*self* trials) or to the Receiver (*other* trials). Money always was received by the Decider. Participants were told to imagine that the electric shocks were comparable to a painful shot at the doctor's office.

On each trial, participants chose between two options: a harmful option of more shocks and more money, and a helpful option that resulted in fewer shocks and less money. The recipient of the shock was listed at the top (red for *self* and blue for *other* trials; Fig. 4.1A). The harmful and helpful options were randomly allocated to the left or the right, to discourage habitual responding. Participants were instructed to make their choices by clicking their mouse and to respond within a maximum of 10 seconds on each trial. After 10 seconds remaining on the choice trial, participants were prompted to give an answer. To preserve choice independence, participants were told to imagine that one trial would be randomly selected and actually implemented. However, participants were

explicitly told that, in reality, nobody would receive any shock and their participation payment would not be affected by their choices.

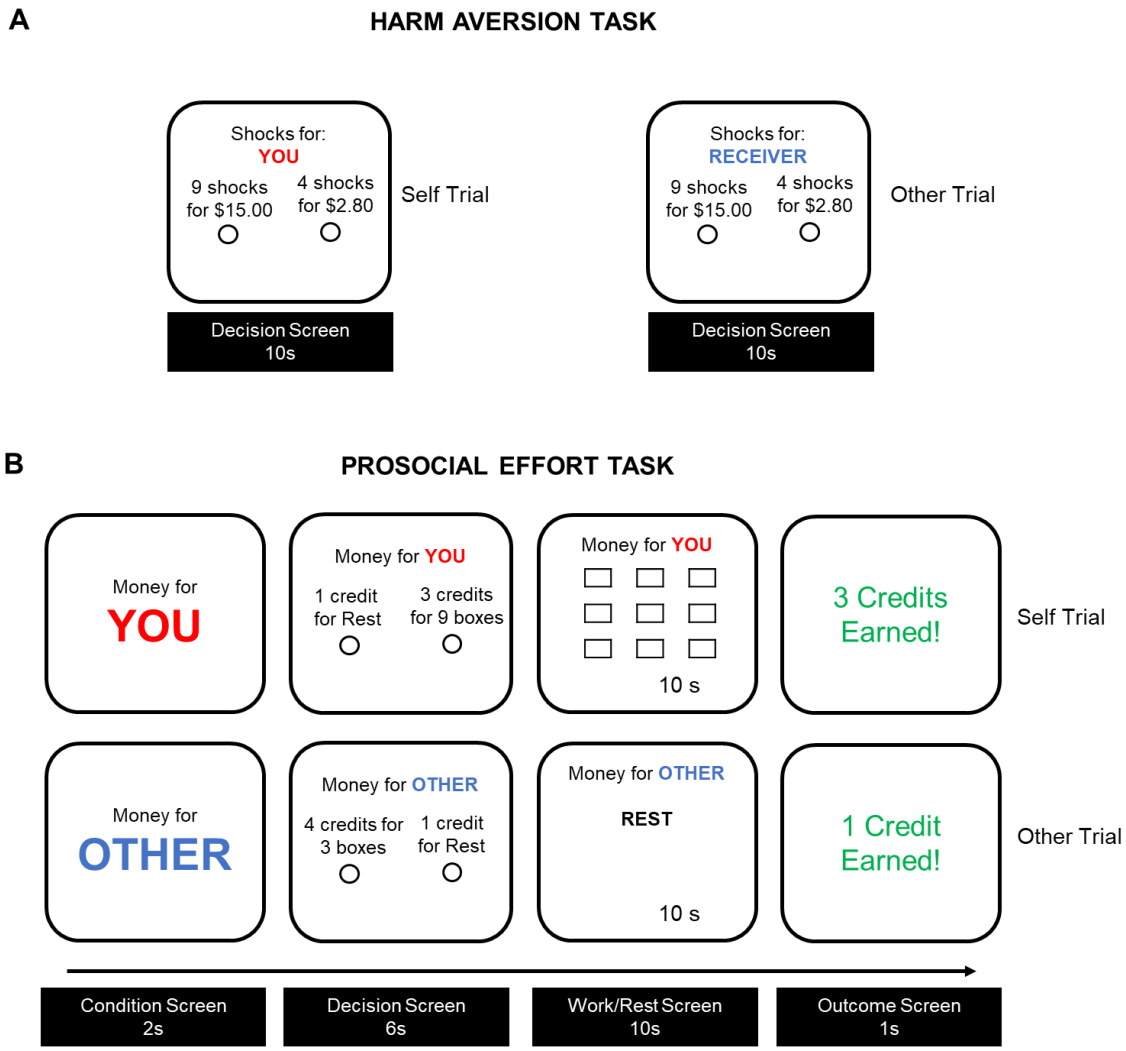


Figure 4.1. Behavioural online tasks testing two aspects of prosocial behaviour. A. *Example of a self (left panel) and other (right panel) trials in the hypothetical harm aversion task*. Participants had to decide between high and low profit in exchange for a high or low number of electric shocks that could be delivered to either themselves or to an unknown person. The task was hypothetical such that no money or electric shocks were actually delivered. B. *Samples of a self (top panel) and other (low panel) trials of the prosocial effort task*. Participants made choices between a baseline rest and low reward option, and a work offer of a variable higher effort (30, 50, 70 and 90% of the boxes ticked in a calibration phase) and higher reward (2-4 credits). If the work offer was chosen, participants had to tick in 10s the specific number of boxes required to obtain the reward on offer (top panel), otherwise 0 credits were delivered. If the baseline was chosen, participants rested for 10s and received 1 credit (low panel). On half of the trials, credits were allocated to participants themselves (*self* trials), and in the other half they were allocated to the other person (*other* trials).

Participants were randomly assigned to one of three trial sets following the procedure described in Chapter 2. These three different trial sets help to ensure that results are not a product of the choice set. Each trial set was optimised to give the most efficient estimates of potential participants' harm aversion parameters. Participants completed 35 trials in both the self and other conditions, with 70 trials in total. Four attention-check trials were included (as exclusion criteria and not included in the final analysis, see above), two for self and two for the other conditions.

4.3.4.2 Prosocial effort task

In this task, participants made decisions about their willingness to exert effort to obtain real monetary rewards for themselves or for another unknown person. On each trial participants were presented with two options: a 'work' offer, associated to different levels of effort (30, 50, 70 and 90% of the maximum number of boxes that a participant could ticked in 10 seconds) and rewards (2-4 credits), and a 'rest' option related to no effort and only one credit. All trials were the same duration regardless of what choice was made, ensuring that choices were not influenced by temporal discounting (Critchfield & Kollins, 2001). On half of the 48 trials participants made choices and exerted effort to benefit themselves (*self* trials), while on the other half they exerted it to benefit another unknown person (*other* trials). Participants were explicitly told that money on *other* trials would be delivered to a participant in another study, that their identity would remain confidential and that the other participants did not know that they were receiving a bonus earned by another person's effort.

Each trial started with a screen indicating the condition (red for *self* and blue for *other* trials) for two seconds, followed by a decision screen where participants had to make their choice (Fig. 4.1B). Participants had six seconds to make a decision. If the participants chose the 'work' offer, a screen with the number of boxes to tick appeared. In order to earn the corresponding number of credits, the participant had to tick all boxes

within 10 seconds, from top to bottom, left to right. If all the boxes were not ticked, then 0 credits would be earned. If the rest option was chosen, participants saw a screen with the word “Rest” at the top, and they were not required to execute any task for 10 seconds. A timer starting from 10 at the bottom of the screen indicated the time on all trials. Once 10 seconds had passed, a screen indicating the number of credits earned appeared for one second. To incentivise participants to pay attention to the offers in the task and make active decisions on each trial, two measures were undertaken. Firstly, participants had six seconds to make their initial decision to work or rest, otherwise they would receive zero credits. Secondly, the location of the work and rest offers was presented in a counterbalanced fashion on the left or right of the screen.

Prior to the beginning of the task, the number of tick boxes used for the effort levels was calibrated to each participant’s skill level, similar to its laboratory version described in Chapter 2. In this calibration stage, participants were asked to tick as many boxes as they could within 10 seconds, up to 25, from top to bottom and from left to right. This task was repeated three times. Participant were told that they would receive higher bonus payments if they ticked a higher number of boxes. The third trial of calibration was used as a subject-specific threshold for the different levels of effort in the decision task.

4.3.5 Computational Modelling

As described in Chapters 2 and 3, I used computational models to precisely quantify decisions in both tasks. For the harm aversion task, the model explains choices by differences in subjective values between harmful and helpful options. The model describes this difference as follows:

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

$$\kappa = \begin{cases} \kappa_{self} \; if \; self \; trial \\ \kappa_{other} \; if \; other \; trial \end{cases}$$

Where ΔV is the difference between the subjective value of the harmful and helpful option, and Δm and Δs are objective differences in money and shocks between harmful and helpful options respectively. In this model, Δm and Δs are weighted by a free parameter κ, which represents harm aversion. When κ is 0, any profit will be accepted regardless the amount of shocks associated to it. When κ approaches to 1, participants become harm averse, sacrificing an increasing amount of profit to avoid additional shocks. Depending on who is receiving the shocks, κ_{other} captures the subjective cost of harming the Receiver (*other* trials), and κ_{self} explains the cost of harming oneself (*self* trials).

On the other hand, choices in the prosocial effort task were fitted using a model where the devaluation of reward by effort for self and other can be represented by the following formula:

$$SV = R - \lambda E^2$$

$$\lambda = \begin{cases} \lambda_{self} \; if \; self \; trial \\ \lambda_{other} \; if \; other \; trial \end{cases}$$

Where SV is the subjective value of the work offer with a certain reward R and effort E. The degree by which reward is discounted by effort is weighted by the discount parameter λ and is different for *self* and *other* trials. When λ is 0, the participant does not discount reward by effort. However, when λ approaches to 0.5, the reward is increasingly discounted by effort. I set λ with a maximum of 0.5 as for values beyond this constraint there is no change in the predicted choice behaviour in the task, i.e. all offered values are below the value of resting (i.e. 1 credit for 0 effort).

For both tasks a softmax function was used to transform trial-by-trial differences in value into choice probabilities. All parameters used in both the hypothetical harm aversion and the prosocial effort tasks were estimated individually for each participant.

4.3.6 Self-Report scales

The Standardised Assessment of Personality – Abbreviated Scale (SAPAS; Moran et al., 2003): The SAPAS is a short screening test that provides a valid measure of the likely presence of general PD. The SAPAS does not focus on one specific dimension of PD, essentially covering the broad multidimensional domain of PDs. Participants respond to eight dichotomously rated (Yes/No) items about the person's normal personality, scoring 0 if the PD trait is absent, and 1 if it is present. A score greater than or equal to 3 is used as a cut-off for PD.

McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD; Zanarini et al., 2003): The MSI-BPD consists of 10 yes/no items that cover the 9 diagnostic criteria for BPD according to the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013). A score equal to or over 7 would suggest the diagnosis of BPD.

Self-report Psychopathy scale – Short Form (SRP-4; Paulhus, Neumann, & Hare, 2009): The SRP-4 is a 29-item scale that measures psychopathic traits in non-institutionalised samples. Participants had to agree or disagree with each item in a 5-point Likert scale. Four subscales are included in this scale representing different dimensions of psychopathy: interpersonal, affective, lifestyle and antisocial

Depression and Anxiety Scales (extracted from DASS-21): The Depression Anxiety Stress Scales 21 (DASS-21) is a short form of the original 42-item self-report measure of depression, anxiety, and stress scale (DASS; Lovibond & Lovibond, 1995). Here, I included only the first two subscales, Depression, associated to dysphoric mood, and Anxiety, related to physical arousal. Each subscale consists of 7 items where participants have to rate how frequently and severely each statement has been experienced over the previous week in a 4-point scale.

Apathy-Motivation Index (AMI; Ang, Lockwood, Apps, Muhammed, & Husain, 2017): The AMI is an 18-item scale that measures apathy-motivation. Participants had to indicate their level of agreement with each item using a Likert scale from 0 to 4. Higher scores mean higher levels of apathy. The AMI comprises three subscales representing different domains of apathy: behavioural activation, social motivation and emotional sensitivity.

Toronto Alexithymia Scale (TAS-20; Bagby, Parker, & Taylor, 1994): The TAS-20 has 20 items rated using a 4 point scale and divided in 3 subscales: difficulty describing feelings, difficulty identifying feelings and externally-oriented thinking.

Interpersonal Reactivity Index (IRI; Davis, 1980): The IRI is a 28-item scale that measures empathy multidimensionally. Participants respond to each item using a five-point Likert scale. The IRI comprises 4 subscales: perspective taking, fantasy, personal distress and empathic concern.

4.3.7 Statistical analysis of behavioural data.

Non-parametric Wilcoxon signed-rank tests were used to test differences in *self* and *other* parameters in both tasks as these measures were not normally distributed. Likewise, Spearman's rank correlation coefficient was used to correlate parameters across tasks. For the prosocial effort task, two different ANOVAs were also performed to examine differences between conditions in choice behaviour for effort and reward and *self* versus *other* trials.

To test the relationship between choice parameters and psychiatric and affective traits, I performed a CCA to examine associations in a transdiagnostic manner across different traits. CCA is a data-driven multivariate approach that finds the maximal correlation between two sets of variables (Sherry & Henson, 2005), generating statistically independent pairs of synthetic variables referred to as canonical variates

(CV). Thus, CCA is a form of data dimensionality reduction, which has similarities to factor analysis and principal components analysis. However, the aim of CCA is to find what combination of variables (i.e. a CV) maximally correlates with another combination of variables (i.e. another CV), while factor analysis groups similar variables into dimensions (i.e. factors) with the aim of reducing data complexity.

The current study aimed to identify a CV comprised of affective and psychiatric traits that was maximally correlated with a CV comprised of the prosocial behaviour measures. I included SAPAS, MSI-BPD, DASS-21's Depression and Anxiety scores, as well as the subscales scores of TAS-20, AMI, SRP-4 and IRI for the affective and psychiatric trait CV. The prosocial domain CV included hyperaltruism (κ_{other} - κ_{self}) and prosocial effort (reverse-coded prosocial apathy, $\lambda_{other} - \lambda_{self}$) parameters in the computational models that quantified degrees of prosociality. The overlap from both sets of variables combines to create a canonical correlation coefficient (r) indexing the size of the relationship between the two CVs. By simultaneously considering all the variables in a single analysis, CCA allows us to understand how each trait affects prosocial behaviour while avoiding the inflation of experiment wise Type I error rates that typically occur when multiple univariate analyses are conducted (Sherry & Henson, 2005).

Using CCA it is possible to determine which variables contribute to explaining a CV, which variables are collinear within a CV, and which may have suppressor effects. Such an approach therefore has significant benefits over simply correlating a range of traits to individual task behaviours. To interpret a CCA, three sets of values are important (Sherry & Henson, 2005). First are the canonical loadings (structure coefficients) of each set of variables onto their own CV. These coefficients are analogous to the structure coefficients used in factor analysis structure matrix. Second are the canonical cross-loadings, which are the linear correlations between an observed variable from one set and the other CV. Cross-loadings are a more direct measure of how a variable (e.g. an affective or psychiatric trait) affects the predicted dimension (e.g. prosocial behaviour in both tasks).

Third are the canonical weights, the standardised canonical function coefficients that are used in each linear equation to combine the observed variables for each set into two corresponding CVs, maximising the correlation between the two CVs. These coefficients are analogous to beta weights in multiple regression. Importantly, discrepancies between these indicators are informative of multicollinearity and suppressor effects (Capraro & Capraro, 2001; Kuylen & Verhallen, 1981; Nimon et al., 2010). Thus, variables with high loadings but low weights may be indicative of multicollinearity, i.e., variance in these variables has been explained by other variables in the function coefficients. On the other hand, low loadings and high weights may mean that these variables are supressing irrelevant variance, i.e. variance shared with other variables in their own CV but not with the other CV.

I complemented CCA results with a commonality analysis (Nimon et al., 2010) with those variables that show either high loadings, high weights or both. This analysis provides the partition of the variance in two different forms: the explanatory power of the unique contribution of each variable, and the explanatory ability of that variable that is common with the other variables. In doing so it is possible to support identification of multicollinearity and suppressors. Prior to running the CCA, questionnaire scores and parameter estimates were z-scored across participants. I considered moderate to high loadings ($x \geq 0.3$ and $x \leq -0.3$) for interpretation (Lambert & Durand, 1975). Furthermore, I additionally performed a permutation analysis for statistical significance of the CCA using the *p.perm* function implemented in the CCP package in R (Roy's Largest Root test, 1000 permutations). Two-tailed tests were used for all analyses.

4.4 Results

4.4.1 Study 1: Hyperaltruism and prosocial apathy effects are replicated in online samples.

I first tested whether hyperaltruism – greater aversion to harming others for money than oneself - was present in the online hypothetical data. Results revealed that participants chose the helpful option significantly more in *other* (M = 46% of trials, SEM = 3%) compared to *self* (M: 37%, SEM: 2%) trials ($z = -2.82$, $p < 0.01$, Fig. 4.2A), replicating the previously observed hyperaltruism effect. Furthermore, using computational models, I found greater harm aversion for others (M = 0.45, SEM = 0.03) than self (M = 0.37, SEM = 0.02; $\kappa_{other} > \kappa_{self}$, $z = -2.28$, $p < 0.03$; Fig. 4.2B), suggesting that even when the pain and money are hypothetical, participants give up more money to prevent harm to others compared to self.

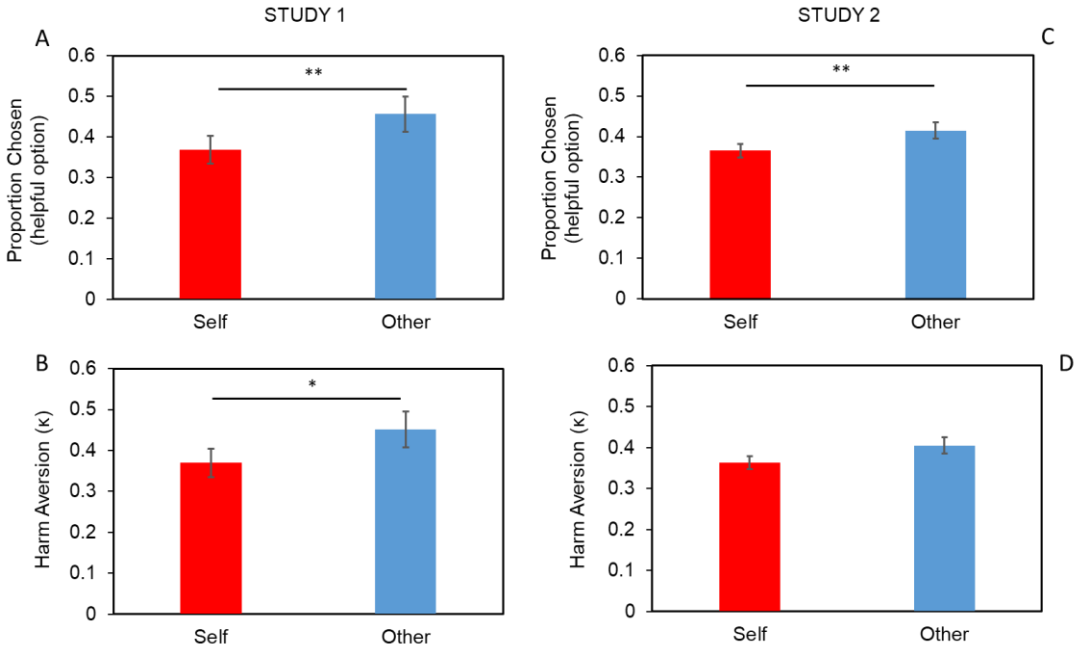


Figure 4.2. Results in the hypothetical harm aversion task for studies 1 and 2. A. Proportion of helpful option chosen over the harmful option in study 1. Participants chose in a higher frequency the helpful option for others than themselves. B. Participants were more averse to harm others compared to themselves in study 1 (i.e. $\kappa_{other} > \kappa_{self}$). C. Study 2 replicates behavioural (i.e. proportion of helpful option chosen) results of study 1. D. However, even though participants were more averse to harm others compared to themselves in study 2 (i.e. $\kappa_{other} > \kappa_{self}$), this result was not significant using computational parameters. Error bars depict S.E.M. * $p < 0.05$; ** $p < 0.01$.

For the prosocial effort task, I used ANOVAs with proportion of choosing the 'work' option as a dependent variable to test whether choosing effortful actions, over rest, to get rewards depends on the beneficiary (*self* vs *other*). Results revealed an effort x beneficiary interaction ($F_{2.1, 236.9} = 7.74$, $p < 0.001$, Fig. 4.3A), and main effects of effort ($F_{1.7, 185} = 91.43$, $p = < 0.001$) and beneficiary ($F_{1, 112} = 93.55$, $p < 0.001$). Participants were less willing to choose the 'work' option when the levels of effort increased. This effect was amplified when the other person was the beneficiary, with participants less willing to exert effortful actions when the reward was received by the other person compared to themselves. On the other hand, main effects of reward ($F_{1.5, 163} = 48.77$, $p < 0.001$) and beneficiary ($F_{1, 112} = 93.55$, $p < 0.001$) were also found, but no interaction between reward x beneficiary ($F_{1.7, 189.3} = 3.03$, $p = 0.59$; Fig. 4.3B). Finally, participants showed higher devaluation of others' reward by effort (M = 0.11, SEM = 0.02) compared to self (M = 0.02, SEM = 0.01; $\lambda_{other} > \lambda_{self}$, $z = -8.34$, $p < 0.001$; Fig. 4.3C) using the computational parameters. Thus, participants chose to work less to benefit others compared to self, replicating the prosocial apathy effect

4.4.2 Study 2: Replicating Hyperaltruism and Prosocial Apathy effects

Study 2 sought to replicate the effects found in study 1. Thus, in the harm aversion task participants chose the helpful option significantly more for the *other* person (M = 42% of trials, SEM = 2%) compared to *self* trials (M = 37% of trials, SEM = 2%; $z = -3.0$, $p < 0.005$, Fig. 4.2C) replicating study 1. The model parameters showed effects in the same direction, such that participants showed higher aversion to harming others ($\kappa_{other} = 0.41$, SEM = 0.02) compared to themselves ($\kappa_{self} = 0.36$, SEM = 0.02), although this trend was not significant ($z = -1.63$, $p = 0.10$; Fig. 4.2D).

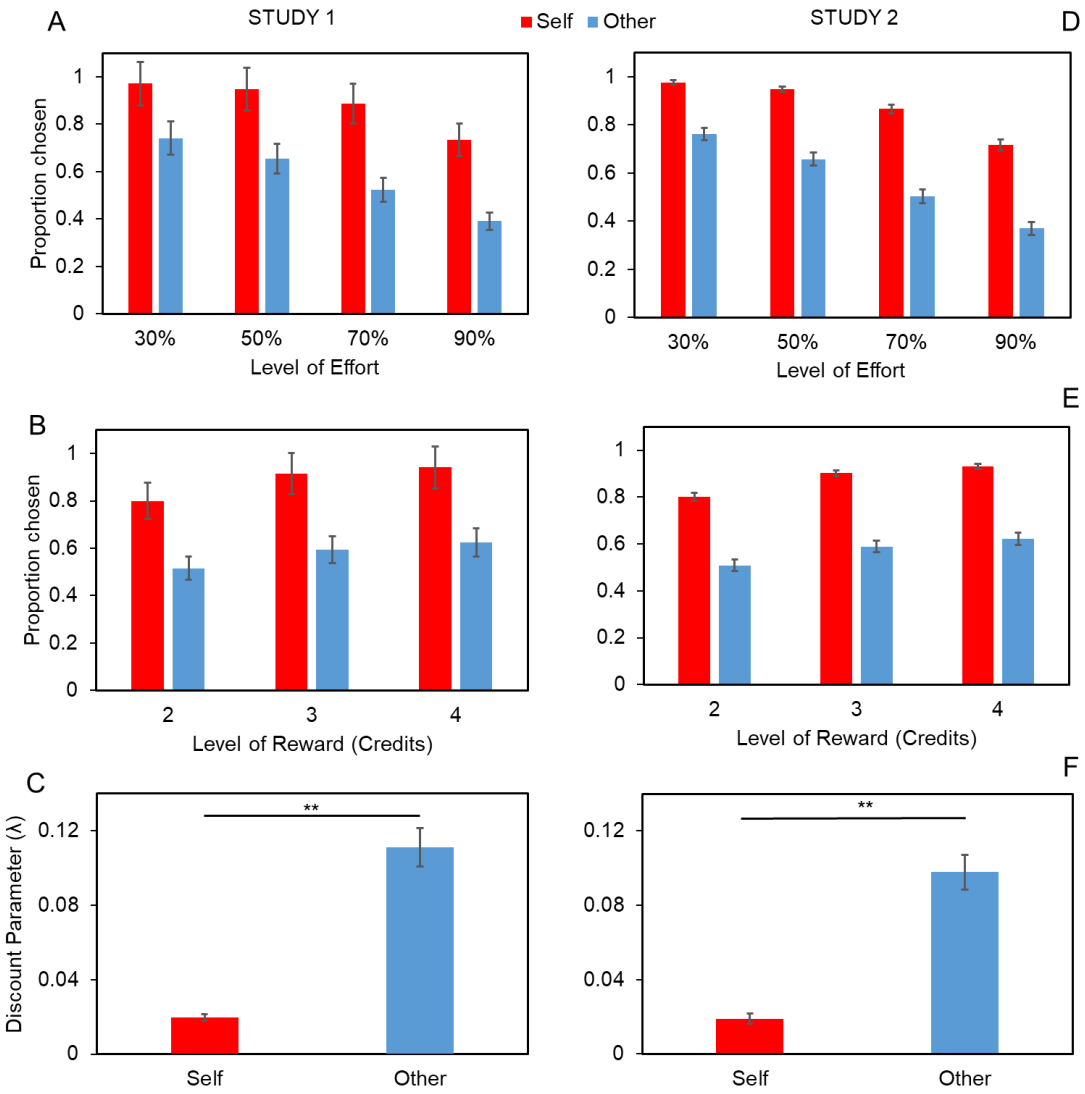


Figure 4.3. Results in the prosocial effort task for study 1 and 2. A. Proportion of 'work' offer chosen over the baseline option plotted against effort in study 1. Participants chose to work for self than other at a higher frequency. This difference increased as the effort level increases. B. Proportion of 'work' offer chosen over the baseline option plotted against reward in study 1. Participants chose in a higher frequency to work for self than other. Furthermore, they chose more often higher reward options, but there was no interaction between the target of the reward and the number of credits associated to the option. C. Participants discounted reward to a higher degree by effort when the receiver was the beneficiary compared to themselves in study 1 (i.e. $\lambda_{other} > \lambda_{self}$). D, E and F. Study 2 fully replicates behavioural (D and E) and computational (F) results of study 1. Error bars depict S.E.M. ** $p < 0.01$.

For the prosocial effort task all effects of study 1 were replicated. There were main effects of effort ($F_{1.7,\ 363.1} = 163.03$, $p < 0.001$), beneficiary ($F_{1,\ 211} = 187.18$ $p < 0.001$), and more importantly, an interaction between effort and beneficiary ($F_{2.3,\ 475.75} = 20.67$, $p < 0.001$, Fig. 4.3D) on the proportion of 'work' choices. Main effects of beneficiary ($F_{1,\ 211}$=

187.2, $p < 0.001$) and reward ($F_{1.51, 317.9} = 111.3$, $p < 0.001$) were also found, but no interaction between these variables ($F_{1.9, 398.41} = 1.3$, $p = 0.27$; Fig. 4.3C). Finally, results also revealed greater discounting for *other* (M = 0.1, SEM = 0.01) compared to *self* (M = 0.02, SEM = 0.003; $\lambda_{other} > \lambda_{self}$, $z = -11.9$, $p < 0.001$, Fig. 4.3F).

As a control analysis, I performed a 2x2 mixed ANOVA having beneficiary/recipient as a within subject factor (self, other) and task order as a between subject factor (harm aversion task first, prosocial effort task first) to test whether hyperaltruism or prosocial apathy effects were influenced by the order in which the harm aversion and the prosocial effort tasks were completed in Study 2. This analysis was performed on both proportion of choices on one hand, and computational parameters on the other hand for both tasks separately. Proportions of choices and computational parameters revealed similar results. For the prosocial effort task, there were not effects of task order (neither main effect nor interaction, $p > 0.1$), and only a main effect of beneficiary (Proportions: $F_{1,210} = 189.7$, $p < 0.001$; Parameters: $F_{1,210} = 77.9$, $p < 0.001$). However, for the hypothetical harm aversion task I found, together with a main effect of recipient (Proportions: $F_{1,210} = 7.6$, $p < 0.01$; Parameters: $F_{1,210} = 5.1$, $p < 0.05$), a main effect of task order (Proportions: $F_{1,210} = 15.2$, $p < 0.01$; Parameters: $F_{1,210} = 14.8$, $p < 0.01$), with people being more averse to harm themselves and others when this task was completed first (Proportions: M = 0.45, SEM = 0.2; Parameters: M = 0.44, SEM = 0.2) rather than second (Proportions: M = 0.33, SEM = 0.2; Parameters: M = 0.32, SEM = 0.2). Nevertheless, and crucially, there was not an interaction between recipient and task order ($p > 0.6$), suggesting that the hyperaltruism effect was not influenced by the order in which the tasks were completed. Thus, hyperaltruism and prosocial apathy effects were robust across task orders and across the two studies.

4.4.3 Prosocial behaviours in both tasks are correlated

Before testing the correlation between prosocial behaviour in each task, I collapsed both samples in order to maximise statistical power and prepare the data for the CCA. To begin, I examined the hyperaltruism and prosocial apathy effects in the combined sample (n = 325). I found consistent results with those revealed by taking the samples separately in both proportion of choices (Fig. 4.4A for harm aversion task; Fig. 4.4B and 4.4C for prosocial effort task) and computational parameters (Fig. 4.5A and 4.5B for harm aversion task; Fig. 4.5C and 4.5D for prosocial effort task). Thus, participants chose in a higher proportion the helpful option in *other* (M = 43%, SEM = 2%) than *self* trials (M = 37%, SEM = 1%; $z = -4.09$, $p < 0.001$). This was supported by a comparison between κ_{self} and κ_{other} showing that participants were significantly more averse to harming other (M = 0.42, SEM = 0.02) compared to themselves (M = 0.37, SEM = 0.01; $z = -2.62$, $p < 0.01$). Furthermore, in the prosocial effort task, there were main effects of beneficiary ($F_{1, 324} = 281.27$, $p < 0.001$), effort ($F_{1.7, 552.78} = 252.97$, $p < 0.001$) and their interaction ($F_{2.2, 714.68} = 28.13$, $p < 0.001$), as well as main effects of reward ($F_{1.5, 481.38.2} = 158.64$, $p < 0.001$), beneficiary ($F_{1, 324} = 281.27$, $p < 0.001$) and their interaction ($F_{1.84, 596.77} = 3.73$, $p < 0.03$). Comparing model parameters, participants significantly discounted more reward by effort in *other* (M = 0.1, SEM = 0.01) than *self* trials (M = 0.02, SEM = 0.002; $z = -14.51$, $p < 0.001$).

Is there a relationship between prosocial behaviour across both tasks? I reverse-coded the prosocial apathy effect ($\lambda_{other} - \lambda_{self}$) in the prosocial effort task and called this variable prosocial effort. Thus, higher values were associated with more prosocial decisions in both tasks. Using these parameters, I found that hyperaltruism ($\kappa_{other} - \kappa_{self}$) and prosocial effort were positively correlated ($\rho = 0.33$; $p < 0.001$; Fig. 4.5E), such that participants who were prosocial in one task were also prosocial in the other. Importantly, this correlation in prosocial behaviour across both tasks is consistently found if the sample is split in studies 1 ($\rho = 0.38$; $p < 0.001$) and 2 ($\rho = 0.3$; $p < 0.001$), indicating that

this relationship is robust. Thus, these results suggest that there is a common prosocial disposition across tasks - people who avoid (hypothetically) harming others also more readily exert effort to earn rewards for others, replicating the lab-based results reported in Chapter 3 in online samples.

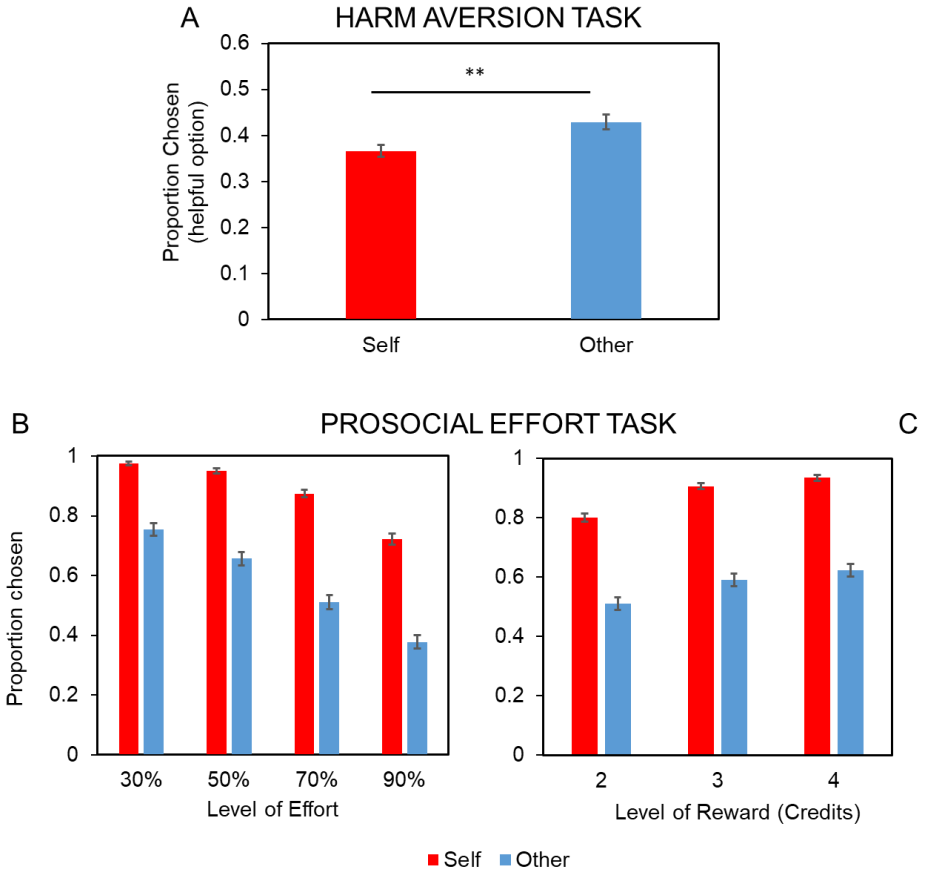


Figure 4.4. Behavioural results of collapsed samples of studies 1 and 2 (n=325). Hyperaltruism and prosocial apathy are manifested in the hypothetical harm aversion and in the prosocial effort tasks respectively. A. Proportion of helpful option chosen over the harmful option in the hypothetical harm aversion task. Participants chose in a higher frequency the helpful option for others than themselves. B. Proportion of 'work' offer chosen over the baseline option plotted against effort in the prosocial effort task. Participants chose in a higher frequency to work for self than others. This difference increased while the effort levels augmented. C. Proportion of 'work' offer chosen over the baseline option plotted against reward in the prosocial effort task. Participants chose to work for self than other at a higher frequency. Furthermore, they chose more often higher reward options, with an interaction between the target of the reward and the number of credits associated to the option. Error bars depict S.E.M. ** $p < 0.01$

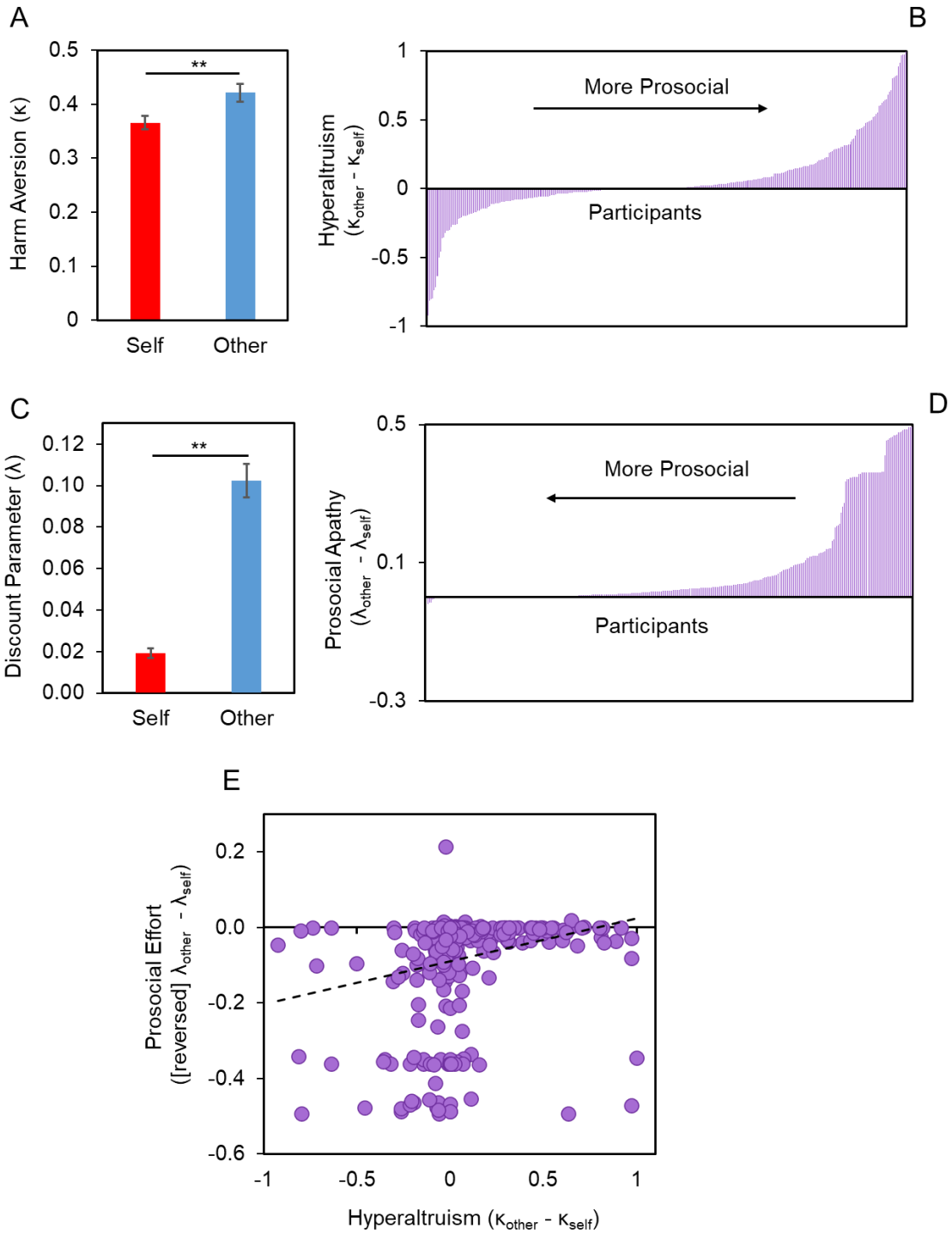


Figure 4.5. Computational results of collapsed samples of studies 1 and 2 (n=325). Hyperaltruism and prosocial apathy in the hypothetical harm aversion task and prosocial effort task respectively are supported by model results. A. Participants were more averse to harm others compared to themselves in the hypothetical harm aversion task (i.e. $\kappa_{other} > \kappa_{self}$). B. Distribution of hyperaltruism (i.e. $\kappa_{other} - \kappa_{self}$) among deciders in the hypothetical harm aversion task. 52.9% of the sample showed the hyperaltruism effect. C. Participants discounted reward to a higher degree by effort when the receiver was the beneficiary compared to themselves in the prosocial effort task (i.e. $\lambda_{other} > \lambda_{self}$,). D. Distribution of prosocial apathy (i.e. $\lambda_{other} - \lambda_{self}$) among deciders in the prosocial effort task. 93.4% of the sample showed the prosocial apathy effect. E. Significant correlation between hyperaltruism and prosocial effort (rho = 0.33; $p < 0.001$), suggesting common prosocial mechanisms across tasks. Error bars depict S.E.M. ** $p < 0.01$.

4.4.4 Transdiagnostic relationship between affective and psychiatric traits and prosocial dispositions

To examine the traits associated with higher prosocial behaviour across contexts, I performed a CCA between participants' scores in the affective and psychiatric scales, and prosocial behaviour in both tasks. I aimed to identify what constellation of affective and psychiatric traits were more associated with a CV comprised of prosocial behaviour measures across tasks. For the psychiatric and affective traits CV, I included general PD, borderline PD, depression and anxiety trait scores, as well as multiple dimensions of apathy, alexithymia, psychopathy and empathy traits. Table 4.1 provides information about the descriptive statistics of each trait in the collapsed sample. Importantly, all variables show considerable variance across the whole scoring range. On the other hand, for the prosocial CV, I included hyperaltruism (κ_{other} - κ_{self}) and prosocial effort (reversed λ_{other} – λ_{self}) parameters in the computational models that quantified degrees of prosociality.

Table 4.1. Descriptive information of psychiatric and affective traits

	Mean	SD	Range	Scoring Range
SAPAS	2.8	1.9	8	0-8
MSI-BPD	2	2.6	10	0-10
SRP-I	12.9	5.4	27	1-35
SRP-Aff	13.4	5	21	1-35
SRP-L	13	5	21	1-35
SRP-Ant	9.7	3.8	21	1-35
DASS-Dep	3.9	5.2	21	0-21
DASS-Anx	2.4	3.5	19	0-21
TAS-DF	11.3	4.4	20	1-25
TAS-IF	12.3	5	21	1-35
TAS-ET	19	5.1	27	1-40
AMI-ES	1.33	0.74	4	0.0-4.0
AMI-BA	1.4	0.87	4	0.0-4.0
AMI-SM	2.09	0.88	4	0.0-4.0
IRI-FT	23.7	6.4	28	1-35
IRI-EC	26.2	6.3	28	1-35
IRI-PT	25.8	6.1	28	1-35
IRI-PD	16.7	6	28	1-35

SRP-I = SRP-Interpersonal; SRP-Aff = SRP Affective; SRP-L = SRP-Lifestyle; SRP-Ant = SRP-Antisocial; DASS-Dep = DASS-21 Depression; DASS-Anx = DASS-21 Anxiety; TAS-DF = TAS20 Difficulty in Describing Feelings; TAS-IF = TAS20 Difficulty in Identifying Feelings; TAS-ET = TAS-20 Externally Oriented Feelings; AMI-ES = AMI Emotional Sensitivity; AMI-BA = AMI Behavioural Activation; AMI-SM = AMI Social Motivation; IRI-FT = IRI Fantasy; IRI-EC = IRI Empathic Concern; IRI-PT = IRI Perspective Taking; IRI-PD = IRI Personal Distress. SAPAS = The Standardised Assessment of Personality – Abbreviated Scale.

Two significant canonical correlations were generated in this analysis, as in CCA the number of canonical correlations is equal to the number of variables in the smaller of the two sets (first canonical correlation: $r = 0.37$; Wilks's lambda = 0.79, $F_{36, 610} = 2.18$; $p < 0.001$, second canonical correlation $r = 0.3$; Wilks's lambda = 0.91, $F_{17, 306} = 1.72$; $p < 0.05$). However, given that the first canonical function explained most of the variance in the canonical solution, and that the second canonical correlation had prosocial behaviour for each task loading in opposite directions (i.e. with one loading positively and the other negatively, not extracting the same variance across tasks), I further analysed and interpreted results related only to the first significant canonical function. In order to add robustness to these results, I performed a permutation Roy's Largest Root test for significance on this first canonical correlation, supporting a significant association between prosocial behaviours and psychiatric and affective traits (1000 permutations; $p < 0.001$).

I looked at loadings (structure coefficients), cross-loadings and weights (function coefficients) to interpret the results revealed by CCA. I further complemented CCA results with a commonality analysis on those variables that showed either high loadings, high weights or both in order to support identification of multicollinearity and suppressor suggested by CCA (see Methods). Consistent with simple correlations, a shared prosocial variance was revealed across tasks, with prosocial effort and hyperaltruism strongly (> 0.3) loaded onto their own CV, suggesting that this dimension encapsulated both measures of prosociality (loadings: hyperaltruism = 0.79; prosocial effort = 0.77).

Furthermore, both showed high weights, consistent with the loading results (weights: hyperaltruism = 0.65; prosocial effort = 0.62).

Regarding the psychiatric and affective CV, results revealed a consistent effect of empathic concern being positively, and emotional apathy together with externally-oriented thoughts (a dimension of alexithymia) negatively loaded onto their own CV and cross-loading with the prosocial CV (Fig. 4.6A and 4.6B). These effects were consistently found looking at the weights and the results shown by the commonality analysis (Fig. 4.6C and 4.6D, Table 4.2). Thus, people with higher levels of empathic concern and lower levels of emotional apathy, together with more attention to their internal feelings, were more averse to harming others and more willing to put in effort to benefit others.

Not all variables showed consistently high values across loadings and weights. Discrepancies between both measures are not uncommon and arise due to multicollinearity between variables, and suppression effects across the trait measures. Weights are the contribution of a variable to the canonical solution as a whole, as such they are sensitive to the presence of multicollinearity. Loadings, on the other hand, are not affected by multicollinearity, as they correspond to the unique contribution of each trait to their own affective/psychiatric CV computed separately. Thus, some traits can have small weights because the variance of these variables could have been explained by other traits in the equation (Kuylen & Verhallen, 1981). Furthermore, high weights do not necessarily mean that there is a high contribution of that trait to the prediction of prosocial behaviour. Variables with high weights but low loadings have been suggested as suppressors in linear equations (Capraro & Capraro, 2001; Nimon et al., 2010), meaning that traits showing such a pattern would not be directly related to prosocial behaviour, but rather they would be subtracting irrelevant variance of other traits in the affective/psychiatric CV to increase their predictive power in relation to the prosocial behaviour CV.

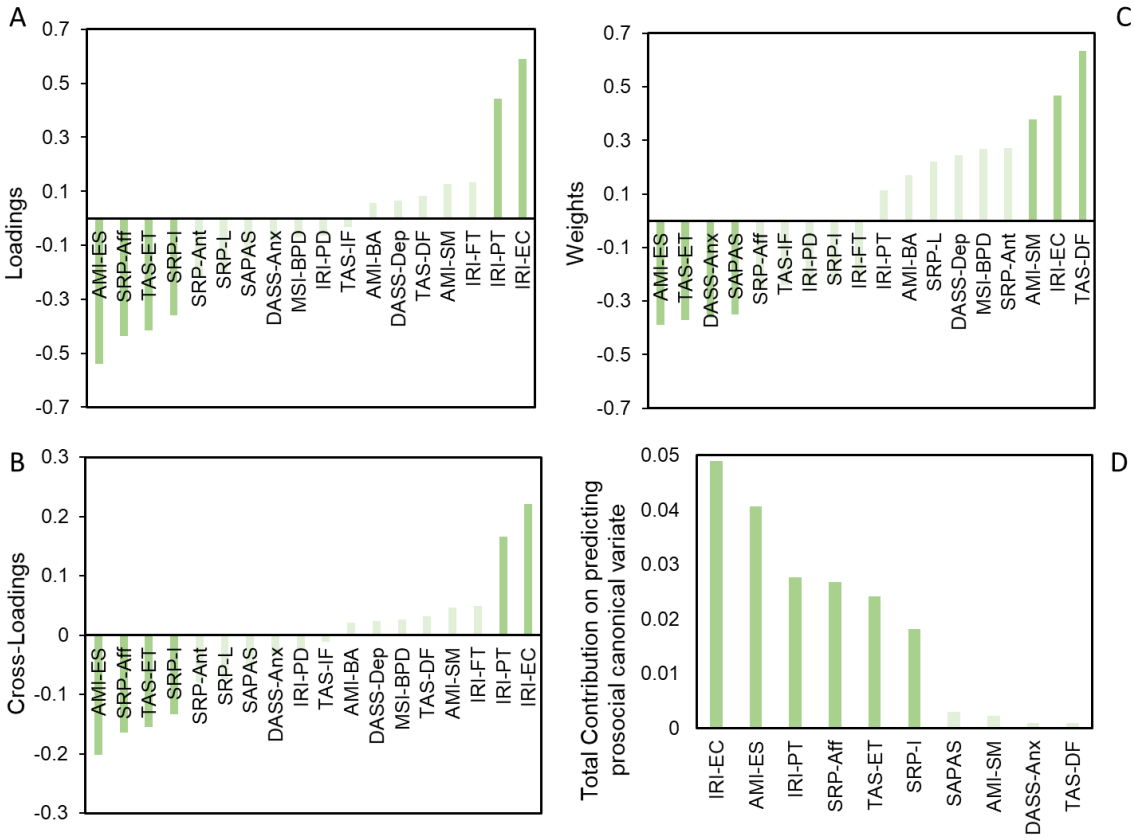


Figure 4.6. Constellation of traits that are correlated with prosocial behaviour. A. Loadings of variables in the affective and psychiatric CV. B. Cross-loadings of variables with the prosocial CV. C. Weights of variables in the affective and psychiatric CV. Note the discrepancies with the loadings suggesting multicollinearity and suppressor effects. D. Total effects of variables with high loadings and weights of the affective and psychiatric traits CV on the prosocial CV. This supports what it was found by weights and loadings: empathic concern and emotional apathy are the most important variables in predicting prosociality. SRP-I = SRP-Interpersonal; SRP-Aff = SRP Affective; SRP-L = SRP-Lifestyle; SRP-Ant = SRP-Antisocial; DASS-Dep = DASS-21 Depression; DASS-Anx = DASS-21 Anxiety; TAS-DF = TAS20 Difficulty in Describing Feelings; TAS-IF = TAS20 Difficulty in Identifying Feelings; TAS-ET = TAS-20 Externally Oriented Feelings; AMI-ES = AMI Emotional Sensitivity; AMI-BA = AMI Behavioural Activation; AMI-SM = AMI Social Motivation; IRI-FT = IRI Fantasy; IRI-EC = IRI Empathic Concern; IRI-PT = IRI Perspective Taking; IRI-PD = IRI Personal Distress. SAPAS = The Standardised Assessment of Personality – Abbreviated Scale.

Results revealed that perspective taking contributed positively to its own CV, while interpersonal and affective psychopathy contributed negatively (Fig. 4.6A). However, when examining the weights and the commonality analysis, it is apparent that this is because these variables were collinear with other variables in the CV (Fig. 4.6C and 4.6D; Table 4.2). At the same time, although general PD, anxiety, social apathy and one dimension of alexithymia showed high weights (Fig. 4.6C), they did not contribute highly

to their CV (Fig. 4.6A), suggesting suppressor effects, which is supported by their negative common effects and marginal total effects in the commonality analysis results (Fig. 4.6D, Table 4.2).

In summary, emotional apathy, empathic concern and externally-oriented thinking appear to have a reliable relationship with prosocial behaviour, with the former two being predominant, such that higher levels of empathic concern and lower degrees of emotional apathy are associated with prosocial behaviour across tasks. Furthermore, and collinear with these variables, perspective-taking and affective dimensions of psychopathy were also related to prosocial behaviour. This suggests that there is a broad cluster of affective traits that is associated with prosocial behaviour across different contexts.

Table 4.2. Commonality analysis results

	Unique	Common	Total
SAPAS	0.0065	-0.0035	0.0029
SRP-I	0.0003	0.0178	0.0181
SRP-Aff	0.0001	0.0267	0.0268
DASS-Anx	0.0025	-0.0016	0.0010
TAS-DF	0.0258	-0.0249	0.0009
TAS-ET	0.0102	0.0140	0.0242
AMI-ES	0.0043	0.0363	0.0406
AMI-SM	0.0112	-0.0090	0.0022
IRI-EC	0.0096	0.0393	0.0489
IRI-PT	0.0010	0.0267	0.0276

4.5 Discussion

In this chapter, I examined the relationship between affective and psychiatric traits with prosocial behaviour manifested in harm aversion and prosocial effort contexts. Using a transdiagnostic approach, I found that higher levels of empathic concern and lower levels of emotional apathy were the traits associated the most with prosocial behaviour

across contexts. Furthermore, affective aspects of psychopathy and alexithymia were associated with less prosocial behaviour, while perspective taking also facilitated prosociality. In contrast no psychiatric traits, except for the affective component of psychopathy, were associated with prosocial behaviour when controlling for suppression and multicollinearity effects. Taken together, these results suggest that affective reactivity across a broad cluster of traits including apathy, psychopathy and empathy are related to common mechanisms in prosocial behaviour.

High empathic concern and low emotional apathy were the strongest predictors amongst several affective and psychiatric traits putatively suggested to relate to prosocial behaviour. Intriguingly, empathy was a facilitating factor even when traits associated with obstructing prosocial behaviour were also included, suggesting that empathic concern is a powerful variable to trigger prosocial behaviour over and above individual differences in psychiatric disorders and other affective traits. This in line with dimensional models for PD used in the DSM-5, where impairment in empathy is a major trait to identify pathological disturbances in interpersonal functioning (American Psychiatric Association, 2013). Moreover, low levels of emotional apathy predicted more prosocial decisions. Thus, this would suggest that motivation to engage emotionally, alongside empathic feelings, contributes to prosocial behaviour (Contreras-Huerta et al., 2020; Decety & Cowell, 2014a, 2014b). In order to recognise and care about others' emotions and act accordingly, people may have to be sensitive to experience emotions and be motivated by them. In this way, emotional apathy and empathy are constructs that may share common processes (Lockwood, Ang, et al., 2017; Zaki, 2014), and together may be necessary to facilitate prosocial behaviour.

Other traits were also related to prosocial behaviour, namely, perspective taking, and aspects of alexithymia and psychopathy. Perspective taking was positively related to prosocial behaviour, similar to empathic concern, consistent with the importance of understanding the emotions of others in order to help them, a more cognitive aspect of

empathic processes (Tamnes, Overbye, Ferschmann, Fjell, & Walhovd, 2018; Tusche et al., 2016). However, empathic concern, a more affective aspect of empathy, had a more predominant role in predicting prosocial behaviour. In fact, emotional components across different traits were associated with prosocial behaviour. Thus, only the emotional aspect of apathy, previously linked to affective but not cognitive empathy (Lockwood, Ang, et al., 2017), was associated to prosocial behaviour in the current results. Furthermore, people who do not avoid thinking about their emotions were more likely to be prosocial. In fact, previous evidence has shown that externally-oriented thinking is associated to callous traits in psychopathy (Lander et al., 2012), low empathic concern (Grynberg et al., 2010; Guttman & Laporte, 2002), less reaction to affective stimuli (Wiebe, Kersting, & Suslow, 2017), and alterations in moral reasoning and prosocial behaviour (Gleichgerrcht, Tomashitis, & Sinay, 2015; Mannarini, Balottin, Toldo, & Gatta, 2016). Finally, the only specific psychiatric trait predicting prosocial behaviour in the present data was the affective components of psychopathy, associated to callous feelings, but not its behavioural aspects (e.g. antisociality, impulsivity), in line with previous studies (Lockwood, Hamonet, et al., 2017; Vahl et al., 2014; White, 2014). Taken together, the transdiagnostic approach adopted in this study identified a common prosocial disposition characterised mostly by affective sensitivity – people who engaged affectively with themselves and others were more likely to assume costs to benefit other people.

No significant effects of specific psychiatric traits in prosociality were apparent in these results, except for the callous component of psychopathy. Psychopathy is not a clear category in classic psychiatric classification, being traditionally associated with antisocial PD that typically focuses on more behavioural aspects of psychopathy rather than its callous symptomatology (Crego & Widiger, 2015). However, new dimensional conceptualisations of PD, included also in the DSM-5, that represent extremes along a trait-dispositional continua, have captured callous traits, among others, to define psychopathic symptomatology (American Psychiatric Association, 2013; Miller, Lamkin,

Maples-Keller, Sleep, & Lynam, 2018; Strickland, Drislane, Lucy, Krueger, & Patrick, 2013). These results support this conceptualisation, suggesting that strict clinical categorisations may be not necessary to understand variability in prosocial behaviour. Rather, a constellation of broad traits associated to affective sensitivity across psychiatric and healthy population may better predict prosociality, suggesting that extreme cases of low affective reactivity could result in social dysfunction. Thus, these results may shed light on identifying risk and protective factors in a wide spectrum of traits for keeping healthy social relationships. Social dysfunction is a transversal symptom across many psychiatric disorders, and these findings could help illuminate a different perspective to understand and approach to this problem through the propensity for prosocial behaviour.

In this chapter, I used an analytical approach that allows for the multicollinearity that exists across different traits to be accounted for, a problem that has been identified in the psychiatric literature. For instance, the present findings suggest that some relationships between traits and prosocial behaviour in this study, e.g. PD and anxiety, are due to shared variance with other traits but not with prosocial behaviour itself, consistent with some previous findings (Gleichgerrcht & Young, 2013; Lander et al., 2012; Mannarini et al., 2016; Vahl et al., 2014; White, 2014). Thus, broad affective traits may be better explaining variability in prosocial behaviour than categorical psychiatric disorders. Furthermore, affective traits such as empathic concern and emotional apathy are collinear with other traits but at the same time show strong associations with prosocial behaviour. Therefore, these results open a potential line of research to disentangle the complex relationship between traits and other traits, traits and behaviour, and traits and specific psychiatric disorders. Even though this topic is out of the scope of this thesis, future research may shed light on how the interaction across different traits can influence ultimate prosocial decisions.

In summary, in this chapter I have shown that people display prosocial behaviour in two different contexts that are modulated by individual differences in broad affective traits.

People who showed a configuration of traits characterised by high empathy and affective reactivity displayed more aversion to harming others and were more willing to exert effort to benefit others. Reduced empathic feelings and affective sensitivity are present in a variety of psychiatric disorders, suggesting that these features may be associated with the poor social relationships seen in a high number of mental health conditions. Thus, these findings shed light on the traits that may underlie social dysfunction in general and clinical population, with the potential of aligning clinical categorisation with cognitive-based evidence, and thereby improving diagnostic accuracy, identifying high-risk populations, and setting more guided systematic treatments that improve people's social relationships and mental health.

Chapter 5

Motivational aspects of moral hypocrisy

5.1 Summary

People often act hypocritically. One form of moral hypocrisy occurs when people blame others for transgressing moral principles that they themselves have violated in the past. Philosophical theories suggest that people simply apply different moral standards to other people than to themselves. Yet, many aspects of social cognition are under the influence of motivation. We are not always willing to exert the effort required to obtain desired outcomes. In this chapter, I hypothesise that hypocrites are simply not willing to exert the effort required to live up to moral principles. I used a computational model of a measure of moral hypocrisy to quantify the discrepancy between how willing people are to profit from another's harm, and how much they blame somebody else for similarly profiting. Participants completed a decision task where their moral preferences were measured. One week later, participants performed another task where they judged the blameworthiness of identical decisions made by others. Furthermore, participants completed a prosocial effort task that measured (i) participants' willingness to effortfully help others, and (ii) how much effort they exerted when they decided to help. Using these tasks, results revealed that moral hypocrisy is associated with reduced motivation in these two aspects, suggesting that the discrepancy between moral standards and actions could be related to reduced willingness to overcome prosocial costs.

5.2 Introduction

Moral principles often guide how people act and are reinforced in society through the judgements people make of others. Judging someone as morally virtuous or vicious, depending how much they deviate from moral principles, serves to perpetuate them as norms. It has often been assumed that people guide both their moral decisions and their moral judgements according to the same standards. However, this is not usually the case (Alicke et al., 2013; Batson et al., 2002; Coates & Tognazzini, 2013; Effron, et al., 2018; Merritt & Monin, 2012). In fact, it is common that people show, at least to some degree, discrepancies between their actions and how harshly they judge others. For example, consider a politician who publicly condemns violations to Covid-19 restrictions as selfish, but then violates those same rules when it is in their interest. There is a clear discrepancy between the politician's moral judgements and their moral actions. This discrepancy has been identified by philosophers, psychologists, and the general public as a form of moral hypocrisy (Alicke et al., 2013; Dover, 2019; Effron et al., 2018; Howe & Monin, 2017; Laurent & Clark, 2019).

Philosophical accounts have proposed that such hypocrisy is due to having stricter moral standards for others, but softer for ourselves, i.e. a double standard for self and other actions, involving thereby deceptive intentions - actively putting on a moral appearance to hide one's immoral reality (Dover, 2019; Fritz & Miller, 2018; Graham, Meindl, Koleva, Iyer, & Johnson, 2015; Howe & Monin, 2017; Kittay, 1982; Laurent & Clark, 2019; Szabados & Soifer, 2004; Todd, 2019; Valdesolo & DeSteno, 2008; Wallace, 2010). However, the psychology underlying hypocrisy is still poorly understood, and theories have not addressed the cognitive mechanisms underlying individual differences in the degree in which hypocrisy occurs. One potential factor that could underlie a judgement-action discrepancy in hypocrisy is how motivated the person is to incur in self-costs for the benefit of others (Batson & Thompson, 2001; Batson et al., 2002; Lindenberg, Steg, Milovanovic, & Schipper, 2018). The politician in the example above

could have violated the Covid-19 rules simply to avoid how demanding the effort of complying is. From this perspective, the hypocrite struggles to conform to the moral norm due to a failure in motivation. Thus, even though hypocrites could genuinely share the moral standards they use to judge and blame others, self-costs could overpower their good faith (Batson et al., 2002; Lindenberg et al., 2018; Watson & Sheikh, 2008), failing to live up to their genuine moral standards (Bartel, 2019; Batson & Thompson, 2001; Dover, 2019; Mele, 1989).

There is some evidence supporting a motivational account of moral hypocrisy. People who are more morally self-motivated show less moral hypocrisy compared with people who are driven to act morally only because of social compliance (Carpenter & Marshall, 2009; Graham et al., 2015). Furthermore, if the costs (e.g. effort) to behave morally increase, people fail to live up to their moral standards (Batson & Thompson, 2001; Lindenberg et al., 2018). Importantly, actively thinking about how others benefit from participants' moral actions can help overcome cost avoidance (Batson et al., 2003), suggesting that a switch in information saliency can be enough for people to comply with their values and overcome the costs. Therefore, compliance with moral standards could be taken as a motivated decision - people need to feel sufficiently incentivised by others' benefits to overcome self-costs, and in this way retain moral integrity (judgement-behaviour alignment) and avoid moral hypocrisy (judgement-behaviour discrepancy). If so, people who are less motivated to help others may be the most likely to be hypocritical.

In this chapter I investigate the relationship between moral hypocrisy and motivation to help others. Specifically, I focus on one of the most common costs in goal-directed behaviour – effort (Contreras-Huerta et al., 2020). In many situations, helping others requires effort, and people must trade-off these effort costs against the social benefits of those actions. If moral hypocrisy is associated with avoiding self-costs, then people with wider gaps between their moral judgements and actions would putatively have less motivation to put in effort to help others. Here, I operationalise individual differences in

moral hypocrisy as the discrepancy between decisions in a moral task (harm aversion task, Crockett et al., 2014, 2015, 2017) and the blame assigned on those same decisions made by others. The discrepancy between actions and judgment has been defined in the past as a type of moral hypocrisy (Alicke et al., 2013; Barden et al., 2005; Dover, 2019; Effron, Markus, Jackman, Muramoto, & Muluk, 2018; Effron, O'Connor, et al., 2018; Howe & Monin, 2017; Jordan et al., 2017; Laurent & Clark, 2019; O'Connor et al., 2020), regardless of whether the moral standards are deceptive or genuinely believed. Here, I only focus on this behavioural manifestation of hypocrisy and its relationship with motivation.

Prosocial motivation was measured using the prosocial effort task, in which participants trade-off different levels of physical effort in exchange of different magnitudes of reward received by either participants themselves or another unknown person (Lockwood, Hamonet, et al., 2017, Lockwood et al., 2021). Crucially, this task allows me to test two aspects of prosocial motivation that are relevant to moral hypocrisy. One aspect is how willing participants are to make efforts to help others compared to self. If moral hypocrisy is in part explained by avoidance to incur in behavioural costs such as effort, it would be expected that hypocrites show less willingness to work regardless of who the beneficiary may be. In contrast, if hypocrites show low motivation only when the cost is to benefit another person, it suggests that moral hypocrisy is specific to social decisions, avoiding self-costs only when the benefits are not for oneself.

Secondly, this task also allows me to look at participants' exertion of effort. In each trial, participants need to physically invigorate the required action to then obtain the reward. There are no task-related reasons whatsoever for participants to perform differently for self and other as (i) they freely decide whether to work or not, and (ii) the required effort levels are the same for both conditions. Thus, looking at the force exerted for each trial in which participants decided to work and how successful participants were when working for self and other, I can also test whether the discrepancy between moral

judgement and actions is associated with invigorating in a less degree actions that benefit others compared with self. In doing this, I could assess whether hypocrites have a superficial interest to seem moral without paying the costs associated with that action (Batson et al., 1997; Merritt & Monin, 2012).

5.3 Methods

5.3.1 Participants

This study used the same sample of 66 participants reported in Chapter 3. However, three participants produced judgment data that was uninterpretable (i.e. delivering more shocks for less money was judged less blameworthy) and were excluded. One further participant did not finish the judgement task, and another participant was also excluded as they never chose to help others in the prosocial effort task, meaning an absence of sufficient variance of force data crucial for hypothesis testing. Therefore, 61 participants were included in the analyses reported in this chapter (age M = 22.6, SD = 3.8, 34 females).

5.3.2 Harm Aversion Task.

This task was used to measure participants' moral behaviour. Participants completed the harm aversion task while in the scanner as part of the first experimental session of a multi-stage study. In this task, participants traded-off a certain number of electric shocks against financial profit. Crucially, in half of the trials the electric shocks were delivered to another unknown person, the Receiver, while in the other half shocks were delivered to participants themselves, the Deciders. Money, on the other hand, was always obtained by the participants/deciders. In every trial, participants decided between two options: a harmful option, associated with more shocks for more profit, and a helpful option,

associated with less shocks but in exchange of less money. Details of this task can be found in Chapter 2.

5.3.3 Moral Judgement Task

To operationalise moral hypocrisy, decisions in the harm aversion task must be contrasted with moral judgements for the same actions. Thus, the moral judgement task aimed to examine how much people rate actions that they have previously performed as blameworthy when they are performed by others. At least one week after participants completed the harm aversion task, they had an experimental session where they undertook the moral judgment task (Fig. 5.1A). In this task, participants were presented with a subset of the decisions that they faced in the scanning session, namely all the trials in the *other* condition where the money was for the decider and the shocks were for the receiver (72 trials). On each trial of the moral judgment task, the harmful option (i.e., more money for the decider and more shocks for the receiver) was highlighted. Participants were asked to judge the blameworthiness of each harmful choice on a visual analogue scale ranging from “not at all blameworthy” to “extremely blameworthy”. Using this task, I could calculate each participant’s degree of moral hypocrisy, which quantifies discrepancies between their current blame judgments and the choices they made a week earlier (see below).

5.3.4 Prosocial Effort Task.

Participants completed the prosocial effort task after the moral judgement task in the same experimental session. Participants traded-off physical effort for monetary rewards in the form of credits. Crucially, in half of the trials the money was received by another unknown person, the Receiver, while in the other half profit was obtained by the participants themselves, the Deciders. In every trial, participants chose between two

options: a baseline option, associated with no exertion of effort (i.e. rest) for a low reward level of one credit, and a work offer, associated with higher amounts of reward (2-10 credits) for variable levels of effort exertion (30-70% of their maximum voluntary contraction) to obtain them. Once participants made their choice, they were required to perform the specific effort level squeezing a handle with their dominant hand. This task allowed me to measure two important aspects of motivation: prosocial effort-based decisions, and action performance for *self* and *other*. For the latter, two indexes were used: (i) the force exerted for each trial where participants decided to work, and (ii) participants' success in accomplishing the required level of physical effort to obtain the reward on offer. Details of this task can be found in Chapter 2.

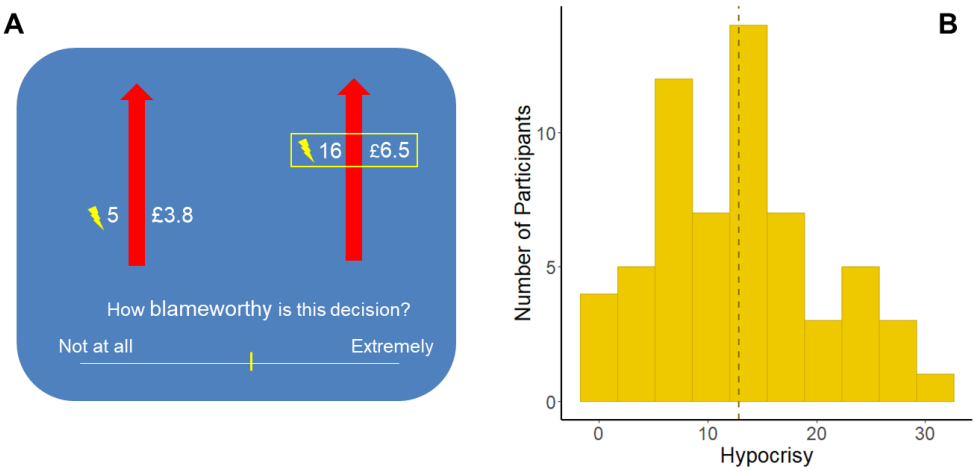


Figure 5.1. Measuring Moral Hypocrisy. *A. Moral Judgement Task*. At least one week after completing the harm aversion task, participants judged how blameworthy a series of decisions were if someone chooses the harmful options on those trials (*other* condition only, one block of 72 trials). Unbeknownst to the participants, this was the same set of trials that they had faced during the harm aversion task (Fig. 2.3B). *B. Moral hypocrisy scores follow a normal distribution*. Histogram with hypocrisy scores in the x-axis. The dash line indicates the mean.

5.3.5 Analysis of the behavioural data

5.3.5.1 Calculating Moral Hypocrisy

To calculate levels of moral hypocrisy for each participant, I estimated the discrepancy between participant's own aversion to harming others and the blame that

they assigned when others make the same actions. Because participants' actual choices were assessed in the harm aversion task, they could be compared with participants' moral judgments to others' decisions in the same set of trials. With this I could quantify each participant's degree of hypocrisy by comparing (i) their own likelihood of making a harmful decision on each trial of the harm aversion task with (ii) the blame that they assigned to others for a harmful decision on each trial of the moral judgment task. Specifically, for (i), I computed each participant's likelihood of harming another person on a given trial with the computational model described in detail in Chapter 2 and that has been used extensively in previous work (Crockett et al., 2014, 2015, 2017). In this model, the probability of choosing to inflict pain for profit is a softmax transformation of the subjective value of the harmful relative to the helpful option, ΔV. For participant *j* and trial *i* then:

$$\Delta V_{j(i)} = \left(1 - \kappa_j\right) \Delta m_{j(i)} - \kappa_j \, \Delta s_{j(i)}$$

$$P(harm)_{j(i)} = \frac{1}{1 + exp^{-\beta_j \, \Delta V_{j(i)}}}$$

Here Δm and Δs denote the additional money and shocks associated with the harmful relative to the helpful option. The harm aversion parameter, κ, indicates the relative weight participants place on shocks over money in the valuation process. κ takes on different values in *self* and *other* conditions, reflecting different decision preferences when harming oneself versus harming another person. I used a softmax function to convert ΔV into the probability of choosing the harmful option over the helpful option. Here, β is a participant-specific inverse temperature parameter that characterises the steepness of the softmax curve. To calculate hypocrisy for each subject, I focused here solely on participants' decisions to harm others for profit.

After computing each participant's likelihood of choosing the harmful option for each trial in the decision-making task, I computed a moral hypocrisy score for each participant by combining the data from the moral judgment task with the data from the harm aversion

task. I computed hypocrisy by summing across trials the amount of blame assigned on each trial of the judgment task, weighted by the participants' own likelihoods of choosing selfishly on the same trial in the decision task, such that:

$$Hypocrisy_j = \sum_i blame_{j(i)} * P(harm)_{j(i)}$$

Here, $blame_{j(i)}$ denotes participant *j*'s blameworthiness judgment on trial *i* of the judgment task. $P(harm)_{j(i)}$ is the likelihood that participant *j* chooses the harmful option on the same trial in the moral decision-making task derived from the softmax function. Weighting the blame judgments by the choice likelihoods captures the logic that it is more hypocritical to blame someone for an action that oneself has previously taken confidently than tentatively or accidentally (Alicke et al., 2013; Dover, 2019; Laurent & Clark, 2019; Wallace, 2010).

5.3.5.2 Testing the relationship between moral hypocrisy and prosocial motivation

In this study, I used an analytical approach that allows me to examine how moral hypocrisy could interact separately either with effort or reward on predicting decisions to work. If moral hypocrisy is indeed associated with motivation, disentangling the effects of reward and effort can give more detailed information about the motivational profile of hypocrites. Moral hypocrisy could be related to effort costs but not reward, suggesting that hypocrites are more effort averse. In contrast, it could also be that hypocrisy modulates reward but not effort, suggesting that hypocrites are less incentivised by the benefit. Using mixed-effects models (MMs) with trial-by-trial decisions as dependant variable, I could test potential effects of hypocrisy on effort and reward separately. For this reason, computational parameters λs were not used in this study (see Chapter 2) as they consider rewards in interaction with effort, without disentangling the effects of the two. Thus, the following MMs were built using the *glmer* function in R:

Simple model

$$DW_i = \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 R_i B_i + \beta_5 E_i B_i + \beta_6 E_i R_i + \beta_7 E_i R_i B_i$$

Hypocrisy Model

$$DW_i = \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 H + \beta_5 R_i E_i + \beta_6 R_i B_i + \beta_7 R_i H + \beta_8 E_i B_i + \beta_9 E_i H + \beta_{10} B_i H + \beta_{11} R_i E_i B_i + \beta_{12} R_i B_i H + \beta_{13} E_i B_i H + \beta_{14} R_i E_i H + \beta_{15} R_i E_i B_i H$$

In the simple model, decision to work *DW* in the trial *i* is predicted by the fixed effects of reward *R*, effort *E*, beneficiary *B*, and their interactions. DW is a binary, factor variable in this logistic MM. The hypocrisy model adds the hypocrisy index *H* as an independent variable together with its interaction with all the other variables. Random intercepts were clustered in subjects *j*, and random slopes on effort and reward were included as it is expected that participants vary in their sensitivities to this information. Furthermore, AIC values confirmed that using random slopes in these models improved model fitting (see Chapter 3). Crucially, AICs for each model were used to test whether the addition of hypocrisy improved the model. This was complemented with a loglikelihood test using the *anova* function in R to test whether differences in model fitting were significant, given that the simple and the hypocrisy models were nested.

To test the influence of hypocrisy on performance, I used similar MMs to the ones described above, but instead of using DW as dependant variable, I used the force exerted for each trial in which participants chose to work, using the *lmer* function in R. Force was normalised per each trial as a proportion of participant's maximum to account for between-subject variability in the force exerted and calculated the area under the curve (AUC) for the three seconds window in which they exerted force. Thus, normalised force was a continuous variable in these subject-level random-intercept models. Trials where participants chose to rest were excluded from analysis. Unlike the decision models,

random slopes for effort and reward were not used because (i) participants had different number of trials per each different condition, and (ii) high individual variability in the force exerted between levels of effort and reward was not expected - the force required to obtain the reward on offer was set for each participant's specific maximum voluntary contraction and showed limited variability in existing data (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021). Furthermore, interactions between effort x reward x beneficiary were not included, as participants, regardless of hypocrisy, chose to work for others compared to self significantly less often, especially in low-reward/high-effort trials, meaning that a potential 3-way interaction would be unstable because of the small trial sampling at certain combinations of effort and reward.

Finally, the effects of moral hypocrisy on performance were also tested with a model that predicted whether each trial was succeeded or failed, i.e. whether people achieved the specific effort requirement for each work trial and therefore whether they obtained the reward on offer. Here, success was a binary factor variable in a logistic random-intercept MM which did not have random slopes because of the same reasons given above. Notice that even though models predicting success can give important insights about the motivational profile of hypocrites, these results must be taken cautiously, as the success rates in this task are very high regardless the condition (the task was designed with that intention – see Chapter 2 for details), and therefore, the effects of hypocrisy on trial success could be triggered by mainly a few participants/trials.

5.4 Results

5.4.1 Moral hypocrisy is common and varies across participants

As outlined in previous chapters, participants were more averse to harm others relative to themselves for profit in the harm aversion task. To calculate moral hypocrisy, only harm aversion for others was considered. Moral hypocrisy was defined as the

discrepancy between the probability to harm in each trial and how harsh participants blame others for the same action. Figures 5.2 illustrates what the hypocrisy index reflects. Participants who assign a high level of blame on trials where they themselves are likely to choose the harmful option have a high hypocrisy score based on the model described above (Fig. 5.2A-B). In contrast, participants who almost never assign blame on the trials where they themselves are likely to choose the harmful option have a low hypocrisy score (Fig. 5.2C-D). With this operationalisation of moral hypocrisy, 97% of participants displayed at least some level of hypocrisy, with a wide range of individual variation in the degree in which it is manifested, following a normal distribution (M±SD = 12.9±7.5; Kolmogorov-Smirnov normality test, Fig. 5.1B). Since moral hypocrisy was operationalised using *value* from the computational model described in Chapter 2, it is not surprising that this index is negatively correlated with κ_{other} (Spearman $\rho = -0.57$, $p < 0.001$). Therefore, for the following models where moral hypocrisy predicts motivation, I controlled for κ_{other} adding it as a variable of no interest (i.e. hypocrisy model + κ_{other}).

5.4.2 Moral hypocrisy is associated with prosocial apathy

First, I tested whether adding the moral hypocrisy index as independent variable improved model fitting in predicting decisions to work. Two models were compared: one where effort, reward, beneficiary and their interactions were included as independent variables predicting decisions to work (simple model), and another model where hypocrisy was added as a main effect, together with its interactions with the other variables (hypocrisy model). Model comparisons revealed that adding hypocrisy improved model fitting (AIC: simple model= 5457.7; hypocrisy model = 5446.8, Fig. 5.3A). A loglikelihood ratio test was performed to assess whether the more complex, hypocrisy model was significantly better than the simple model. This analysis confirmed that, indeed, the hypocrisy model significantly improved the model fit ($\chi^2_{diff} = 28.93$, $df_{diff} = 9$, $p < 0.001$) suggesting that moral hypocrisy is important to predict decisions to work.

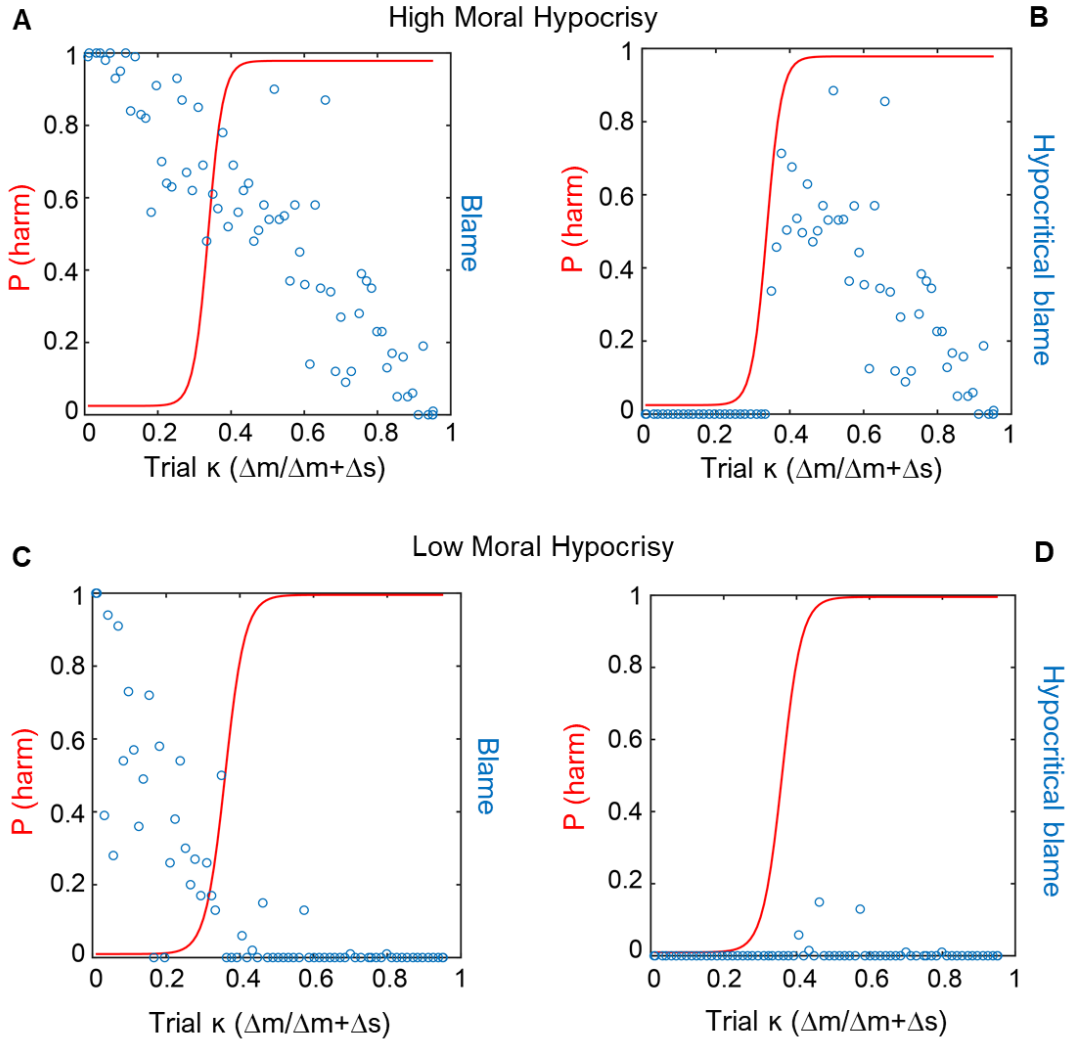


Figure 5.2. Samples of two participants with different hypocrisy scores but similar harm aversion. A-B corresponds to a high hypocrite (hypocrisy = 16.8), while C-D depicts a low hypocritical participant (hypocrisy = 0.7). Both participants have comparable harm aversion parameters in the *other* condition (κ_{other} = 0.39 and κ_{other} = 0.37, respectively). The x-axis indicates the κ value [Δm/(Δm + Δs)] of each trial, where higher trial κ value meaning that choosing the more harmful option confer a large amount of profit for the decider by increasing the receiver's harm by a small degree. The y-axes on the left indicate the model-derived probability of choosing the harmful option for each participant. The red curves are the best-fit softmax curves. In A and C, the blue dots indicate participants' blameworthiness judgment ratings. In B and D, the blue dots represent the product of blameworthiness judgment and the model-derived participant-specific probability of choosing the more harmful option on the same trial (weighted blame). Participant's moral hypocrisy index is the sum of weighted blame across all the trials. As it is apparent, the participant described in B was more hypocritical by this operationalisation than the participant described in D, because participant in A indicated substantial degree of blameworthiness judgment on the trials where they were very likely to choose the harmful option themselves.

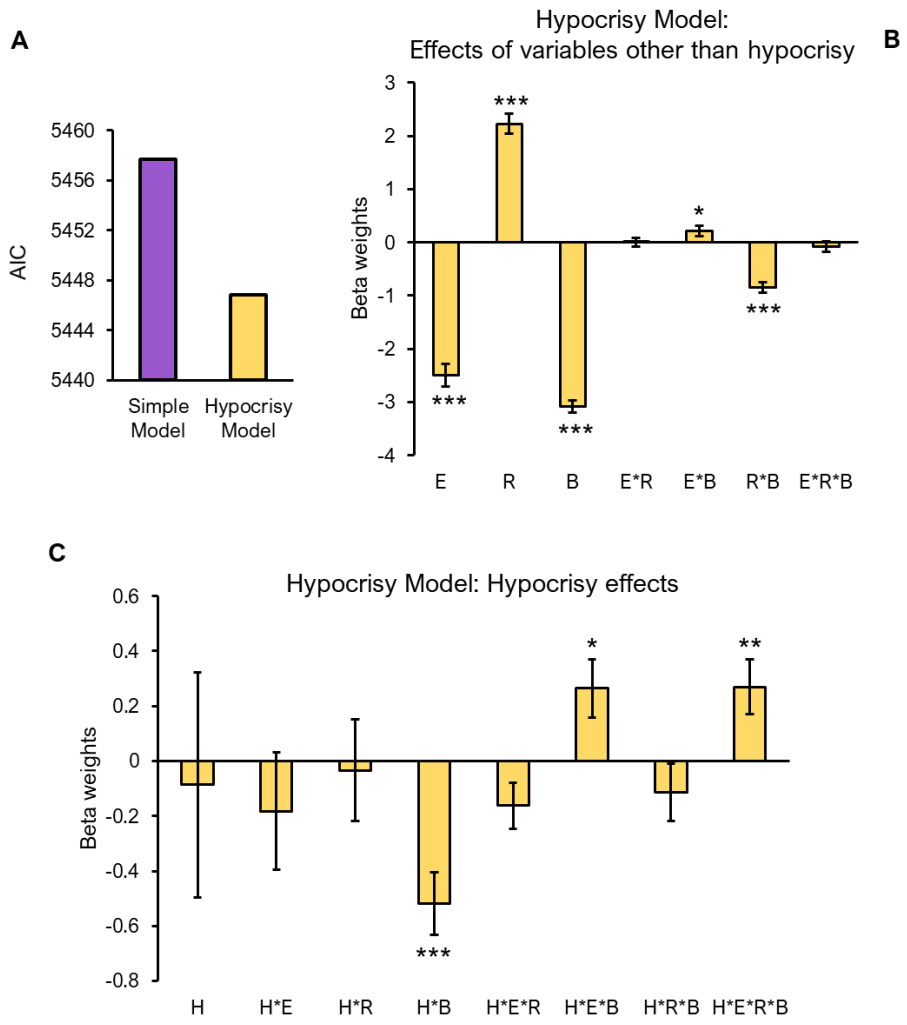


Figure 5.3. Moral hypocrisy influences effort-based decisions. *A. Moral hypocrisy improves model fitting*. Adding hypocrisy index (H) as an independent variable predicting decisions to work improves model fitting according to AIC values (Akaike Information Criterion, y-axis). The simple model only includes effort (E), reward (R) and beneficiary (B) as independent variables. *B. Previous results are replicated in a model having hypocrisy as an independent variable.* Effects not related to hypocrisy are consistent in the hypocrisy model compared with simple models used in previous work. Thus, main effects of E, R and B are significant, together with E*B and R*B interactions. Beta values of each effect is depicted in the y-axis. *C. Effects of hypocrisy on effort-based decisions.* In the same hypocrisy model, hypocrisy shows significant interactions with beneficiary, with effort and beneficiary, and with effort, reward and beneficiary in predicting decisions to work. These results suggest that hypocrisy modulation on motivation to work varies according to who is getting the reward on offer. Error bars depict SEM. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

In the hypocrisy model, a four-way interaction was found between hypocrisy x effort x reward x beneficiary ($b = 0.27$, $SEM = 0.1$, $z = 2.7$, $p < 0.007$; Fig. 5.3C). Thus, while there was not much difference between high and low hypocrisy for the *self* condition in decisions to work (Fig. 5.4A), people who were more hypocritical were less willing to work

for others across almost all reward and effort levels (Fig. 5.4B). Comparing motivation for self vs other in high vs low hypocrisy (Fig. 5.4C), it is apparent that, in general and across effort/reward levels, people who have higher hypocrisy scores have higher differences in their motivation for *self* and *other* compared to those with low scores, who seem to behave more similarly across beneficiaries. These effects seem to be greater in the 'easy offers', i.e. low effort for high rewards. These results suggest that hypocrites are less prosocially motivated in general, and especially in easy trials. However, the latter must be taken cautiously as this effect may be driven mainly by ceiling effects in the *self* condition, where participants, regardless hypocrisy and consistent with previous studies, chose almost always to work for themselves.

Besides this interaction, and consistent with the above results, the hypocrisy index also showed a significant three-way interaction with beneficiary and effort ($b = 0.27$, $SEM = 0.11$, $z = 2.5$, $p < 0.02$, Fig. 5.3C) and a two-way interaction with beneficiary ($b = -0.52$, $SEM = 0.11$, $z = -4.54$, $p < 0.0001$, Fig. 5.3C), but not a main effect by its own. These results support the four-way interaction described above - as participants scored higher in moral hypocrisy, they worked less for others compared with self (Fig. 5.5A), even if the level of effort required to obtain the reward were low (Fig. 5.5B and 5.5C). Finally, and as a control result, the hypocrisy model replicated the general effects revealed by models that do not include hypocrisy as independent variable, showing that participants, regardless of their hypocrisy levels, showed less willingness to work for others than for self, especially in high effort trials, and in trials with low rewards on offer (Fig. 5.3B; Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021, Chapters 3 and 4).

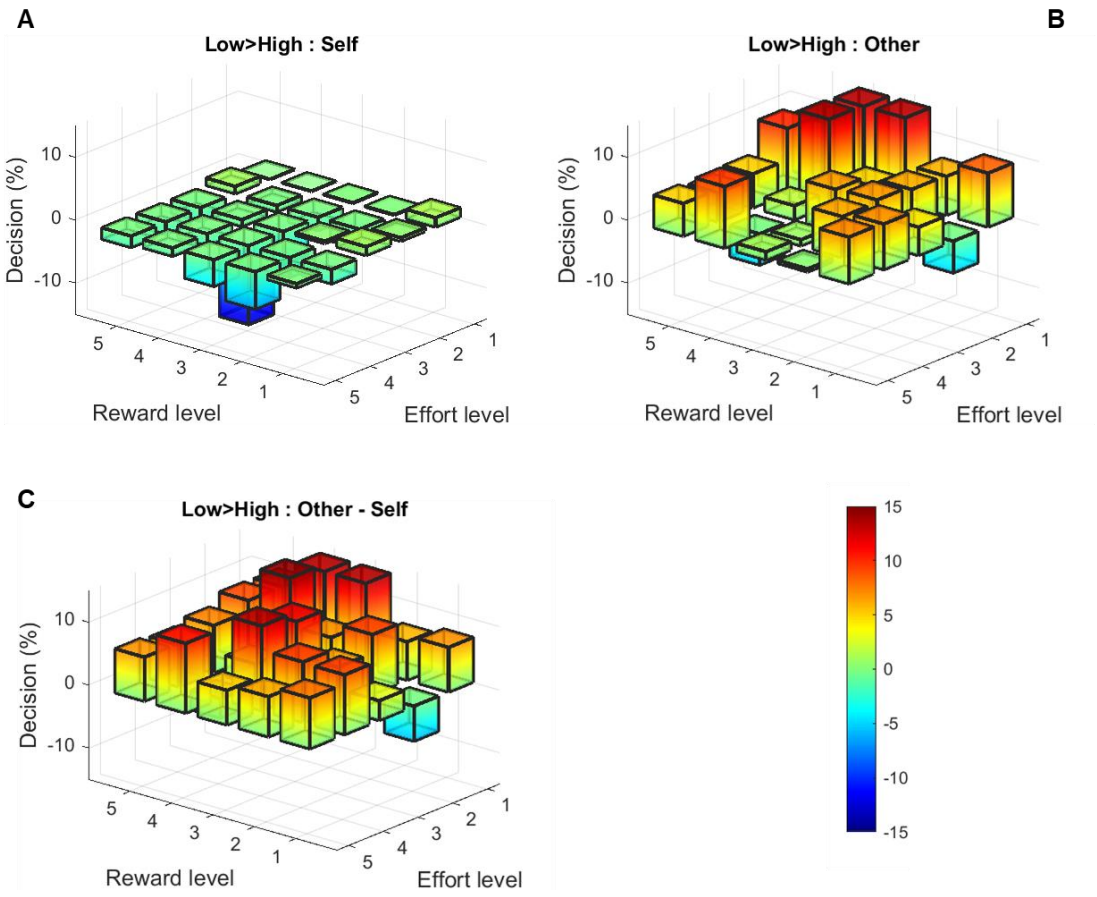


Figure 5.4. Moral hypocrisy interacts with effort, reward and beneficiary. *A. 3D plot of choices to work vs. rest across different effort and reward levels in low compared to high hypocritical participants for the self condition*. Participants who have high and low scores in hypocrisy (median split) show similar patterns of decisions in self trials across different effort and reward levels. *B. 3D plot of choices to work vs. rest across different effort and reward levels in low compared to high hypocritical participants for the other condition.* Participants who score low in hypocrisy are more motivated to work for others compared with high hypocrites in almost all the spectrum of effort and reward combinations. This seems especially true in 'easy trials', where high reward is obtained by low effort. *C. Differences between low and high hypocrisy individuals in choices to work for other vs self.* Participants who score low in hypocrisy are more similarly motivated to work for self and others compared with high hypocrites in almost all the spectrum of effort and reward combinations. People with a high hypocrisy score seems to me more biased to work for self than other, especially in trials where high reward is obtained by low effort. *Note*: Median split dividing participants in high and low hypocrites was only used for illustrative purposes. All statistical analyses used hypocrisy as a continuous variable.

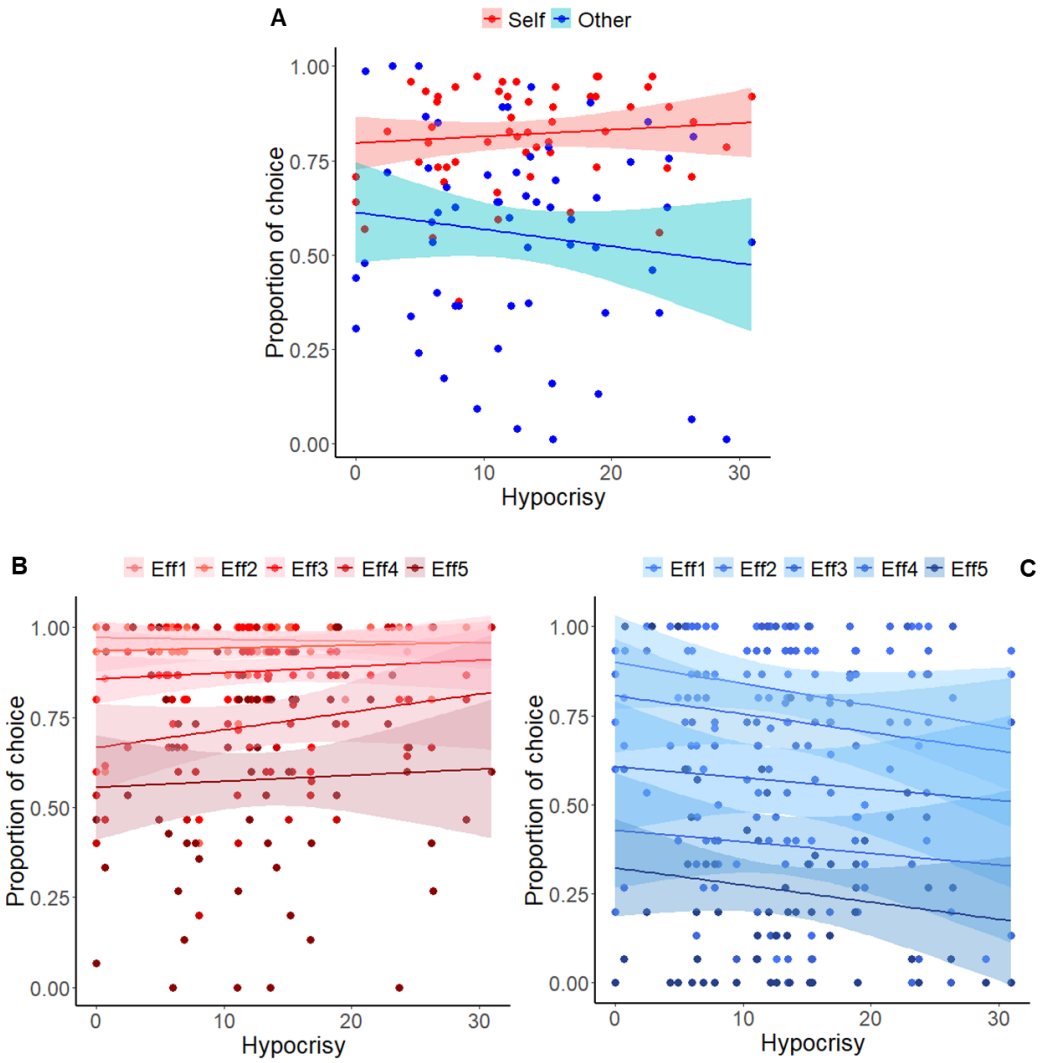


Figure 5.5. Moral hypocrisy is associated with less willingness to work for others compared to self. *A. Moral hypocrisy is associated with less motivation to work for others*. The hypocrisy model revealed a hypocrisy*beneficiary interaction, suggesting that people who are more hypocritical are less willing to work for others compared to self. This effect seems to be specifically modulating behaviour in *other* trials. Y-axis shows proportion of choosing the work vs the rest option. *B-C. Effects of hypocrisy on decisions according to effort and beneficiary.* Hypocrisy seems to be generally associated with less motivation for others across effort levels, including the lowest ones (C). Furthermore, hypocrisy seems not to modulate decisions in the *self* condition (B), although a trend of more motivation for *self* in higher hypocrites seems to occur in higher levels of effort. Dots depict each participant.

5.4.3 Moral hypocrisy is associated with trial performance.

After establishing an association between hypocrisy and prosocial decisions, I tested whether a similar relationship could be found in how much force participants exerted into actions when helping others. In principle, there should not be differences between *self*

and *other* as participants have the alternative of resting for a low reward, and as each effort level is adjusted to a percentage of each participant's MVC, which ensure a high level of success. Thus, if participants choose to work for a specific amount of effort level, they should exert, on average, the same amount of force in that effort level regardless of whether the reward is for themselves or not. However, participants in fact generally exert less force into the same actions when they benefit others than self (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021). This 'superficiality' in *other* trials performance could be associated with discrepancy between moral judgement and decisions, with moral hypocrites tending to invigorate in a lesser degree actions that do not bring self-benefits, such as prosocial effort.

To test this hypothesis, I first compared two models to see whether including the hypocrisy index as an independent variable improved model fitting. Similar to what is described in the analysis above, a simple model was built where effort, reward, beneficiary and their interactions were predicting normalised force (AUC) trial-by-trial. A hypocrisy model was compared to the simple one, in which the hypocrisy index was added as a main effect and in interaction with the variables in the simple model. Comparing the AIC values for each model, it is apparent that the hypocrisy model outperformed the simple one (simple model = -17179, hypocrisy model = -17193, Fig. 5.6A). This was later confirmed by a loglikelihood test that revealed a significant model fitting improvement by the hypocrisy model ($\chi^2_{diff} = 27.59$, $df_{diff} = 7$, $p < 0.001$). Crucially, the hypocrisy model revealed an interaction between hypocrisy and beneficiary ($b = -0.01$, $SEM = 0.002$, $t = -3.9$, $p < 0.001$; Fig. 5.6B), indicating that as participants have higher hypocrisy scores, the difference in force exertion between self and other increased (Fig. 5.6C), such that they exerted less force for actions that benefitted others relative to self than participants low in hypocrisy. Thus, the discrepancy between moral judgements and moral actions are associated with higher gaps between *self* and *other* action invigoration. Furthermore, this model revealed main effects of effort (high effort levels

require more force exertion), reward (participants exerted more force when the reward was high), and beneficiary, together with an effort x beneficiary interaction (Fig. 5.7A), showing that participants generally exerted less force for others than for self, especially in high effort trials, replicating previous studies (Fig. 5.7B).

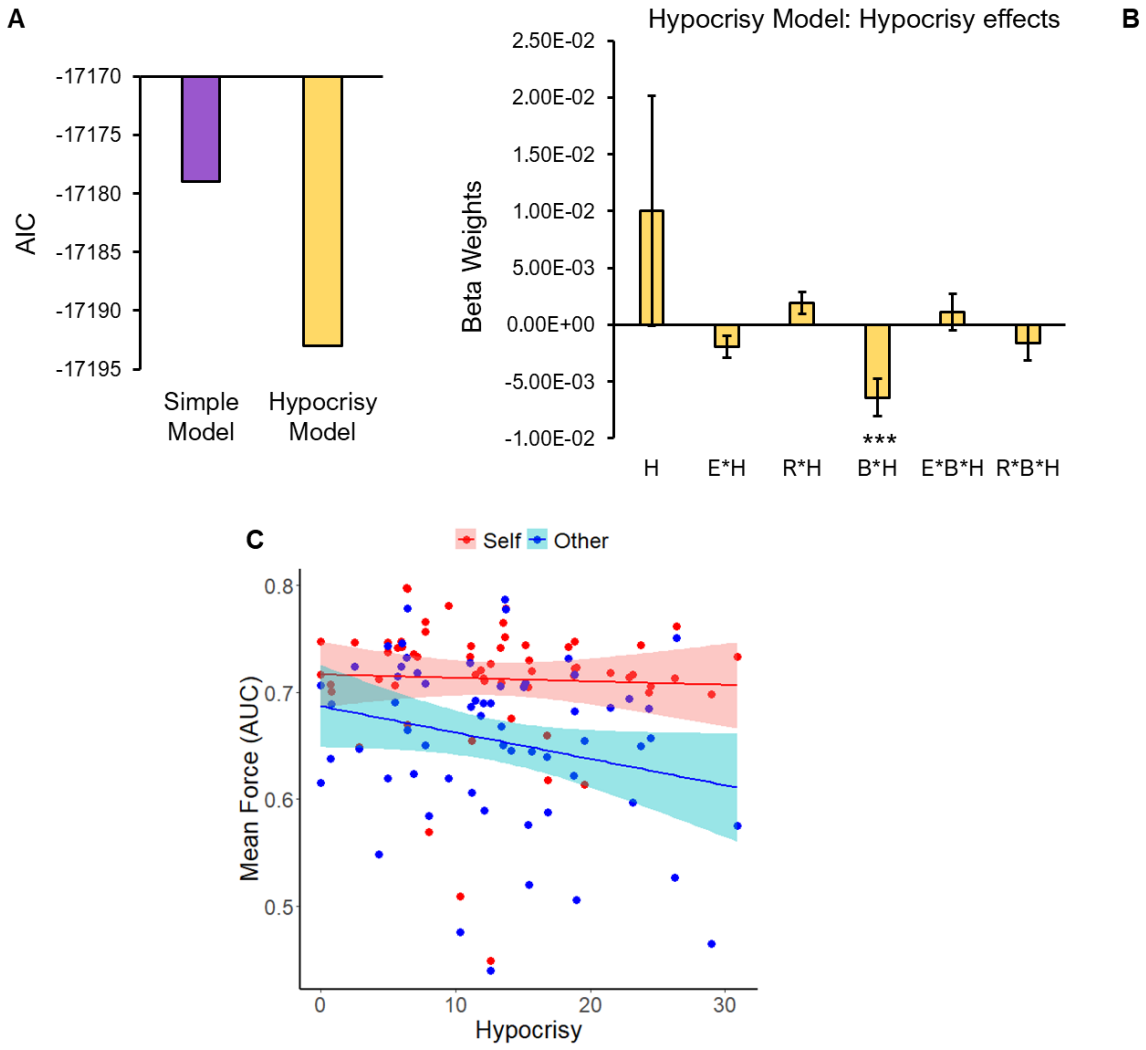


Figure 5.6. Moral hypocrisy modulates force exertion. *A. Moral hypocrisy improves model fitting predicting force exertion*. Adding hypocrisy index (H) as an independent variable predicting force exertion outperformed a simpler model that only includes effort (E), reward (R) and beneficiary (B) as independent variables according to AIC values (Akaike Information Criterion, y-axis). *B. Moral hypocrisy interacts with beneficiary on predicting force*. Moral hypocrisy modulates how much force is exerted for self and other across trials. Y-axis depicts the beta values for the effects found in the hypocrisy model. *C. Moral hypocrisy is associated with less force exerted for others compared with self.* Participants who scored high in hypocrisy showed more differences in how they exerted force for others compared to self when similar effort was required, suggesting more superficiality in their prosocial actions. Y-axis depicts the mean area under the curve (AUC) during the 3s force period normalised to participants maximum level of force exerted across trials. Dots corresponds to each participant. Error bars correspond to SEM. *** $p < 0.001$

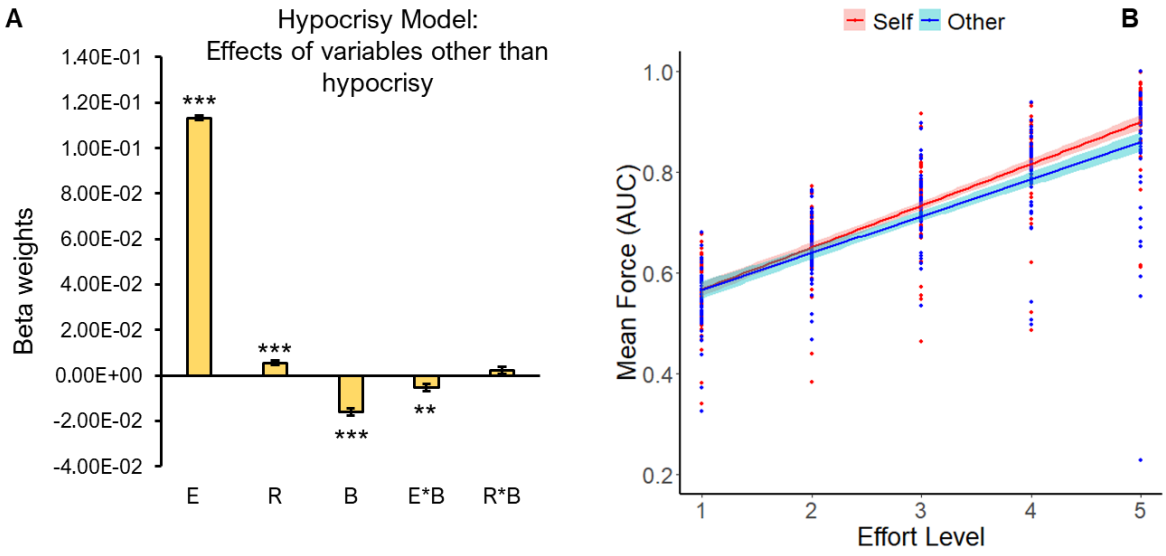


Figure 5.7. The hypocrisy model reveals that people exerted less force when actions benefitted others than self. *A. Participants exerted less force for other than self trials.* The hypocrisy model shows a main effect of beneficiary, and a beneficiary*effort interaction, suggesting that people exerted less force when the beneficiary was another person relative to self. Y-axis depicts the beta values for the effects found in the hypocrisy model. *B. Participants showed higher self-bias in force exertion in high effort trials.* Effort and beneficiary interaction indicates that participants exerted less force for others compared to self especially in high effort trials. Y-axis corresponds to the mean area under the curve (AUC) during the 3s force period across effort levels normalised to participants maximum level of force exerted across trials. Error bars correspond to SEM. *** $p < 0.001$, ** $p < 0.01$.

Since participants, regardless of their hypocrisy score, chose to benefit others much more when the effort required was low, these effects could be accounted for by a reduced sampling of trials for higher level of effort in the *other* condition. Random-intercepts in the MMs can address some of the problems associated with the difference in trial sampling, as data points (trials) are clustered for each participant. However, I performed further analyses to confirm the robustness of the effect of hypocrisy on force exertion. Importantly, given that the effect of hypocrisy on force exertion is in interaction with beneficiary *across* effort levels, it should remain even if higher effort levels are removed from the model, having in this way more similar number of trials per condition. Consequently, I removed consecutively the three highest effort levels from the model testing each time whether the effect was still present. The hypocrisy x beneficiary interaction proved to be a robust effect and it was present even if the effort level 5 (t = -

4.1, $p < 0.001$), the effort levels 4-5 ($t = -2.78$, $p < 0.01$), and the effort levels 3-5 were removed ($t = -2.17$, $p < 0.05$). Thus, for people who show high levels of hypocrisy, higher gaps in force exertion between actions that benefitted self and other were found across all effort levels.

5.4.3.1 Moral hypocrisy predicts success in prosocial actions

As effort levels are set as a percentage of participants' MVC in the prosocial effort task, all effort levels should be attainable, and failure indicates a reduction in the motivation to exert the required effort even when they freely chose to undertake it. In fact, success rates were very high in the prosocial effort task. Participants on average obtained the reward on offer in 98.3% of the *self* trials (SEM = 0.5%), and in 96.9% of the *other* trials (SEM = 0.8%). Is hypocrisy related to failure to succeed at the required effort level? The above results on force exertion suggests that this could be the case. In order to test whether moral hypocrisy modulated the probability to succeed in a given trial, two different models were built using the same approach described above, i.e. a simple model against a hypocrisy model, but now predicting the binary variable 'success'. Note, however, that the results revealed in this analysis could be driven only by a number of participants in only a number of trials due to the high success rates of this task. Results described in this section should be therefore interpreted with caution.

The hypocrisy model improved model fitting relative to the simple model according to the AIC (simple model = 953.7, hypocrisy model = 949.1; Fig. 5.8A). A loglikelihood test confirmed that the hypocrisy model significantly outperformed the simple model ($\chi^2_{diff} = 18.61$, $df_{diff} = 7$, $p < 0.01$). Crucially, this model revealed that moral hypocrisy interacted with beneficiary (hypocrisy x beneficiary interaction, $b = -0.76$, SEM = 0.3, $z = -2.48$, $p < 0.02$, Fig. 5.8B), such that people who scored high in hypocrisy failed more in *other* trials compared with people who scored low, showing bigger differences in success between *self* and *other* (Fig. 5.8C). Moral hypocrisy did not significantly modulate success in *self*

trials. These results indicate that participants with high moral hypocrisy not only exerted less force for others than for themselves for similar actions, but also they did so to the extent that more often failed to energise the required force to obtain the reward on offer. Finally, the hypocrisy model revealed main effects of effort, indicating that people failed more as trials get harder, and beneficiary showing that people failed more in *other* trials compared to *self* trials (Fig. 5.8D).

Given that moral hypocrisy modulated success in *other* trials, it could be argued that its effect on force exertion might have been driven only by those trials where participants exerted such a low amount of force that they did not achieve the required effort goal. However, this seems unlikely, since the effect of hypocrisy on force remained even if highest effort trials (levels 3-5), which are more likely for participants to fail (Fig. 5.8D), are removed from the model. Despite this, and to add robustness to these results, I performed a final analysis to test the effects of hypocrisy on normalised force building again a simple and a hypocrisy model on trial-by-trial force exertion but this time considering only trials where participants succeeded. Again, the hypocrisy model significantly improved model fitting relative to the simple one (AIC: simple model = -18858, hypocrisy model = -18865, $\chi^2_{diff} = 20.39$, $df_{diff} = 7$, $p < 0.005$). Crucially, hypocrisy x beneficiary interaction was maintained ($b = -0.003$, SEM = 0.001, $t = -2.18$, $p < 0.03$), indicating that people who scored high in moral hypocrisy exerted less force for others across trials. A three-way interaction between hypocrisy x beneficiary x effort was also significant in this model with only successful trials ($b = 0.004$, SEM = 0.001, $t = 2.64$, $p < 0.01$), suggesting that the modulation of hypocrisy on the force exerted for *self* and *other* varied across different effort levels. This effect is likely to be triggered by the removal of the failed trials from the analysis, which corresponds to the manifestation of hypocrites' low motivation in force exertion especially at the higher effort levels. Thus, while hypocrisy effects on lower effort levels remained in this model, its effects in the harder ones were diminished by eliminating these failed trials.

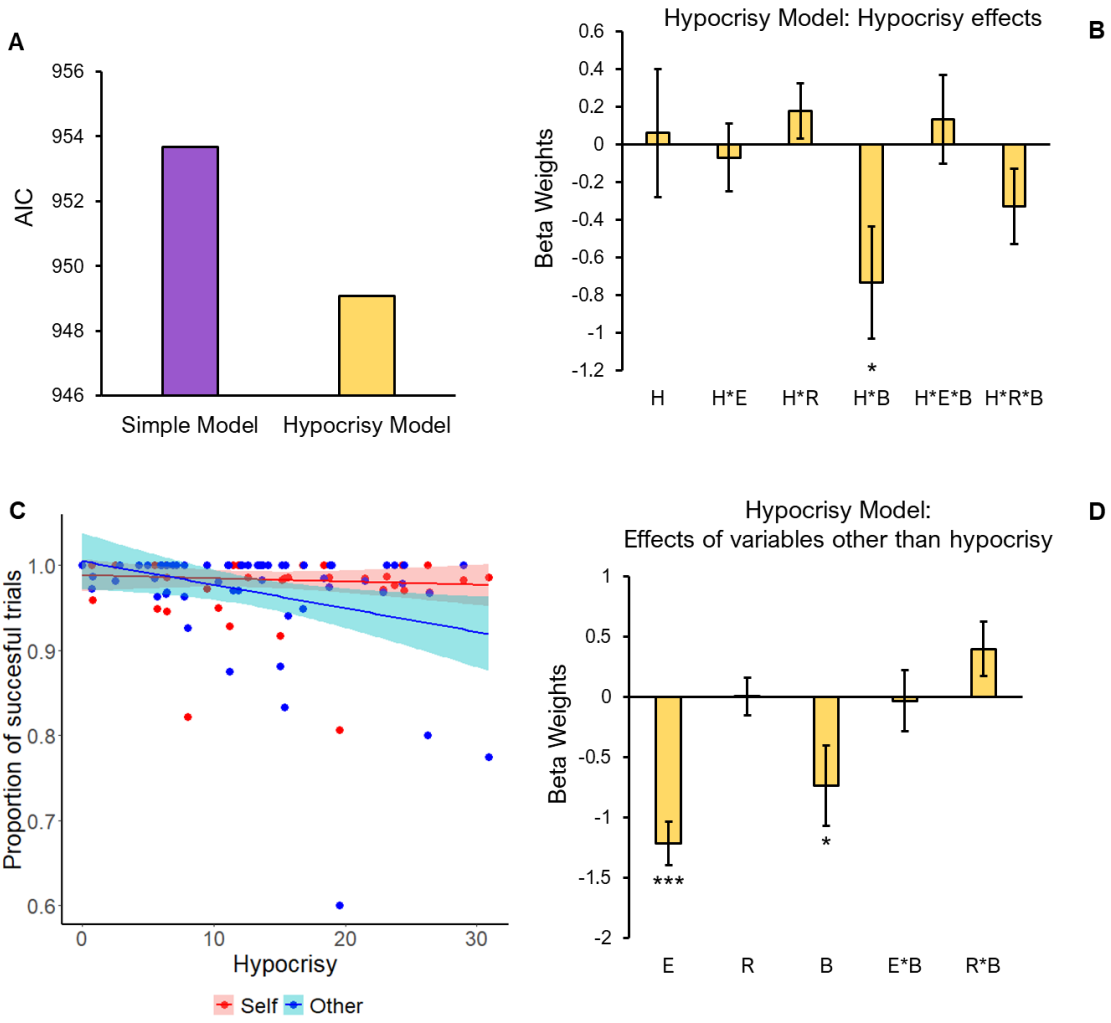


Figure 5.8. Moral hypocrisy influences success in working for others' benefit compared to self. *A. Moral hypocrisy improves model fitting predicting trial success*. Adding hypocrisy index (H) as an independent variable predicting trial success outperformed a simpler model that only includes effort (E), reward (R) and beneficiary (B) as independent variables according to AIC values (Akaike Information Criterion, y-axis). *B. Hypocrisy effects on trial success.* Hypocrisy interacted with beneficiary when predicting the success of the failure of a trial. Y-axis depicts the beta values for the effects found in the hypocrisy model *C. Moral hypocrisy is associated with less success in trials benefiting others.* As people score higher in moral hypocrisy, they fail more in trials where the benefit is received by others compared with themselves. Y-axis depicts the proportion of successful vs failed trials *D. Participants failed more in other trials compared with self in the hypocrisy model.* The model revealed a significant main effect of beneficiary. Furthermore, it revealed a main effect effort, indicating less success in trials with higher effort levels. Y-axis depicts the beta values for the effects found in the hypocrisy model. Error bars correspond to SEM. *** $p < 0.001$, * $p < 0.05$.

5.5 Discussion

Every day people blame others for actions that they judge as morally wrong, denoting in this way their moral standards. Commonly, though, those who blame others experience similar shortcomings to those that they judge, a discrepancy that has been identified as a type of moral hypocrisy. One of the reasons that could underlie this phenomenon is a reduced motivation to overcome the costs of being moral (Batson & Thompson, 2001). Here, I formally investigated this hypothesis using computational modelling together with decision-making tasks, to measure moral hypocrisy and prosocial effort motivation, testing for their association. This approach revealed two novel aspects on the motivational profile of moral hypocrites. Participants with higher levels of hypocrisy, i.e. wider gaps between moral judgements and actions, were (i) less willing to put in effort to benefit others relative to self, and (ii) exerted less force when performing for others the same actions that benefitted themselves. These results suggest that moral hypocrisy might be related to a dissociation between intentions to be moral, manifested in blame judgements, and the motivation to overcome the costs to act accordingly.

The current study revealed that moral hypocrisy is associated with less willingness to incur effort costs specifically for prosocial actions, but not for self-benefitting acts. This could be due to a conflict between self and other interests in moral hypocrites, as suggested by previous work (Tang et al., 2017, 2018). Putting in effort to benefit oneself is not a moral dilemma or a moral problem - it is simply how motivated one is to work, and therefore, it is unaffected by hypocritical tendencies. However, prosocial effort is morally relevant. Doing something to benefit someone else requires a personal cost. People high in hypocrisy appear to be unable to overcome this personal cost. This is consistent with evidence suggesting that low hypocritical people might be more self-driven by the welfare of others, valuing more equally self and other benefits (Tang et al., 2018), in contrast with people who act morally only to conform to the social norm (Dong et al., 2019; Graham et al., 2015; Lönnqvist, Irlenbusch, & Walkowitz, 2014).

People who internalise the welfare of others, according to this view, would be more likely to succeed in living up to moral standards, consistent with evidence showing that people scoring high on moral responsibility are not necessarily less hypocritical, while people who score high in self-motivated moral intentions display more moral integrity (alignment between moral judgements/actions; Batson et al., 1997, 2002; Dong et al., 2019; Graham et al., 2015; Lönnqvist et al., 2014; Tang et al., 2018). The current results support such claims but show that what is argued may be partly due to a lack of willingness to exert effort to obtain benefits for others.

Strikingly, moral hypocrites were less willing to work prosocially especially in 'easy' trials – i.e. where the effort cost was low and the benefit high. At first, this result seems counterintuitive, since it is in the high effort levels where people in general show more reluctance to incur in prosocial effort (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021), and moral hypocrisy could have just augmented this effect. However, people have consistently shown high motivation to work in easy trials regardless of the beneficiary, a ceiling effect that could underlie the current results. Thus, as hypocrites *also* display low motivation in these trials, the difference in this spectrum of the design is more pronounced compared with low hypocrisy participants. Hypocrites might reduce self-costs, even if those costs are small, especially in non-normative contexts like the prosocial effort task - i.e. not helping others is not as socially prohibitive as harming others for profit (Carnes & Janoff-Bulman, 2012; Decety & Cowell, 2017; Gert, 2004; Janoff-Bulman et al., 2009). The motivational aspects related to moral judgement-action discrepancy is likely to be amplified in a non-normative context, with self-interest (avoiding effort) becoming more salient (Lindenberg et al., 2018), and where there is not a clear norm to comply with.

Self-interest in hypocrisy could also be associated with the potential role that blaming others could have in social perception. Blaming others for wrongdoing has a social impact – it increases social image and esteem (Lammers et al., 2010; Szabados & Soifer, 2004).

Thus, hypocrites' interest towards others might be superficial – hypocrites seem moral through their judgements, but they do not commit to the costs of acting accordingly. When faced with the opportunity to act selfishly, they will do so, although in appearance they would maintain their (good) moral status (Batson et al., 1997, 2002). Intriguingly, moral hypocrisy in the current study was not only associated with less willingness to choose to work for others, but also, when they decided to do so, they energised less the prosocial actions compared with self-benefiting behaviour ones. This happened across all effort levels, sometimes failing to achieve the required force to obtain the money for others due to their poorer performance. These results suggest that prosocial motivation in hypocrites is not only reduced but is also superficial, which would be in line with a superficial morality account of hypocrisy.

Even though the latter results, together with a superficial morality framework of hypocrites, can be interpreted as supporting a deceptive perspective of moral hypocrisy, this does not necessarily have to be the case (Bartel, 2019; Dover, 2019; Mele, 1989). First, participants, regardless of hypocrisy, decided to help others in a high number of trials. Second, in the vast majority of the trials, participants, regardless of hypocrisy, achieved the force required to obtain the reward on offer. Therefore, it seems more appropriate to consider moral hypocrisy as the extreme of a continuum, associated to motivational problems with overcoming self-costs to benefit others. Here, hypocrisy was operationalised as a continuous variable (Yu et al., 2019), with almost all participants (59 out of 61) showing some degree of moral judgement/decision discrepancy, consistent with previous findings (Batson et al., 1997, 2002). People with higher levels of hypocrisy still worked for others, only that they did so to a lesser degree than the lower extreme, as they are more sensitive to their self-interest when benefitting others. Whether their moral standards are genuine or deceptive should be addressed directly and empirically in future studies.

Research on goal-directed behaviour can offer an explanatory framework for interpreting the current results. This literature has identified two components in reward processing (Berridge, 1996; Berridge, Robinson, & Aldridge, 2009) - a hedonist impact of pleasure given by the reward, a 'liking' component; and the incentive saliency of the reward, a 'wanting' component, that drives the agent towards a reward-seeking behaviour manifested in attention-grabbing motivational features. Even though these two components are related to each other, they can, in principle, be dissociated. For instance, a person could frequently perform actions that they do not enjoy (e.g. habits or compulsive behaviours) or not seek behaviours that bring rewarding experiences (e.g. being effort averse such as in apathy).The current results could be interpreted as hypocrites 'liking' the idea of being moral – they share moral principles that manifest in their judgements toward others –, but they fail in 'wanting' to be moral, and thus do not overcome self-costs to benefit others. Philosophical perspectives on hypocrisy have claimed that this form of hypocrisy could trigger guilt and frustration in the hypocrite due to this failure (Bartel, 2019; Dover, 2019; Gilead, 1999; Mele, 1989; Polman & Ruttan, 2012). Future studies could directly test the viability of these hypotheses, designing experiments that test the predictions from this framework.

Future research could extend this work and control for some limitations of the current study. Firstly, and following the previous point, it could test whether hypocrisy is associated with deceptive or genuine moral intentions. Secondly, it could also examine whether the current results could be extended to different scenarios, where the cost to be overcome by participants is other than effort, where hypocrisy is manifested as other types of moral discrepancy (e.g. attitudes/actions, judgements for self/other actions), where judgments and actions are made publicly vs privately, and where hypocrisy is embedded in real social contexts (Batson et al., 1997; Bruneau, Kteily, & Urbiola, 2020; Warach, Josephs, & Gorman, 2019). Furthermore, future research could shed light on what personality, affective and social traits could modulate the motivational aspects of

moral hypocrisy (Batson et al., 1999, 2002; Carpenter & Marshall, 2009; Dong et al., 2019; Lammers et al., 2010; Polman & Ruttan, 2012; Tang et al., 2018; Valdesolo & DeSteno, 2008). Thus, the present results open a research opportunity for future investigations to illuminate the psychological aspects of moral hypocrisy.

In summary, in this chapter I tested whether moral hypocrisy, a discrepancy between moral judgements and actions, is associated with lower prosocial motivation in the form of higher effort cost sensitivity. Results suggest that this hypothesis is true in at least two aspects: as hypocrisy increases, (i) prosocial effort decisions to help others decrease, (ii) and helping actions get more superficial. These results open an investigative door into a motivational, goal-directed component of moral hypocrisy.

Chapter 6

Neural self/other reward representations are linked to interoception and prosocial motivation

6.1 Summary

The ACC and AI are two key regions in self and vicarious reward processing, as well as interoception, i.e. the ability to perceive internal bodily signals. Previous studies on self-processes suggest that while AI activity predicts interoceptive ability and affective experience, ACC is mainly involved in motivated actions, proposing distinct roles for ACC and AI in reward processing. However, it is unknown whether the same distinction can be identified when processing vicarious rewards, and how similar these neural representations are to self reward processing. Here, participants underwent fMRI while passively observing cues that indicated monetary gains and losses for self or other. Using this design, I tested the hypothesis that similarity between reward signals for self and other in the ACC and AI were linked to (i) people's sensitivity to internal bodily signals (interoceptive sensibility) and (ii) motivation to exert effort to obtain rewards for self or others (prosocial motivation). Behaviourally, greater reliance on interoceptive signals was linked to increased incentivisation by others' rewards but not for self-benefitting behaviours. fMRI results revealed that higher reliance on interoceptive signals was associated with more dissimilar neural reward representations between self and other in the right AI. However, participants who were more motivated to exert effort for others' rewards, regardless of their interoceptive sensibility, showed more similar BOLD activation patterns between self and other rewards in the ACC. Thus, these results suggest that while ACC representations of others' rewards are tied to prosocial motivation, AI representations are anchored to bodily information.

6.2 Introduction

The anterior cingulate cortex (ACC) and the anterior insula (AI) have been identified as processing salient stimuli in multiple domains (Apps, Rushworth, & Chang, 2016; Craig, 2002; Devinsky, Morrell, & Vogt, 1995; Dosenbach et al., 2007; Lamm, Decety, & Singer, 2011; Lockwood, 2016; Menon & Uddin, 2010; Seeley et al., 2007). In reward evaluation, AI is predominantly involved in the affective experience of a reward, while ACC processes its motivational aspects (Craig, 2002; Dillon et al., 2008; Holroyd & Yeung, 2012; Paulus, 2007; Paulus & Stewart, 2014; Rolls, 2016; Williams, Bush, Rauch, Cosgrove, & Eskandar, 2004). Importantly, states of the body informed by interoception can modulate both aspects of reward processing (Naqvi & Bechara, 2009; Paulus, 2007; Paulus & Stewart, 2014). For instance, how hungry and tired an agent *feels* will influence its eating experience and its motivation to get food. Crucially, even though both ACC and AI are involved in interoception, activity in the AI, especially in its right dorsal portion, is more consistently linked to interoceptive ability (Craig, 2009; Craig, 2002, 2003; Critchley, 2005; Critchley, Wiens, Rotshtein, Öhman, & Dolan, 2004; Ernst, Northoff, Böker, Seifritz, & Grimm, 2013; Naqvi & Bechara, 2009; Terasawa, Fukushima, & Umeda, 2013; Terasawa et al., 2021). Thus, the ability of an individual to be aware of their body signals might relate to affective reward processing in AI (Craig, 2009; Craig, 2002; Critchley, 2005; Naqvi & Bechara, 2009).

ACC and AI are involved not only in processing rewards for self but also for others (Apps & Ramnani, 2014; Apps et al., 2016; Lockwood, 2016; Lockwood, Apps, Roiser, & Viding, 2015; Mobbs et al., 2009; Morelli, Sacchet, & Zaki, 2015), and individual differences in their vicarious reward responses can predict prosocial behaviour (Fukuda et al., 2019; Greening et al., 2014; Harbaugh et al., 2007). However, it is still unclear whether there are distinctive functions of ACC and AI that emulate reward processing for self. Importantly, AI interoceptive activity overlaps with its responses to socio-affective tasks, and individual differences in interoception and alexithymia, a close related

subclinical trait, modulate AI activity to social stimuli (Bird et al., 2010; Ernst et al., 2013; FeldmanHall et al., 2013; Silani et al., 2008; Terasawa et al., 2013, 2021; Zaki et al., 2012). Furthermore, higher interoceptive ability has been associated with greater emotion recognition, affective and cognitive empathy, and helping behaviour (Grynberg & Pollatos, 2015; Lischke et al., 2020; Piech et al., 2017; Shah et al., 2017; Terasawa et al., 2013). Thus, AI might have vicarious reward representations linked to interoception in a similar manner that self rewards.

How similar are self and vicarious reward representations in the ACC and AI? Action-perception theories on empathy predict that vicarious representations of outcomes and affective states are isomorphic to first-hand experiences (De Waal & Preston, 2017; Preston & de Waal, 2002). Nevertheless, right dAI might be crucial in distinguishing the target of an outcome, suggested by its key role in interoception, self-awareness and body ownership (Craig, 2009; Craig, 2002, 2003; Critchley, 2005; Critchley et al., 2004). From this perspective, people who have higher interoceptive ability might have more dissimilar patterns of activation between self and other rewards in right dAI. Previous studies have indirectly supported this claim, revealing that the right AI, and dAI, do not show similar mechanisms between self and other outcomes in other modalities such as pain (Corradi-Dell'Acqua et al., 2016; Crockett & Lockwood, 2018; Hu et al., 2021; Rütgen et al., 2015). On the other hand, it is unclear whether vicarious representations in the ACC, in particular the ACCg, portion specifically associated with social processes (Apps & Ramnani, 2014; Apps et al., 2016; Behrens, Hunt, Woolrich, & Rushworth, 2008; Chang, Gariepy, & Platt, 2013; Lockwood, Apps, & Chang, 2020; Lockwood et al., 2015) share representational patterns of activation with self, and whether individual differences in these patterns of activation can be associated with prosocial motivation.

Most previous studies on neural overlapping between *self* and *other* processes have used univariate analysis, where the strength of activity in certain brain location is correlated between *self* and *other* conditions. However, in the last decade, multivariate

pattern analysis techniques in fMRI have allowed to disentangle the spatial representation of an outcome within an area, allowing to calculate the similarity of this representation with another event pattern (Kriegeskorte, Mur, & Bandettini, 2008). Thus, how much people vary in their similarity patterns for self and other representations of rewards in AI and ACC might uncover functional distinctions and mechanisms in their relationship with interoception and prosocial motivation respectively.

Crucially, individual differences in interoceptive ability not only could unveil distinctive neural functionality but also reveal the influence of bodily signals on self-benefitting and prosocial behaviour. Even though interoception has been linked to self and social decision-making, the mechanisms underlying these links are still unknown. Often, to obtain a reward, either for self or for other, an agent needs to make some effort, e.g. travel long distances to collect food and share with the community. How hungry the agent feels, and how hungry they perceive the community is, through interoceptive information and empathetic representations, could modulate their motivation to act (Gao, Ping, & Chen, 2019; Ondobaka et al., 2017; Palmer & Tsakiris, 2018; Seth, 2013). Likewise, the agent's bodily signals also inform the state of its muscles, having an evaluation of the effort costs involved in travelling for food (Herbert, Ulbrich, & Schandry, 2007; Iodice, Porciello, Bufalari, Barca, & Pezzulo, 2019; Machado et al., 2019; McMorris, Barwood, & Corbett, 2018). Therefore, while individual differences in interoceptive ability can modulate how an agent perceives self and others' benefits (rewards), facilitating behaviour, it could also influence the evaluation of the costs associated with the action, refraining the agent to act.

In the present study, I aimed to examine three questions. First, whether vicarious reward processing in AI and ACC can be functionally disentangled, hypothesising that AI will be associated with interoception, while ACC with prosocial motivation. Secondly, whether these potential associations are linked to similar or dissimilar representations between self and other rewards. Finally, whether interoceptive ability modulates prosocial

and self-benefitting decisions through the benefit, the cost, or both elements involved. To research these questions, participants completed three tasks, measuring (i) neural reward representational similarity between *self* and *other* in AI and ACC through an fMRI task where participants passively witnessed monetary gains and losses for *self* and *other*; (ii) reliance on external versus internal bodily signals through a recently developed respiratory task (Murphy, Catmur, et al., 2018); and (iii) willingness to obtain rewards for others and for self through effortful actions, disentangling effort costs from reward benefits in motivated decisions. Using this design, I aimed to examine how vicarious reward representations in the ACC and AI relate to individual differences in reliance on interoceptive signals and prosocial motivation.

6.3 Methods

6.3.1 Participants

The study was approved by the University of Oxford ethics committee. 80 healthy volunteers aged 18-38 years, recruited from the University and the city of Oxford, participated in the study, gave informed consent and were compensated for their time (£15 per hour for the fMRI session, £8 per hour for behavioural session, plus extra potential bonuses earned through the reward-related tasks). Participants were excluded if they had a history of neurological or neuropsychiatric disorders, psychoactive medication or drug use. Pregnant participants were also excluded from the study, as well as those who had participated in studies involving social interaction and who had studied psychology, addressing concerns that prior experience could influence psychological and brain processes.

From this pool of 80 participants, 61 participants (age M = 22.6, SD = 3.8, 32 females) were included in the behavioural analysis – four participants failed to show-up for the behavioural session, two participants expressed doubts about the social

manipulation and the existence of the receiver, nine did not complete either the interoception respiratory task or the prosocial effort, and four were excluded from analysis because of poor quality data in either of the two tasks (see below).

From these 61 participants, 56 were included for the fMRI analysis (age M = 22.3, SD = 3.3, 30 females). One participant's neuroimaging data was not registered due to technical issue of the scanner, one participant showed excessive head motion in the scanner, and three participants were excluded from analysis due to poor quality data in the value representation task (see below).

6.3.2 Procedure

Participants undertook different sessions, including an fMRI and a behavioural session. The value representation task was measured while in the fMRI scanner, whereas the interoception respiratory task and the prosocial effort task were tested in a desktop computer as part of the behavioural session.

6.3.2.1 fMRI session: Value representation task

Before going into the scanner, participants were randomly assigned the role of the Decider while another unknown person, a confederate, was assigned the role of the Receiver, following the social manipulation described in Chapter 2. Participants witnessed six trials as practice and were then guided to the fMRI scanner. Participants were instructed that, in each trial, a computer would randomly select a value to add or to subtract to their total earnings (*self* trials) or to the earnings of the receiver (*other* trials). Thus, participants passively observed screens indicating that themselves or the anonymous receiver had gained or lost different amounts of money (Fig. 6.1A). From this set of trials, participants were told that one of them would be randomly selected and added or subtracted to the receivers' or to their own earnings at the end of the study.

There were twelve different types of trials in the task according to a 2 (target: *self*, *other*) x 6 (value: -£15, -£9, -£3, £3, £9, £15) design. Each trial consisted of a screen indicating "You *OR* Receiver / Won *OR* Lost / Value" presented on a light grey background for 1400 ms followed by a 100 ms interstimulus interval (Fig. 6.1). An equivalent number of catch trials were also presented to ensure attention to the task. Furthermore, an equivalent number of null trials were also presented for 3000 ms. The trial presentation order was pseudo-randomised using a "type 1 index 1" sequence which ensures that each stimulus is preceded and followed by all other stimuli (Aguirre, 2007; Drucker & Aguirre, 2009). For the purpose of this experiment, the four most efficient "type 1 index 1" sequences were chosen among 1000 randomly generated sequences using the Python code available online[4]. *Efficiency* is the ability of a simulated sequence to create a temporal pattern of neural activity that passes through the assumed hemodynamic response function (HRF), considering the dissimilarity matrix of the events in the design which accounts for potential neural habituation effects. Each of the four chosen sequences included 14 presentations of each stimulus, the blank screen (null) and catch trials for a total of 196 trials. The full experiment therefore included a grand total of 56 presentations of each stimulus and a scanning time of approximately 21 minutes across four runs. The presentation order of the four sequences was counterbalanced across participants.

To ensure that participants maintained attention across all trials, two alternative forced choice catch trials were included in the sequences. During these catch trials, participants were asked to indicate as quickly as possible if the previous trial was targeted toward the self or the receiver (i.e. "You or Receiver?") or to identify the correct amount that was awarded in the previous trial compared to a randomly generated amount (e.g. "Won £15 or Lost £9") using a response box. The catch trial prompt was presented for the same duration as regular trials, 1400 ms. This task was added to ensure participants paid

[4] https://cfn.upenn.edu/aguirre/wiki/public:t1i1_sequences/

attention to the stimuli without influencing the processing of the magnitude and the target of the gain or loss. Accuracy at the catch trial task was computed by counting the number of correct answers and dividing by the total number of catch trials. Participants who had an accuracy below 50% of the catch trials were excluded from analysis (n = 3).
Once participants finished the value representation task, and outside the scanner, they were informed about the random trial selected from the value representation task. If it was a *self* trial, the amount of money earned or lost was added or subtracted to their earnings at the end of the behavioural session. If the trial was an *other* one, participants' own earnings did not change.

The value representation task included reward gains and losses for self and other to calculate similarity between beneficiaries/targets across events in the AI and ACC. This approach was taken for three reasons. First, previous studies have shown that these regions respond to negative and positive outcomes for self and others in a wide range of formats, signalling the saliency of the stimuli, and guiding behaviour irrespective of their valence. Second, and consistently, in the current data similarity between gains and losses was significantly above zero in these areas, suggesting that responses to rewards was similar irrespective of their valence. Finally, since the design plan was to correlate brain similarity with individual differences in motivation and interoception in two different tasks, and given that gains and losses had similar spatial representation in the current data, I chose to use similarity across events for *self* and *other* to gain statistical power and avoid Type II error, sacrificing in the process a degree of resolution, as my hypotheses were not specific to magnitudes of reward.

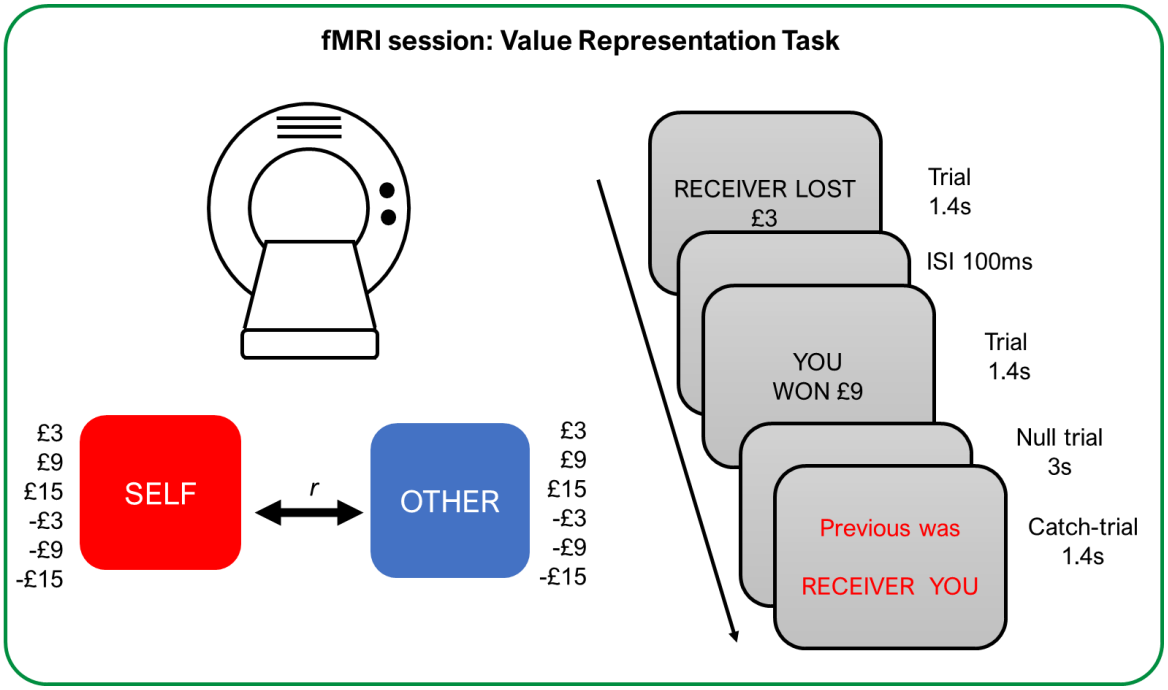


Figure 6.1. fMRI session – The value representation task. *Bottom left*. Neural similarity between self and other representation to reward gains and losses was measured through the value representation task. Six levels of rewards were defined, being half gains and half losses for both self and other. Similarity was the Pearson correlation between the spatial representation of self and other activity across events. *Right*. While in the scanner, participants completed the value representation task, where they passively witnessed financial gains and losses for themselves and for another unknown person, the receiver. In addition, two catch trials were used to exclude participants who did not pay attention to the task. These catch trials asked for either who was the target of the last outcome, or what was the magnitude of the reward in the previous trial.

6.3.2.2 Behavioural session: Prosocial effort task

At least one week after the fMRI session, participants attended a behavioural session where they completed a series of tasks. One of them was the prosocial effort task, which measures how motivated participants are to obtain rewards for self and others (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021, Fig. 6.2A). This task, which was extensively described in Chapter 2, consists in trade-offs between financial rewards (represented by number of credits) in exchange of different levels of physical effort (grip force). For each trial, participants chose between two options: a rest baseline associated with no effort and low reward (1 credit); and a work offer, with higher monetary gain (2-10 credits) for higher effort that varied across 30-70% of each participant's maximum voluntary contraction (MVC). Importantly, in half of the trials the reward was obtained by

participants themselves (*self* trials), while in the other half rewards was received by another unknown person, the receiver (*other* trials). If the work option was chosen, participants needed to make the effort required to obtain the credits on offer. They received zero credit if they failed to do so. If the rest option was chosen, participants rested for the same amount of time and obtained the baseline credit in exchange. Participants who missed more than 10% of the trials in the prosocial effort task were excluded from analysis (n = 1).

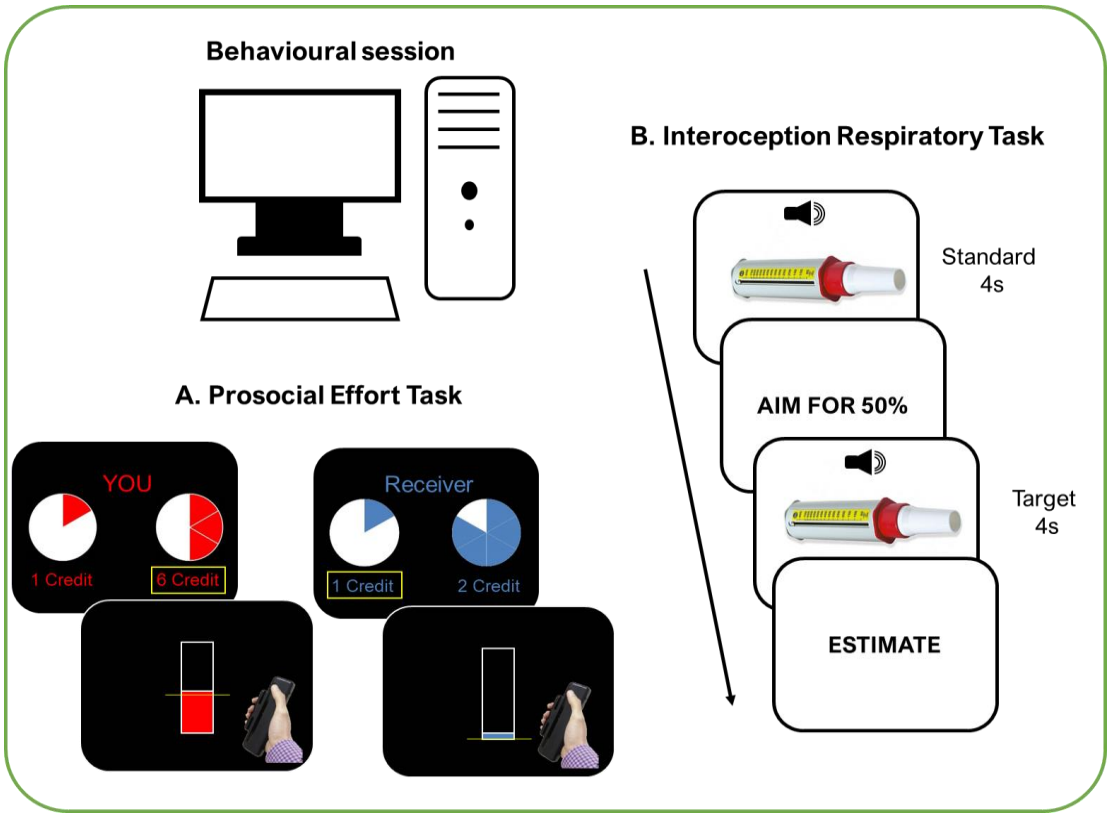


Figure 6.2. Behavioural Session. At least one week after the fMRI session, participants undertook a behavioural session where the prosocial effort task (prosocial motivation) and the interoception respiratory task (reliance on internal relative to external cues) were measured. A. *Prosocial Effort task.* Participants traded-off different level of physical effort against variable amounts of rewards that could be delivered to either participants themselves or an unknown person*. B. Interoception Respiratory Task.* In each trial participants blew a peak-flow meter twice, the first blow being the standard for that trial, and the second, the target. In the target blow, participants aimed to achieve a percentage of the standard, with participants estimating their actual performance at the end of the trial. Note that sequence of the screens is an illustration of the actual test, as they were blindfolded in the task.

6.3.2.3 Behavioural session: Interoception respiratory task

Participants completed the interoception respiratory task (Murphy, Catmur, et al., 2018) at the end of the behavioural session. The goal of this task was to assess how much participants rely on their internal signals relative to exteroceptive cues, measuring the propensity to be aware of interoceptive information, i.e. interoceptive sensibility. This task was preferred over the often used heartbeat counting task (Dale & Anderson, 1978; Schandry, 1981), where participants count their heartbeats within a specified time interval that compare to an objective measure, because of three reasons. First, in everyday life, people are more aware of their breathing patterns than their heartbeats, making the respiratory domain more familiar and closer to real life. Second, the respiratory task focuses on the extant in which people are aware of their internal signal, which is directly related to some of the hypotheses of this study (self-awareness and body ownership function of AI). On the other hand, the heartbeat task is focused on interoceptive accuracy, i.e. how accurate a person is to perceive internal bodily changes, which is not directly related to the hypotheses of the current study. Finally, the heartbeat counting task has received some criticism in recent studies, rising questions to its validity (Khalsa & Lapidus, 2016; Murphy, Brewer, et al., 2018; Ring & Brener, 2018). The design of the respiratory task controls for potential exteroceptive explanations of people's interoceptive abilities, addressing some of the problems of the heartbeat task. Taken together, the respiratory interoception task is more appropriate for the aims of this study.

To assess participants' speed of exhalation, a standard peak flow meter, a medical device that measures the maximal peak of air flow in litres per minute, was used. Due to this, participants who had a history of breathing difficulties did not undertake this task. On each trial, participants were first required to make a large exhalation into the peak flow meter. This first exhalation was taken as the 'standard' for that trial, i.e. 100% performance (Fig. 6.2B). Participants were then required to perform a second exhalation. In this second, 'target' exhalation, participants were told to aim to perform a percentage of

the standard exhalation (e.g. 30%) from that trial. There were four possible targets for each trial, i.e. 30, 50 70 or 90% of the standard. Once the participant had performed the target, they were asked by the experimenter to estimate their performance as a percentage of the standard blow.

Trials in the interoception respiratory task were completed under two conditions. In the *internal condition*, participants performed the standard and the target exhalations while hearing background white noise through headphones connected to a laptop (~79 decibels), ensuring they were unable to access to auditory cues related to their exhalation noise, and relied only on their internal, interoceptive signals. In the *external condition*, participants performed each exhalation with a background white noise coming from a laptop loudspeaker (~79 decibels) located approximately one meter from their right side, allowing them access to the external cues (i.e. their exhalation noise) while maintaining the distracting background noise. For both conditions, the white noise started approximately one second before the exhalation, lasting for four seconds. Contrasting the performance in each of these conditions therefore allows measurement of the extent in which participants rely on interoceptive relative to exteroceptive signals.

During the task, and across conditions, participants were blindfolded, to prevent them from checking the values on the gauge, from attending to any visual cues about the body (e.g. chest movements), and thus from assessing their performance. For both internal and external conditions, participants completed six repetitions of the four (i.e. 30, 50, 70, 90%) targets presented in a random order, completing a total of 24 trials per condition. The order in which the internal and the external conditions were presented was counterbalanced across participants. To ensure that the 30% target trial could be measured, participants were required to surpass a threshold of 200 L/min in their standard exhalation. In case they did not accomplish this, the standard blow was repeated until the threshold was surpassed. No feedback about participants' performance was provided across the experiment.

Before starting the experimental trials, the experimenter explained the task to the participants, and demonstrated how to perform the exhalations using a mouthpiece that was not attached to the gauge. After this demonstration, participants were invited to perform one internal trial as practice with a target of 50%, where they also confirmed that the volume level of the white noise was comfortable. In cases where the practice trial was not well executed, the trial was repeated. The gauge in which participants perform their exhalations was gently secured in a horizontal position using a vice clamp and elevated in line with each participant's mouth using a stand. Participants were instructed to keep their hands resting at the bottom of the stand during exhalations, using their hands only to locate the gauge prior to performing exhalations. Moreover, they were instructed to be still in the chair and sat upright, without pushing the mouthpiece forward while exhaling. Trials in which these conditions were not accomplished were repeated or removed from analysis. Participants with more than 10% of missed trials in either of the two conditions were excluded (n = 3).

6.3.3 Analysis of behavioural data

6.3.3.1 Interoception score: reliance on internal relative to external signals

To calculate participants' reliance on their interoceptive signals, the absolute error scores were computed for each trial of the interoception respiratory task, such that:

$$AE_{ij} = \left[\frac{T_{ij} - E_{ij}}{T_{ij}}\right]$$

$$\overline{AE_j}\begin{cases}\overline{AE}_{int}\ for\ internal\ trials\\ \overline{AE}_{ext}\ for\ external\ trial\end{cases}$$

Where the absolute error *AE* in the trial *i* for the participant *j* is the absolute difference between the actual performance of the participant in target *T* as a percentage of the standard, and their estimation *E*, divided by their actual performance T. The mean of the

AEs is computed separately for the internal and the external conditions for each participant. The interoception score was then the difference in performance between the internal and the external condition (Fig. 6.3A), such that:

$$Interoception_j = \overline{AE}_{ext[j]} - \overline{AE}_{int[j]}$$

Where the interoception score for the participant j is the difference between their mean AEs in the external condition and in the internal condition. Scores above zero indicate that a participant made more mistakes in the external than the internal conditions, suggesting that they are more internally focused. On the other hand, negative values mean better performance in the external than the internal condition. The interoception score therefore reflects how much participants rely on internal versus external signals.

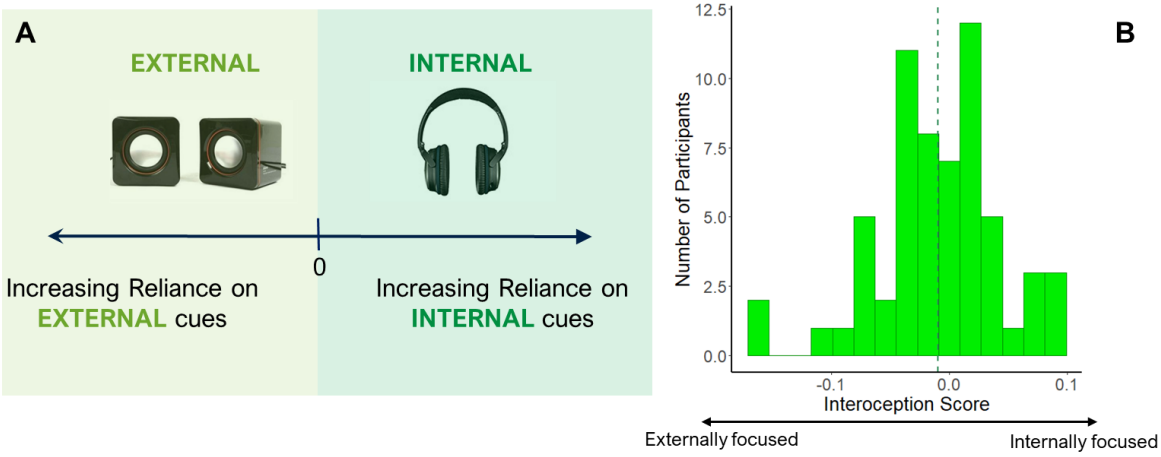


Figure 6.3. Calculating reliance on internal versus external signal. *A.* The interoception score was computed considering the absolute error between participants' actual performance in the target blow and their estimation in two conditions. An internal condition - where participants performed the task with white noise coming through headphones placed over their ears, relying only on their internal signal; and an external condition, where the white noise came from loudspeakers, allowing them to hear external cues such as their exhalation noise. Interoception scores were computed as the difference between the errors of these two conditions, such that positive values reflected more reliance on internal comparative to external cues. *B.* The interoception score followed a normal distribution with an average mean of -0.01 (indicated by the green dash-line).

6.3.3.2 Testing the links between interoception and motivation for self and other rewards

For the prosocial effort task, choices for the work offer relative to rest were taken as an index for motivation to obtain self and other rewards. Since the main hypotheses in this study were related to reward vs effort sensitivity, the computational parameters λ (see Chapter 2 for details) were not used, as they consider a specific relationship between both informational aspects of the motivated decision, i.e. effort and reward. I used mixed-effects models (MMs) predicting trial-by-trial decisions using the *glmer* function in R. Thus, two models were built to test the effects of interoception on motivation.

Simple model

$$DW_i = \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 R_i E_i + \beta_5 R_i B_i + \beta_6 E_i B_i + \beta_7 R_i E_i B_i$$

Interoception model

$$\begin{aligned} DW_i = {} & \beta_{0j[i]} + \beta_{1j[i]}R_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 I + \beta_5 R_i E_i + \beta_6 R_i B_i + \beta_7 R_i I + \beta_8 E_i B_i \\ & + \beta_9 E_i I + \beta_{10} B_i I + \beta_{11} R_i E_i B_i + \beta_{12} R_i B_i I + \beta_{13} E_i B_i I + \beta_{14} R_i E_i I \\ & + \beta_{15} R_i E_i B_i I \end{aligned}$$

In the simple model, decision to work *DW* in the trial *i* is predicted by the fixed effects of reward *R*, effort *E*, beneficiary *B* and their interaction, with a random intercept clustered in each subject *j*. In this logistic MM, DW is a binary, factor variable. Random slopes for R and E were included as it is expected for these variables to vary across participants. AIC values confirmed that the random-slopes model outperformed the random-intercept only MM. The interoception model adds the interoception score *I* as a predictor together with its interaction with the other independent variables. I examined whether the addition of interoception into the simple model improved model fitting comparing the AIC values between models, and also using loglikelihood tests.

6.3.4 fMRI acquisition and preprocessing

MRI data were acquired using a 3 Tesla Siemens Prisma MRI scanner. The structural scan was acquired at the start of the session using a magnetization prepared rapid gradient echo sequence with 192 slices; slice thickness: 1 mm; repetition time (TR): 1900 ms; echo time (TE): 97 ms; field of view (FOV): 192 × 192 mm; voxel size: 1 × 1 × 1 mm. Field map images were obtained immediately before the value representation task using a double-echo spoiled gradient echo sequence (TR: 488 ms; TE: 4.92/7.38 ms; voxel size: 3x3x3 mm; flip angle: 46°). Four runs of multiband T2*-weighted echo-planar imaging (EPI) volumes with BOLD contrast were collected during the value representation task (72 slices in interleaved ascending order; repetition time: 1570 ms; echo time: 30 ms; flip angle: 70°; field of view: 216x216 mm; matrix size: 108x108; voxel size: 2x2x2 mm with 1 mm gap; MB factor: 3; in-plane acceleration factor: 2). Each run was 5:46 minutes in length, during which 220 functional volumes were acquired.

Preprocessing was performed using fMRIPrep 20.1.0rc1 (Esteban et al., 2019), which is based on Nipype 1.4.2 (Gorgolewski et al., 2011). The following section was partly automatically generated by FMRIPrep.

6.3.4.1 Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2011), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9; Zhang, Brady, & Smith, 2001). Volume-based

spatial normalization to the MNI152NLin6Asym standard space was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalisation: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* (Evans, Janke, Collins, & Baillet, 2012)

6.3.4.2 Functional data preprocessing.

For each of the four BOLD runs per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. A deformation field to correct for susceptibility distortions was estimated based on a field map that was co-registered to the BOLD reference, using a custom workflow of *fMRIPrep* derived from Greve's epidewarp.fsl script[5] and further improvements of HCP Pipelines (Glasser et al., 2013). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson, Bannister, Brady, & Smith, 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI (Cox & Hyde, 1997). The BOLD time-series were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled to

[5] https://www.nmr.mgh.harvard.edu/~greve/fbirn/b0/epidewarp.fsl

MNI152NLin6Asym standard space and spatially smoothed using a 6mm full-width half maximum Gaussian kernel. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals were extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi, Restom, Liau, & Liu, 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation).

6.3.5 fMRI analyses

6.3.5.1 First level model.

First, a general linear model was created on the BOLD response at the participant level using SPM12, with regressors modelling the onset of each event, i.e. -£15, -£9, -£3, £3, £9, £15 for *self* and *other*, and the onset of the catch trials. The regressors were generated using delta functions convolved with SPM12's canonical haemodynamic response function. This first level model also included nuisance regressors modelling non-neural sources: (i) 24 movement-related regressors comprising the six estimated head movement parameters (x, y, z, roll, pitch, yaw), their first temporal derivatives, their

squares, and their squared derivatives; (ii) the first six components of the anatomical CompCor extracted using the CompCor method (Behzadi et al., 2007) implemented in *fMRIprep*; and (iii) a variable number of single event regressors indicating volumes exceeding a threshold of 0.5 mm framewise displacement or 1.5 standardized DVARS and considered as motion outliers (mean % of volumes that were motion outliers = 1.04%, sd = 0.91% , range = 0-4.05%). The four runs were concatenated in the model and constant regressors were added to model each of the four run's mean. The null trials were used as the un-modelled baseline in all models. A high-pass filter with a cut-off of 128 s and SPM12's *AR(1)* correction for serial correlation were applied during the model estimation.

6.3.5.2 Similarity analysis

To examine similarity in neural signals, I created representational maps between self and other rewards. This analysis allows to measure the correlation between reward signals for *self* and *other* across space, differing from the classical univariate analysis of BOLD signal. Thus, individual differences in this analysis do not indicate the magnitude of the BOLD response to an outcome. Rather, the individual differences reported here corresponds to the degree in which the pattern of activity for self rewards within a specific ROI resembles the pattern of activity for other rewards. This allows me to test whether similarity, or dissimilarity, is associated with either interoception or prosocial motivation. Thus, I measured multivariate similarity between self and other trials by calculating the Pearson correlation between the average parametric maps for each condition of interest in seven ROIs. I carried out correlations between events in *self*, and events in *other* trials for each ROI. These correlation coefficients for each participant were used to assess the relationship between individual differences on self/other neural similarity and performance on the interoception respiratory task and the prosocial effort task.

I based my analyses on different portions of the AI and the ACC as ROIs that have been associated with reward processing, interoception and social cognition (Apps et al., 2016; Apps, 2018; Craig, 2009; Craig, 2002; Critchley, 2005; Critchley et al., 2004). For the AI portions, I used the parcellations created by Deen et al., 2011 derived from a voxel-wise k-means clustering approach applied to resting-state. These parcellations divide the insular cortex in three different complexes for each hemisphere: posterior insula, dorsal AI (dAI) and ventral (vAI). This subdivision of the insula has been replicated in other work (Chang et al., 2012; Nomi et al., 2016) and used in other studies (Uddin et al., 2014; Nomi et al., 2018). The 6 insula ROIs were acquired in 2 mm Montreal Neurological Institute (MNI) space directly from the author website[6]. Only right and left vAI (RvAI, LvAI) and dAI (RdAI, LdAI) ROIs were used. I used lateralised AI instead of bilateral regions as I had specific hypothesis about right dAI association with interoception.

For the ACC region, I used thresholded masks taken from the resting-state connectivity-based parcellation by Neubert et al., 2015. This parcellation divides the frontal cortex into 21 different regions. The parcels were acquired in 2 mm Montreal Neurological Institute (MNI) space directly from the author website[7]. I created the three different masks corresponding to bilateral ACCg, area d32 and area v32 modifying the original parcels using the *imcalc* function in SPM12 in the following manner - (i) I duplicated the left hemispheres parcels onto the right hemisphere to create bilateral masks (the atlas was created in participants with a paracingulate sulcus only in the left hemisphere), as I did not have a hypothesis for lateral functional specificity. (ii) For the ACCg region, I used the area 24ab from the atlas, and I removed those voxels located in the posterior portion of the mask in order to capture more closely what the literature has identified as ACCg (Apps et al., 2016; Apps and Sallet, 2017). Thus, multivariate similarity values between *self* and *other* images in each of these ROIs were extracted and

[6] https://bendeen.com/data/
[7] http://www.rbmars.dds.nl/CBPatlases.htm

correlated with the behavioural measures of motivation and interoception, controlling for multiple comparison across these seven regions.

6.3.5.3. Association between neural representations and behavioural measures.

Similarity values obtained across the seven ROIs described above were correlated with reliance on interoceptive signals and motivation for rewards. For the former, I used the interoception score. For the latter, I used the reward and effort beta values obtained for each participant from two MMs where decisions to work are predicted by effort, rewards and their interaction. Crucially, each MM considered beneficiaries separately, with one predicting decision to work only for s*elf*, and the second predicting motivation for *other*. In this way, I could dissociate sensitivity for rewards and effort separately for *self* and *other*, and test for their specific or general relationship with similarity in the ROIs. These two models were clustered with random-intercepts for each participant, and had random-slopes for reward and effort, such that:

$$DW_{self} = \beta_{0j[self]} + \beta_{1j[self]}R_i + \beta_{2j[self]}E_i + \beta_{3[self]}E_iR_i$$

$$DW_{other} = \beta_{0j[other]} + \beta_{1j[other]}R_i + \beta_{2j[other]}E_i + \beta_{3[other]}E_iR_i$$

Where DW is specific for *self* and for *other* trials separately. Thus, the similarity values in ACCg, v32, d32, RdAI, RvAI, LdAI and LvAI were correlated with reliance on internal signal on one hand, represented by participants' interoception scores, and with motivation to work on the other hand, represented by participants' beta estimates for reward and effort for *self* and *other*.

Pearson correlations were used to correlate brain similarity and behavioural measures. Visual inspection of the brain data revealed that there were some evident outliers in the similarity values, which were located multiple standard deviations from the

mean in the distribution. Consequently, for each correlation between ROI self/other similarity and behavioural measures, participants who had values below and above three standard deviations were excluded. Furthermore, box-cox transformations were applied to those ROIs that showed deviation from normal distribution. All behavioural measures were normally distributed. Results obtained from these correlations were corrected for multiple comparison across the seven ROIs using false discovery rate (FDR, Benjamini & Hochberg, 1995; Benjamini &Yekutieli ,2001).

6.4 Results

6.4.1 Behavioural results: Interoceptive sensibility facilitates prosocial motivation.

To test whether reliance on internal, interoceptive signals is associated with motivation for self or others' rewards, behavioural indexes from the interoception respiratory and the prosocial effort tasks were used. Reliance on internal relative to external signals were computed through the difference in performance between internal and external condition in the interoception respiratory task (Fig. 6.3A), with positive values meaning better performance in the internal condition while negative values meaning better performance in the external condition. The interoception score was normally distributed across participants (Kolmogorov-Smirnov normality test), with an average close to zero (M = -0.01, SD =0.05, Fig. 6.3B).

To test whether the interoception score was linked with prosocial motivation, two different MMs were built, having decision to work as the dependant variable. One of the models, the simple model, was a full model with effort, reward and beneficiary, together with their interactions, predicting decisions to work. A second model included the interoception score as an independent variable as well as its interaction with all the other

variables. Model comparison revealed that the interoception model outperformed the simple model according to their AIC (AIC: simple model = 5115.5, interoception model = 5079.3, Fig. 6.4A). Furthermore, a loglikelihood ratio test showed that adding the interoception score into the simple model significantly improved model fit (χ^2_{diff} = 52.28, df_{diff} = 8, p < 0.001). Thus, decisions to work in the prosocial effort task are better explained by a statistical model that includes people's reliance on internal signals.

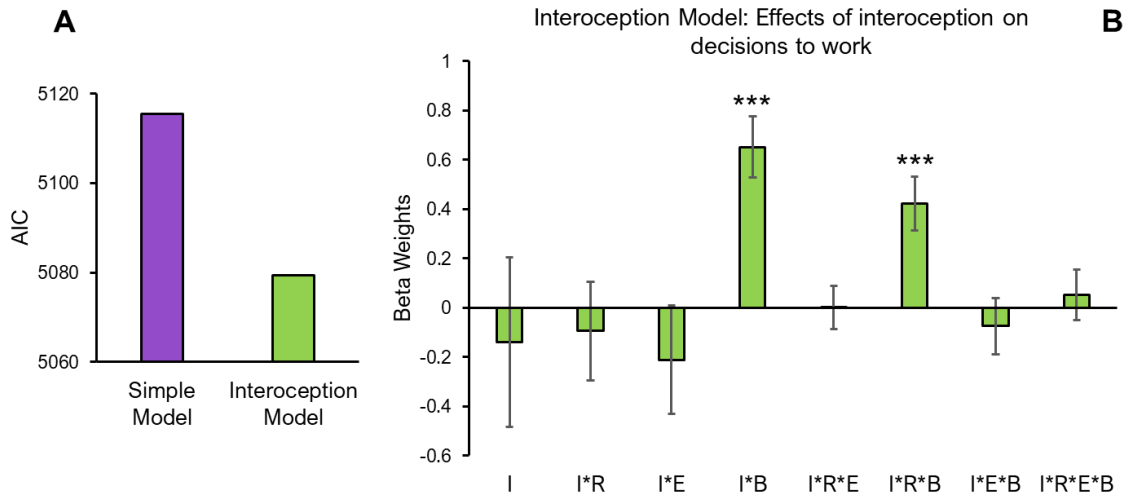


Figure 6.4. Interoception score is associated with prosocial motivation. *A. Interoception score improves model fitting*. Adding interoception score (I) as an independent variable predicting decisions to work improves model fitting according to AICs (Akaike Information Criterion, y-axis). The simple model only includes effort (E), reward (R) and beneficiary (B) as independent variables*. B. Effects of reliance on internal vs external signal on effort-based decisions.* Beta weights (y-axis) from the interoception model revealed that interoception score had significant interactions with beneficiary, and with reward and beneficiary in predicting effort decisions, suggesting that the modulation of interoception on motivation varies according to who is getting the reward on offer. Note that only interoception effects on decisions are depicted in the plot. Error bars depict SEM. *** p < 0.001.

The interoception model revealed a three-way interaction between interoception, reward and beneficiary (b = 0.42, SEM = 0.11, z = 3.87, p < 0.001, Fig. 6.4A), such that people who relied more on their internal relative to external signals, chose to work consistently more as the reward on offer for the receiver increased (Fig. 6.5B). Participants more externally oriented, on the other hand, were more reluctant to work for others regardless of the reward on offer. Importantly, this effect seems to be specific to prosocial motivation, as the interoception score did not modulate behaviour in *self* trials

(Fig. 6.5A). Therefore, people who are more internally focused are more sensitive to others' rewards and guide their behaviour accordingly, showing similar motivation to obtain reward for others than themselves, whereas externally focused people are less incentivised by others' rewards.

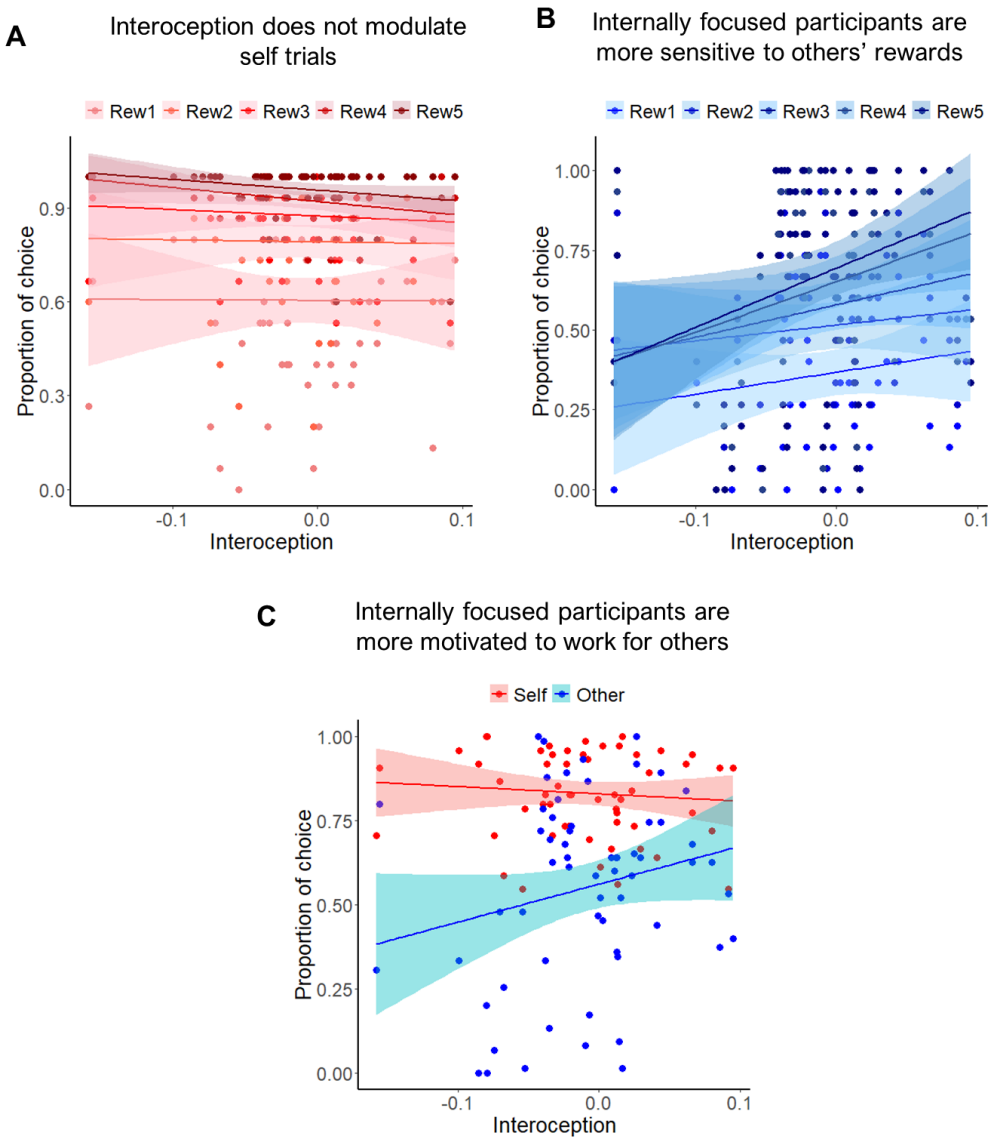


Figure 6.5. Interoception facilitates prosocial choices through sensitivity to others' rewards. A. *Interoception did not influence motivation for rewards in self trials*. Participants did not show different patterns in their choice proportions to work over rest (y-axis) depending on their level of reliance on internal vs external signal (x-axis) across different reward magnitudes. B. *Interoception influences motivation for others' rewards.* As participants relied more on internal than external cues (x-axis), they chose to work consistently more as the reward to be received by another person increased. Participants who relied more on external cues were more reluctant to work to benefit others regardless the reward on offer. Y-axis depicts proportion of choices to work over rest. C. As participants relied more on internal than external cues (x-axis), they chose to work more to benefit others, resembling their choice pattern when working to benefit themselves. Y-axis depicts proportion of choices to work over rest.

A two-way interaction effect between interoception and beneficiary was also found (b = 0.65, SEM = 0.12, z = 5.3, p < 0.001), confirming that participants who relied more on their internal than external signals, were more willing to work for others, with a more similar level of motivation to work for *self* and *other* (Fig. 6.5C). It is important to note that adding the interoception score into the model did not alter the results reported in Chapters 3 to 5 that replicated previous studies (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021). Thus, there were main effects of effort, reward, and beneficiary, together with interactions between beneficiary x effort and beneficiary x reward.

6.4.1.1 Reliance on internal signal relates to reward sensitivity for others

To correlate similarity in the brain with indexes of motivation, I used beta values for reward and effort in each participant extracted from two different MMs that takes decisions to work for self and other separately, capturing participants' sensitivity to benefit and costs in individual and prosocial decisions. Before correlating behavioural measures with brain similarity, I tested whether the above-mentioned effects of interoception on motivation were replicated using this different approach. Thus, I correlated the interoception score with reward and effort betas for *self*, *other* and their difference. Results revealed a significant negative correlation between interoception and the difference in reward betas, such that people who relied more on internal signals were similarly motivated for self and other rewards, while people who were more externally focused valued more their own rewards compared to others' (r = -0.39, p < 0.002, Fig. 6.6A). This effect was not found on differences in effort betas, and this correlation was significantly different from the correlation between interoception and sensitivity to reward (Fisher's *Z* transformation, z = -3, p < 0.003), suggesting specificity on interoceptive effects. The effect of interoception on reward seems to be triggered by variability in *other* trials, as interoception positively correlated with sensitivity to others' rewards (r = 0.27, p

< 0.04, Fig. 6.6C), but not with self (Fig. 6.6B), and these effects were significantly different from each other (z = 2.23, p < 0.03). Furthermore, the effects of interoception on others' rewards were also found as significantly different from its effects on effort for *self* (z = 2.19, p < 0.03), and as having a non-significant trend from its effects on effort for *other* (z = 1.92, p = 0.055). Taken together, all these results reveal that people who are more internally versus externally focused are specifically more driven by others' rewards when deciding to put in effort and act prosocially.

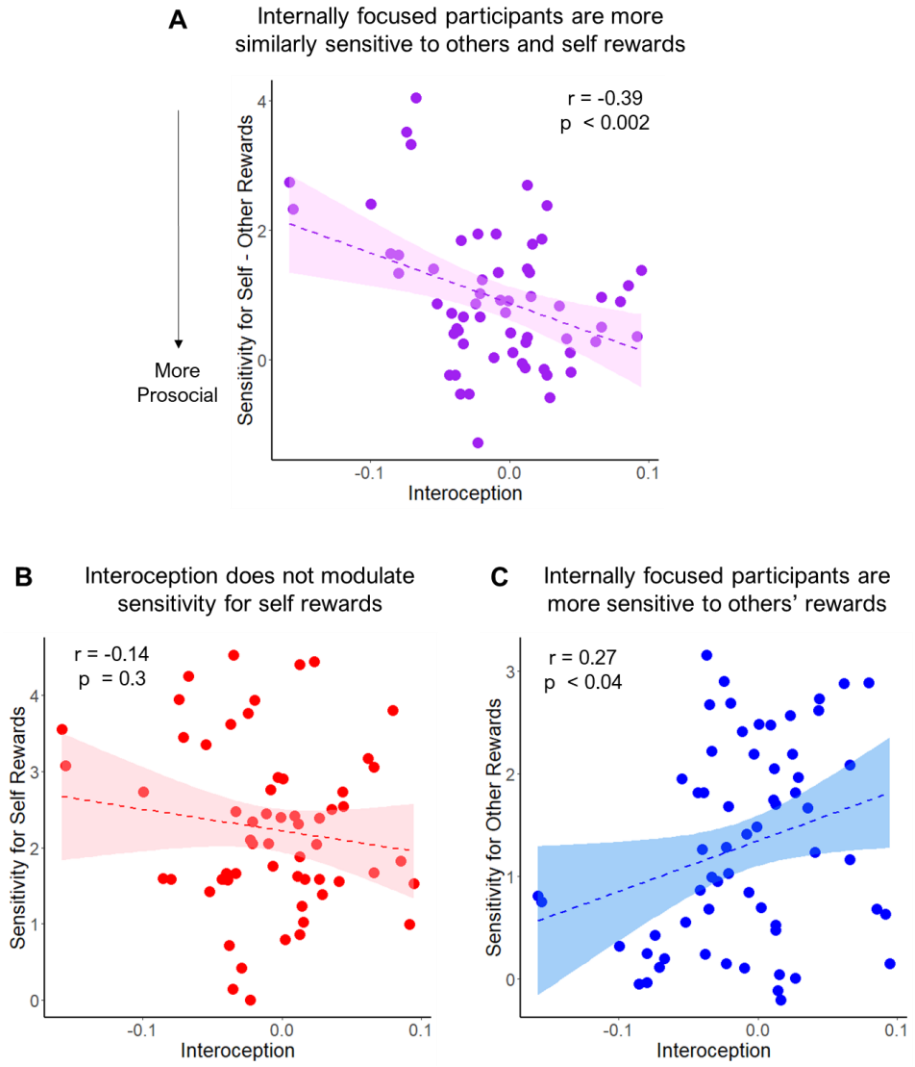


Figure 6.6. Interoception is associated with sensitivity to others' rewards. A. People who rely more on their internal vs external signals (x-axis), weighted rewards more similarly for self and other when making decisions to work to obtain rewards. Y-axis depicts the difference between self and other betas from two mixed models predicting choices separately for self and other. Higher values mean valuing more self rewards than others'. B. Null correlation between interoception (x-axis) and reward betas in the self condition (y-axis). C. However, participants who relied more on internal vs external cues (x-axis), were more sensitive to others' rewards (beta estimates, y-axis).

6.4.2 fMRI results: Different roles of ACCg and RdAI in reward processing.

Next, I focused on whether similarity between self and other rewards in ACC and AI can shed light on different functions relating to vicarious reward processing. I examined whether similarity between self and other rewards in AI and ACC is specifically associated with interoception and motivation respectively. Multivariate similarity analysis was applied between conditions calculating the Pearson correlation between the average parametric map for self and other rewards. Thus, the similarity index represented how spatially similar were the pattern of activation between *self* and *other* across the different reward events. Similarity was calculated for seven ROIs - ACCg, v32, d32, RdAI, RvAI, LvAI and LdAI. These anatomical masks were extracted from parcellations developed in previous studies (Deen et al., 2011; Neubert et al.,2016). Similarity values were then correlated with interoception on one hand, and motivational indexes on the other, controlling for multiple comparisons across the seven ROIs using FDR.

I first tested for correlations between the interoception score and similarity across the seven ROIs. The only area that survived multiple comparison correction was the RdAI ($r = -0.41$, $p < 0.02$ FDR corrected), such that as people relied more on internal relative to external signals, they showed less similar neural patterns between self and other rewards (Fig. 6.7A). No effects were found in any of the other masks. Crucially, the effect of interoception on RdAI was significantly different from correlations between this area and all motivational indexes, i.e. sensitivity to effort and reward for *self* and *other* (Fisher-Z transformation, $p < 0.05$). Thus, the association between reward representations in the RdAI and interoception is specific, and different for self and other.

Next, I tested whether similarity in these seven regions was associated with motivation. Specifically, beta weights for effort and reward separately for *self* and *other*, and their difference, were considered in this analysis. This analysis did not reveal significant results correcting for multiple comparison. However, the correlation between

sensitivity to others' rewards and similarity in the ACCg was significant at an uncorrected threshold and was very close to significant when correcting for multiple comparisons ($r = 0.35$, $p = 0.009$ uncorrected, $p < 0.06$ FDR corrected Fig. 6.7B). Participants who showed more motivation to work for others depicted higher neural similarity between self and other rewards in this region. However, this effect was not significantly different from correlations between ACCg similarity and sensitivity to self rewards, nor with interoception. Yet, it was significantly different from correlations between ACCg similarity and effort betas (Fisher-Z transformation, $p < 0.05$).

For completeness, a whole-brain analysis was performed, where self/other reward similarity was computed for 200 different ROIs, which covered 66% of the brain, using a functional parcellation based on meta-analytic coactivation from Neurosynth[8]. Each of these ROIs were tested for correlations with both interoception and sensitivity to others' rewards following the same protocol used for the other seven ACC and AI ROIs. Uncorrected results ($p < 0.01$) can be found in Appendix B. Importantly, portions associated with AI showed correlation with interoception in this parcellation defined by functional meta-analysis, rather than defined by anatomical properties reported above.

Taken together, RdAI and ACCg appear to be associated with two different aspects on vicarious rewards. Dissimilar reward representations between self and other in the RdAI is associated with reliance on internal signals, while similarity in ACCg was related to prosocial motivation.

[8] https://identifiers.org/neurovault.collection:2099

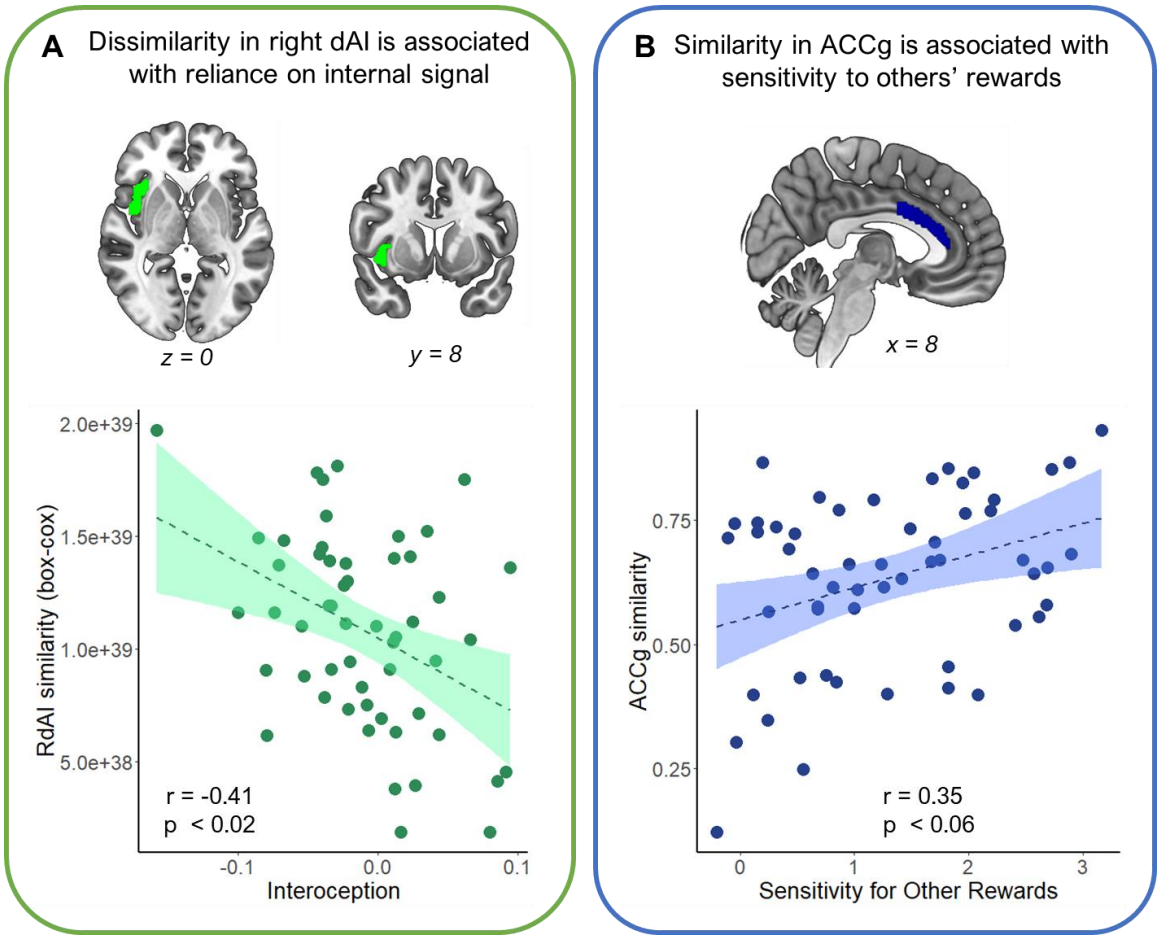


Figure 6.7. Right dorsal anterior insula (RdAI) and the gyrus portion of the anterior cingulate cortex (ACCg) have different roles in vicarious reward processing. A. Interoception was associated with multivariate similarity between self and other rewards only in the RdAI (top panel), among a number of ROIs that included ACCg. People who relied more on internal relative to external signals showed more dissimilar neural responses to reward between self and other. This effect was specific, as it was significantly different from motivational effects on RdAI. Y-axis depicts the similarity values transformed using box-cox, given its non-normal distribution. Higher values mean more similarity between self and other reward representation. B. Prosocial motivation showed a non-significant trend (correcting for multiple comparison) in its association with multivariate similarity between self and other rewards only in the ACCg (top panel), among a number of ROIs that included RdAI. Participants who were more motivated to work for others' rewards showed more similar neural patterns between self and other rewards in the ACCg. However, this effect was not significantly different from interoception effects on ACCg. X-axis corresponds to the reward betas obtained from a mixed-model where decisions to work for others was predicted by effort, reward and their interaction. Positive values mean higher weights for others' rewards.

6.5. Discussion

Interoception has been associated to reward evaluation, emotional processing, and social cognition (Craig, 2002; Critchley, 2005; Gao et al., 2019; Naqvi & Bechara, 2009;

Palmer & Tsakiris, 2018; Paulus, 2007). Two neural structures, the ACC and the AI, have been not only identified as the core structures of interoception, but have also been shown to co-activate in reward, affective and social processes (Craig, 2002, 2009; Critchley, 2005; Critchley et al., 2004; Fan et al., 2011; Lamm et al., 2011; Lockwood, 2016; Paulus, 2007; Paulus & Stewart, 2014). Here, I aimed to shed light on the relationship between interoception and vicarious reward processing, examining in turn whether AI and ACC have distinctive or common roles in processing rewards obtained by other people. Using neural similarity between self and other rewards and individual differences in interoception and prosocial motivation, I found two novel aspects in the links to vicarious reward processing. First, behaviourally I found that interoception is associated with prosocial motivation to benefit others, something that has been suggested in the past but rarely tested. Crucially, these two processes were linked to different and specific sub-regions of the ACC and AI involved in vicarious rewards. Dissimilarity in the right dAI representations for self and other rewards was associated with higher interoceptive sensibility, consistent with its role in self-awareness and self/other differentiation (Craig, 2002, 2009; Critchley, 2005; Singer, Critchley, & Preuschoff, 2009). On the other hand, more similar self/other reward representation in ACCg was associated with more willingness to help others. Thus, while ACC representations of others' rewards are tied to prosocial behaviour and motivation, AI representations are attached to bodily signals.

Decades of research have provided evidence that the brain, including ACC and AI, signals not only self but also vicarious outcomes, which has long been suggested trigger prosocial behaviour (Apps et al., 2016; Fukuda et al., 2019; Harbaugh et al., 2007; Hein et al., 2010; Hu et al., 2021). However, as in most cognitive and social processes, high variability exists between people in how they respond to others' outcomes, neurally and behaviourally. Here, I revealed that these individual differences are crucial to understand a functional dissociation between how vicarious reward signals in the AI and ACC link to behaviour. Intriguingly, individuals whose right dAI responded more differently to rewards

for self compared to others were more aware of their interoceptive signals, measured behaviourally one week later. Likewise, similar patterns of reward representation in the ACCg between self and other predicted more prosocial motivation to work for others' rewards, again, measured in a completely different task one week later. These findings suggest that individual differences in the coding of vicarious rewards in the AI and ACC whilst passively seeing them earned or lost, may be linked to different functional processes predicting interoceptive sensibility and prosocial motivation. Dorsal AI result is consistent with the hypothesis that signals from one's body inform one's prosocial behaviour, and that this may be linked to how responsive one is to seeing others receive rewards. As most research has examined vicarious reward processing at the group level, individual differences may have been overlooked in previous studies.

Interoception was also behaviourally associated with the willingness to make effortful actions to reward other people. Interoception has been linked to representations of others' affective and mental states across different contexts (Grynberg & Pollatos, 2015; Lischke et al., 2020; Ondobaka et al., 2017; Palmer & Tsakiris, 2018; Shah et al., 2017; Terasawa et al., 2013), and it has been widely related to emotional awareness and experience (Craig, 2002, 2009; Critchley & Garfinkel, 2017; Critchley et al., 2004; Seth, 2013). Why should interoception and prosocial behaviour be linked? The current results suggest that interoception facilitates prosocial behaviour through the representation of others' rewards. Thus, interoceptive signals might guide the understanding of others' affective states, motivating actions that bring a positive emotional effect in the other person. This is consistent with the findings reported in Chapter 4, which revealed that people who are more emotionally reactive across affective and subclinical traits manifest more prosocial actions.

Importantly, interoception did not modulate the effects on self costs on prosocial decisions, neither on cost nor benefits in self-benefiting actions, having specific effects on sensitivity to others' rewards. This is despite interoception being associated with effort

evaluation and execution for self-benefitting acts in previous work (Herbert et al., 2007; Iodice et al., 2019; McMorris et al., 2018). However, these previous studies have focused only on effort evaluation, without contrasting it with reward incentives. Furthermore, these studies have mainly assessed interoception either through accuracy tasks (e.g. heartbeat counting task) or physiological measures, without considering interoceptive sensibility and awareness. Interoceptive sensibility is crucial in separating self and social outcomes, (Palmer & Tsakiris, 2018; Shah et al., 2017), and this could have made the representation of others' rewards salient in the prosocial effort task, which includes both types of decisions. Moreover, the interoceptive task used here measures interoception, but specifically to the respiratory system, not heartbeats, controlling for confounds present in the classical heartbeat task (Murphy, Catmur, et al., 2018). This difference between cardiac and respiratory interoceptive domains might have also influenced the modulation of interoception on self processes. Finally, some of these previous studies have focused on how interoceptive signal monitors behaviour while fatigued (McMorris et al., 2018). Fatigue might make interoceptive information salient, facilitating its modulation on effort evaluation and self-related information. Taken together, the current experimental design had a different approach from previous experiments which might affect the relationship between interoception and self costs and benefits.

Even though interoception facilitated prosocial motivation, it seems that the neural substrate underlying these processes are distinct with respect to vicarious reward processing. Interoception was associated with dissimilar representations of self and other rewards in the dorsal portion of the right anterior insula, but not in the ACC. The insular cortex has been strongly associated with interoception (Craig, 2002, 2009; Critchley et al., 2004). Its more posterior portion is involved in primary interoceptive representation of the physiological state of the body through thalamocortical pathways, and it has strong connections with dAI, and both share some of their connectivity fingerprints (Craig, 2002, 2009; Critchley, 2005; Nomi, Schettini, Broce, Dick, & Uddin, 2018; Nomi et al., 2016;

Uddin, Nomi, Hébert-Seropian, Ghaziri, & Boucher, 2017). Thus, it has been argued that the level of integration of the interoceptive information increases in dAI, where emotional, social and cognitive factors are considered through connections with other brain regions, especially the prefrontal cortex, the ACC and the limbic system (Craig, 2002, 2009; Critchley, 2005; Uddin, Kinnison, Pessoa, & Anderson, 2014; Uddin et al., 2017). Thus, this area has been proposed as crucial for self-awareness, where multiple sources of information are integrated to represent body-ownership and sense of agency (Craig, 2002, 2009). This body awareness seems to be crucial then in the distinction between self and other processes (Palmer & Tsakiris, 2018). Furthermore, previous results indirectly suggest that the right AI, including its dorsal portion, have different mechanisms underlying vicarious and self outcomes (Brethel-Haurwitz et al., 2018; Corradi-Dell'Acqua et al., 2016; Crockett & Lockwood, 2018; Hu et al., 2021; Rütgen et al., 2015). Therefore, right dAI might have representations of vicarious rewards anchored to interoceptive sensibility, with increasing distinction in its signal as people are more aware of their bodily changes.

In contrast, prosocial motivation to help others was associated with greater similarity between self/other rewards representations in the ACCg, as opposed to greater dissimilarity which linked interoception and AI. This effect was weaker, and not significantly different from its interoceptive counterpart, and thus inferences here are made with caution. However, previous evidence suggests that the ACCg signals outcomes mainly when they are received by others relative to self (Apps & Ramnani, 2014; Apps et al., 2016; Behrens et al., 2008; Chang et al., 2013; Lockwood et al., 2015). Thus, different from other portions of the ACC, ACCg might process information in a socially-specific manner. However, the current results used a different approach that reveals novel and distinctive aspects of ACCg compared with previous work. First, here the focus was on individual differences, and how variability in ACCg reward activity can predict prosocial behaviour in a completely different task. This approach highlights

therefore the predictive power of ACCg representations of self and other rewards on prosocial motivation, and not ACCg involvement in a specific process at the group level. Second, and linked to the previous point, the focus here was the representational configuration of the reward signal between self and other, and not the significant univariate difference between each condition related activity. It is likely that the ACCg is involved in processing self and other outcomes, with a stronger activation in the latter condition that surpasses statistical significance. However, here I measured *how similar* the *spatial* activation to self and other rewards is, regardless of the strength of this response for each condition. Thus, more similar self/other representations of this area predict more prosocial motivation, which is consistent with behavioural results in this task – people who are more similarly sensitive to self and others' reward are, in fact, more prosocial. Future studies could further examine the association between univariate, multivariate and behavioural methods on reward representation.

In this chapter, using an approach focused on individual differences, I examined the role of interoception in reward processing in two different dimensions – behavioural and neural. I found that reliance on interoceptive vs external signals was associated with more motivation to get reward for others. Furthermore, interoception was linked with dissimilar patterns for self and other reward responses in the right dAI, while similar patterns of activation in the ACCg were associated with prosocial motivation. These results suggest that representation of rewards in the right AI is attached to bodily signals, whereas reward representation in the ACCg appears to be tied to motivational actions.

Chapter 7

Opportunity cost shapes prosocial behaviour: Prosocial Foraging

7.1 Summary

The survival of many species depends on hunter-gatherers - conspecifics foraging prosocially for rewarding resources for others. Research has suggested that people are generally selfish, and value their own rewards more than others'. However, this work has largely failed to consider theories from behavioural ecology that provide rich frameworks to understand how animals make foraging decisions, and how they are shaped by the availability of rewards in the environment. Thus, we have a poor understanding of how people process rewards for others when faced with crucial hunter-gatherer foraging problems. Here, I leverage ecological theories of how to solve one of the most significant foraging problems: patch-leaving. When foraging in a location (patch), as you collect rewards the resource is depleted, and the rate of (foreground) reward accumulation declines over time. When should you leave and spend some time travelling to find another patch? Theoretically, it is optimal to leave when the foreground rate declines to the average rate of rewards that can be obtained in the environment (background). I developed a social patch-leaving task where participants decided when to leave patches where rewards were accrued at different foreground rates, in two different environments. Crucially, in half of blocks participants foraged for themselves, while in the other half the rewards were given to another person. I found that people were less sensitive to changes in foreground reward rates for others compared to self. However, people did not show differences in background rate sensitivity between beneficiaries. These results highlight that people are not always selfish when it comes to rewards, with the antecedents of prosociality lying in our ability to consider the environment when foraging prosocially.

7.2 Introduction

Across experiments in experimental psychology and neuroeconomics, people commonly manifest prosocial behaviour and are willing to incur self-costs to benefit others (Camerer, 2011; Crockett et al., 2014; Fehr & Camerer, 2007; Hein et al., 2010; Hutcherson et al., 2015; Lockwood, Hamonet, et al., 2017). Nevertheless, people typically do not value rewards for others as much as they do for themselves - they are not willing to equally share rewards with others, try to take advantage of others' generosity to maximise their own rewards, do not learn as quickly how to obtain rewards for others, nor do they exert efforts to gain rewards for others as much as they do for themselves (Camerer, 2011; Fehr & Camerer, 2007; Fehr & Fischbacher, 2004; Lockwood et al., 2016; Lockwood, Hamonet, et al., 2017). Most of these findings come from experiments where participants make binary decisions that result in discrete amounts of rewards for another person or group of people, involving in turn some personal cost for the decider. However, this experimental approach, which has revealed valuable insights about the cognitive and neural mechanisms of reward-based prosocial actions, partially ignores the potential function that these kinds of actions likely have from an ecological perspective. In this chapter, I propose a different experimental framework to research prosocial behaviour based on ecological theories that not only consider the reward being traded, but also how the reward environment dynamically shapes its value (Gabay & Apps, 2020; Hall-McMaster & Luyckx, 2019; Mobbs et al., 2018; Rosati, 2017). In this ecological framework, the value of an action is relative to the opportunity cost of missing a more valuable outcome from the alternative options available.

From an ecological perspective, prosocial behaviour is not especially unique among non-human animals. In fact, many species who live in communities forage for resources not only for themselves, but also for conspecifics - insects such as bees collect pollen to share it with their hive, while mammals such as wolves hunt preys to eat together with the pack (Lemanski et al., 2019; Metz et al., 2011; Vucetich, Peterson, & Waite, 2004;

Woodard & Jha, 2017). Therefore, animals who live in societal systems incur self-costs for the benefit of others. Human prosocial behaviour could be a manifestation from these phylogenetic foraging roots (Hill, 2002). Early and present human hunter-gatherer communities have similar forms for searching, collecting and sharing rewards to other animals (Kelly, 2013; Stibbard-Hawkes, 2019). This evolutionary and ecological function of prosocial behaviour has been overlooked by most behavioural decision-making experiments, which assume that agents simply choose between two options or decide on how much of a certain reward they would share (Mobbs et al., 2018). However, the origins of prosocial behaviour may have been shaped by early activities as hunter-gatherers, and previous research has yet to examine how people make decisions about rewards facing problems of the kind hunter-gatherers encountered.

Behavioural ecology has studied these kinds of problems and formalised the decision process of the agent facing it. One of the most common problems faced by animals, due to how their environments are structured, is the *patch-leaving problem* (Charnov, 1976). When should leave where you are collecting rewards to travel somewhere else to collect them? Rewards, such as food, are not located everywhere, evenly. Rather, they are located in patches (e.g. berries in bushes). To maximise the intake of rewards when foraging, an individual must not only pay attention to the reward available in the current location (i.e. a patch), but also to the other rewards available in the environment. If you are in a lean berry bush, which is not giving you much reward, you should not stay until you have collected them all you need to leave, otherwise you could be missing the opportunity to be collecting rewards somewhere else, at a higher rate. Hence, to maximise their reward accumulation (and consequently their chance of survival), foragers must dynamically and continuously compare the reward rate that they are receiving in the current patch (the foreground reward rate), with the average reward rate available in all other locations in the environment, i.e. the background reward rate.

The optimal solution to this problem is encompassed within marginal value theorem (MVT, Charnov, 1976). MVT provides a quantification of how the different variables in the environment should shape decisions of how long to spend foraging, in order to optimally maximise rewards. Key within this theory is that the optimal forager should leave the current patch when the foreground reward rate equals the background reward rate. Crucially, within MVT the influence of foreground and background reward rates on patch leaving times are independent. Thus, within an environment, high yield patches, with a lot of resources, should be left later than low yield patches with few resources. This is because a well-resourced high yield patch will reach the average rate of their environment later than a low yield one. On the other hand, a forager should stay longer in poor reward environment patches, than in a richer reward environment. This is because the depleting reward rate of a patch will reach the high average reward rate of a rich environment sooner than the average reward rate of a poor one. Therefore, there are two key things that make the patch-leaving problem and MVT model different from those in existing economic games examining prosocial behaviour. First, the focus of this problem is on *when* - how agents allocate time to collecting resources, and when do they choose to leave a certain location in function of the environment, i.e. the opportunity cost of staying where you are. Second, and connected to the previous point, such decisions require the agent to change how they value rewards depending on the quality of the environment.

Human social decisions may also be embedded in patch-leaving problems in certain contexts, such as those made by nomadic tribes. Importantly, previous evidence has shown that animals and humans, both in the wilderness and under laboratory conditions, conform broadly to the predictions made by MVT when foraging rewards (Amano et al., 2006; Cassini, Lichtenstein, Ongay, & Kacelnik, 1993; Cibula & Zimmerman, 1984; Constantino & Daw, 2015; Goulson, 2000; Hassell & Southwood, 1978; Krebs, Ryan, & Charnov, 1974; Campbell Le Heron et al., 2020; Lewis, 1982; Rita et al., 1997; Smith et

al., 1983; Venkataraman, Kraft, Dominy, & Endicott, 2017). But, do people's prosocial decisions conform to such principles? Previous findings about cognitive and neural processing of others' rewards have suggested a self-bias but typically for rewards that will be received instantaneously. It is unknown whether environmental reward rates shape people's decisions differently for others compared with self.

Here, I designed a task based on MVT to examine whether the instantaneous foreground reward rate and the average background reward rate influence leaving decisions when collecting rewards for self differently to when collecting for an anonymous other. In two studies, participants collected rewards from patches where the rate of obtaining rewards continuously depleted. Patches could have either high or low yields, such that it would take more (or less) time to reach the same reward rate. Participants encountered patches across a fixed amount of time in different environments, which could be either rich (high background reward rate) or poor (low background reward rate). In study one, this was defined by the proportion of high vs low yield patches. In study two, it was defined by the time spent travelling between patches (increased time decreases the richness of the environment). Crucially, participants visited environments where the reward was collected for themselves, and environments where the reward was collected for another person. The participants' task was to decide when to leave the current patch and spend some time travelling to the next one, with participants not receiving any rewards while traveling.

Using this design, I hypothesised that, for foreground rewards, participants will display a self-bias in their leaving times, such that they will be more sensitive to the instantaneous rewards (whether it has a high or low reward rate) when collecting rewards for themselves than others, consistent with previous behavioural economic studies (Camerer, 2011; Engel, 2011; Fehr & Camerer, 2007). However, for the effect of the environment, I expected participants to behave similarly for self and other, guiding their leaving decisions according to the opportunity cost of the environment (whether the

environment is rich or poor), consistent with previous studies showing that both animals and humans guide their foraging behaviour according to MVT principles. Finally, I predict that, overall, participants will leave patches at a suboptimal reward rate, consistent with previous studies (Constantino & Daw, 2015; Campbell Le Heron et al., 2020), but they will conform broadly to the predictions of MVT on leaving decisions for high vs low yield patches, and poor vs rich environments.

7.3 Methods

7.3.1 Participants

The study was approved by the University of Oxford ethics committee. Volunteers were recruited through university participant databases, they gave informed consent, and they were paid for their time (£8 per hour, plus £4 on average for an extra bonus based on their decisions). Participants were excluded if they had a history of neurological or neuropsychiatric disorders, psychoactive medication or drug use. Pregnant participants were also excluded from the study, as well as those who had participated in studies involving social interaction and who had studied psychology, addressing concerns that prior experience could influence psychological processes in the study.

For study 1, 45 participants were recruited. From this sample, one participant did not complete the entire task, one suffered technical issues while completing the task, two expressed doubts about the social manipulation, and one did not show engagement when collecting rewards for others (leaving times < 500 milliseconds across patches, see below). This resulted in a total sample of 40 participants for study 1 (age M=24.8, SD= 5.18, 26 females).

For study 2, data were collected from 31 volunteers who participated in the study. However, two participants failed to follow instructions and leaving time behaviour was

either very short (<500ms) or long (>3 SD from the sample average) and thus were excluded. Therefore, the final sample in study 2 was 29 (age M=24.7, SD=5, 16 females)

7.3.2 General Procedure

Participants arrived at the lab in one of the facilities of the Experimental Psychology Department, where they answered questions and completed the experimental task using a desktop computer. First, they were given a general overview of the experiment. Participants were told that they were one of two participants in the study to complete a decision-making task. However, the characteristics of the task would vary depending on the result of a role assignment procedure, which would determine who would complete a task involving bonuses according to their performance, or a task that did not involve extra money. Next, once participants answered some demographic questions, the social manipulation explained in detail in Chapter 2 occurred. Participants were randomly assigned one of two roles – Decider and Receiver. In reality, participants were always assigned the role of deciders, as the receiver was a confederate. Participants were told that, as deciders, they would complete a decision-making task where they would collect rewards not only for themselves, but also for the receiver. They were also told that the receiver would complete a different decision-making task that would not involve financial bonuses for anyone. Finally, participants were instructed that their decisions were confidential and anonymous. Once participants finished the task, they answered some debriefing questions about the social manipulation setting and how they made their decisions in the social foraging task. Before leaving, participants received their payment for their participation plus a monetary bonus for their decisions.

7.3.3 The Social Foraging Task.

7.3.3.1 Experimental Design

Participants undertook a computer-based task that emulates a patch-leaving problem (Fig. 7.1), deciding when to leave from the current patch to the next. The task was designed following the principles of MVT (Charnov, 1976), and based on the task used by Le Heron and colleagues (2020), which examined the effect of foreground and background reward rates on decisions to leave when collecting rewards but only for self. In this version, participants were accruing rewards for either themselves or an anonymous other person, in rich or poor environments, with patches of high or low yield.

The social foraging task was framed as a farming game, where participants had to collect berries (reward) on sequentially encountered fields (patches) within a certain farm (reward environment). Participants were told that the berries they accumulated in the experiment would serve as a monetary bonus for themselves and the receiver. There were two berry farms: raspberry and blueberry, where they collected berries to obtain money for themselves and for the receiver, respectively. Crucially, there were also two qualities of farms/environments. The quality of the farms was depicted to participants with coloured border in the screens. In gold farms, the background reward rate was higher than in green farms.

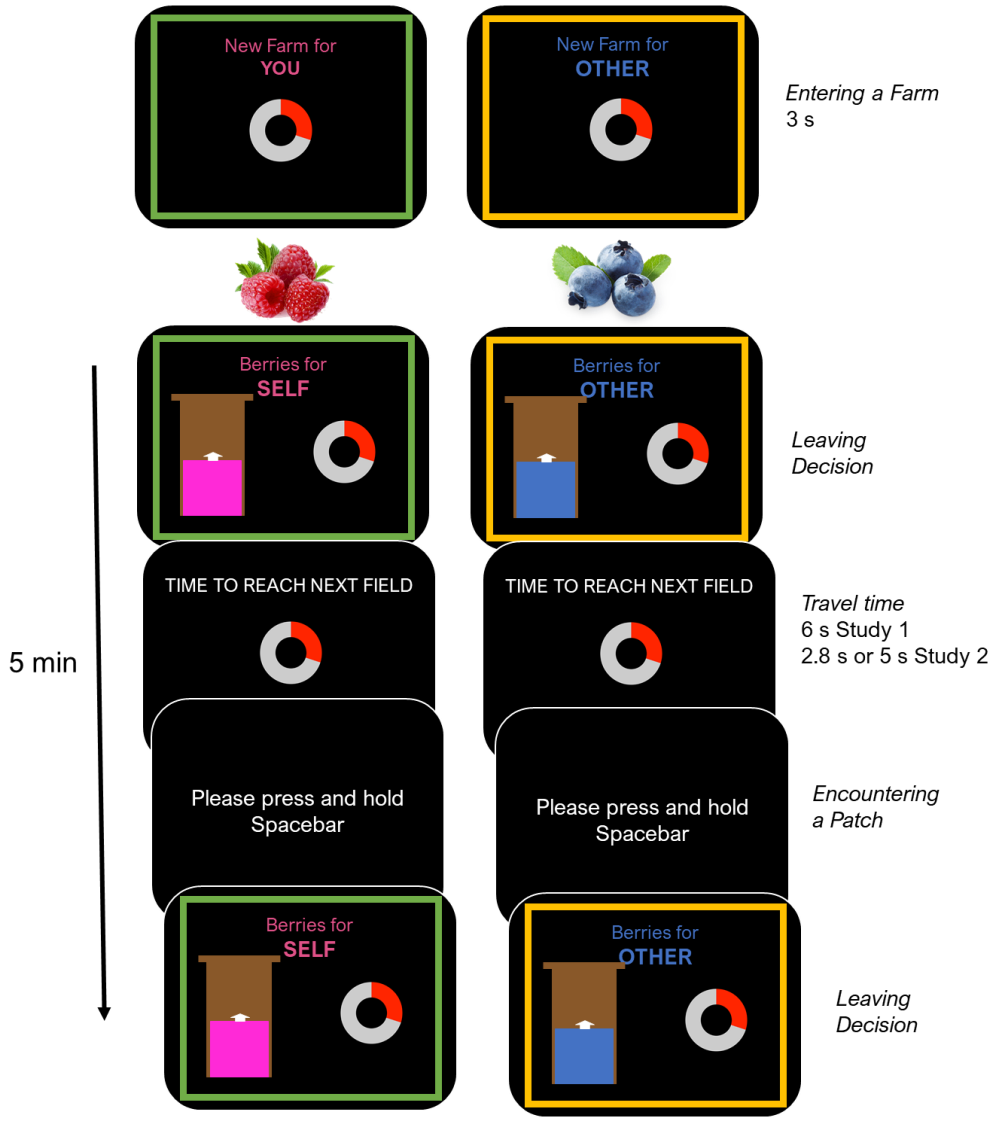


Figure 7.1. Social Foraging Task. Participants visited different types of farms (environments) for 5 minutes. When entering a farm, a screen indicating the quality of the farm (gold or green squares) and the beneficiary (raspberry of blueberry) appeared for 3 seconds. After this screen, participant spent 5 min in the farm, deciding how long to stay in a current field (patch) in which berries (reward) were collected at an exponentially decreasing rate. The reward rate was illustrated by a brown basket being continuously filled by either raspberries or blueberries. Once a decision to leave was made, participants incurred a travel time cost (6s in study 1, variable time in study 2) during which they did not collect rewards. The instantaneous rate of bucket filling indicated the foreground reward rate, while the coloured frame indicated the background reward rate. Crucially, in raspberry farms (left), participants collected rewards for themselves, while in blueberry farms (right) they did for another unknown person. Participants knew about their limited 5 min for each environment, but no cues were shown about the time elapsed, neither cues about their accumulated reward within a farm nor in the whole experiment. Following a decision to leave, a clock ticking down the travel time was presented.

In a block design, participants visited each farm - two qualities (gold/green) x two berries (raspberries/blueberries) = four - for a fixed amount of time of five minutes in a randomised order, where participants collected berries from the fields they encountered. On each patch/field, the rate of reward accumulation depleted at an exponential rate. They could leave the patch to travel to a new one at any time. However, moving from one patch to another had a time-cost associated, a lapse of time where participants did not collect any reward. The number of patches participants could potentially visit was unlimited, but was constrained by the amount of time participants could spend in each specific farm (i.e. 5m). Participants were given instructions outlining the details above, and told that their goal was deciding when to leave the current patch to move to the next one, without any further information. This differs from previous foraging experiments where participants are instructed to maximise rewards (e.g. Le Heron et al., 2020). However, it ensured comparability with other prosocial decision-making studies, where pursuit of reward for other depends on participants' preferences and not on the instruction to maximise others' reward.

When a participant was in a patch, its instantaneous, foreground reward rate decreased exponentially. There were two types of patches depending on their foreground rate (Fig. 7.2A). In high yield patches, rewards had a higher initial rate, but decayed in the same manner, meaning that the rate would remain higher for longer than in low yield patches. These two types of patches were defined by the following patch reward function:

$$g(T) = S * e^{-1.1 * T}$$

Where *g* is the foreground reward rate after *T* seconds. Patches varied in the scaling factor of the reward function *S*. Thus, the high yield patch had an S value of 57.5, and the low yield patch had a value of 34.5. This foreground reward rate was represented by a basket continuously filled with berries while participants were in the patch. The level of the berries in the basket was proportional to the integral of the patch reward function between time zero and T, being updated with a frequency of 20Hz. Even though

participants knew that there were two patches across the experiment, they were not explicitly told the quality of the patch they were collecting berries from. Rather, they had to deduce it from the rate at which the basket was getting filled.

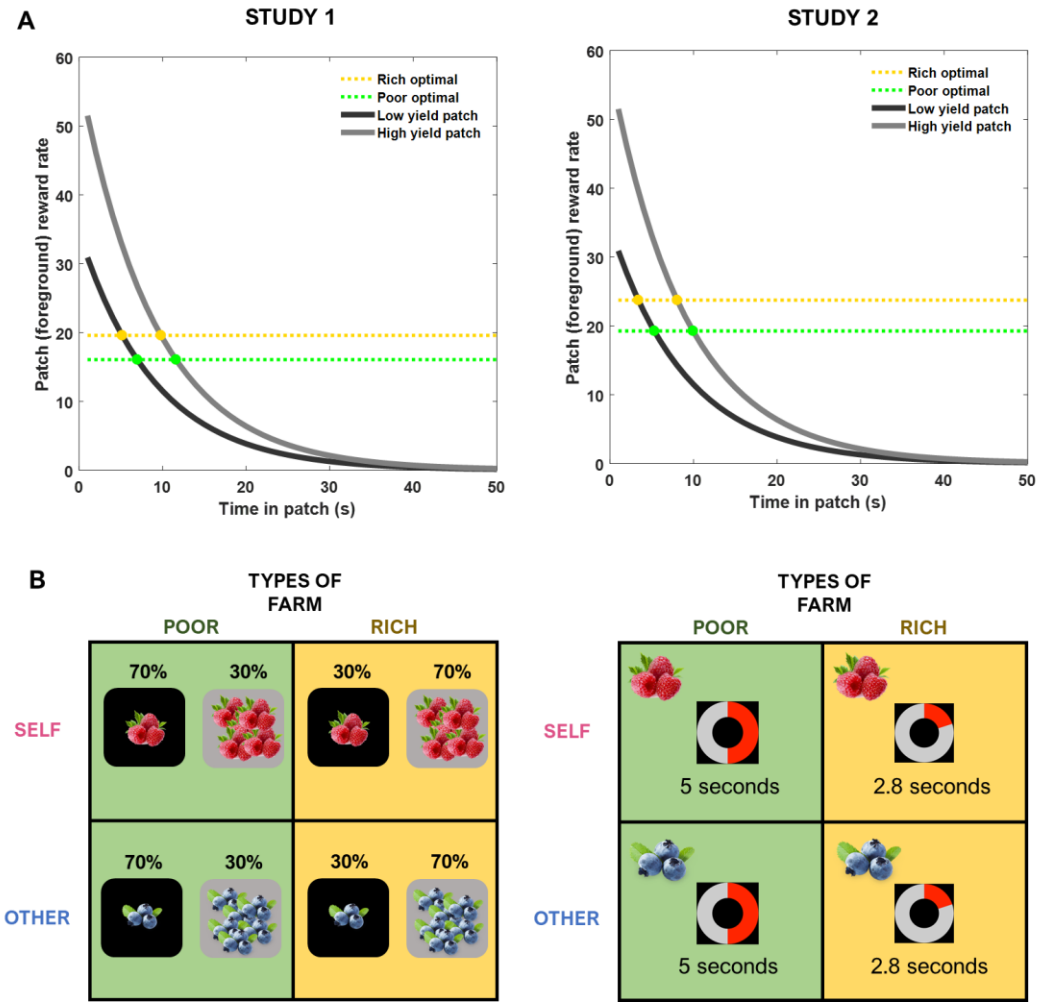


Figure 7.2. Background reward manipulation in studies 1 and 2. A. Continuously decreasing reward rate functions for the high and low yield patch (grey and black lines, respectively), as well as background reward rates for rich and poor environments (gold and green dotted line respectively) are shown. Gold and green dots depict optimal leaving times for all environment and patch yield combinations as predicted by MVT. Importantly, the environment and patch effects are independent from each other. *Left*: Study 1. *Right*, Study 2. B. Participants visited four types of farms in a 2 (beneficiary: raspberry/self, blueberry/other) x 2 (quality of environment: rich/gold, green/poor) blocked design. The gold farm had a higher background reward rate than the green farm. *Left*: In study 1, the background reward rate of the environment was manipulated through the relative proportions of high and low yield patches. The rich environment had a higher proportion of high-yield, and a lower proportion of low-yield patches, while the opposite was true for the poor environment. *Right*: In study 2, the background reward rate of the environment was manipulated through travel costs, such that the rich environment had shorter travel times (2.8 s) than the poor environment (5 s). High and low-yield patches were evenly distributed in study 2 farms.

7.3.3.2 Procedure

Before starting with the task, participants were provided instructions and performed practice trials to familiarise themselves with the kind of decisions they would perform. In this practice set, they made leaving-decisions for eight patches, with half being high yield, and the other half low yield patches. Different from the real experiment (where participants spent five minutes per farm), in these practice trials each patch was in a single farm, meaning that every time a participant left a patch, they also left a farm. Thus, they visited the four types of farms twice, with each containing only one patch. The order of these practice trials was pseudorandomised. After the practice session, participants were asked to summarise the characteristics of the experiment verbally to check whether they understood the different aspects of the task (environments/farms, beneficiaries/berries and patches/fields). Next, participants completed the social foraging task.

Each block of the task lasted 20 minutes, in which participants visited the four types of farms in a random order. Each farm started with a screen lasting 3 seconds giving information about the quality of the farm (gold/green) and the recipient of the berries (raspberry-self/blueberry-other, Fig. 7.1A). This screen only appeared once when entering a farm. After this initial screen, participants were prompted to press and hold the spacebar to automatically find a field and collect berries. Until participants released the spacebar, they were continuously collecting berries determined by the yield of the patch. The amount and type of berries collected per field was shown by a continuously growing pink (raspberry) or blue (blueberry) bar that represented a continuously growing pile of berries in the participant's basket. The rate of berries being collected slowed down exponentially as the patches resources depleted. When the participant decided to leave the field and released the spacebar, they had a variable time cost of a few seconds in which they 'walked to the next berry field'. During travel time, a screen with a timer was displayed. After the timer had run down, a new screen told participants that they now had

arrived at the next field and asked them to press and hold the space bar until they wished to leave this field. Participants did this during the five minutes that a farm lasted. After this time, a screen lasting three seconds informed them that the time had ended, followed by the initial screen informing them about the new fam quality and the beneficiary of the berries. Participants then entered a new farm and performed the patch-leaving decisions.

7.3.3.3 Study 1 – Manipulation of environmental richness by patch distribution

Participants completed two blocks of the experiment, having a short break in between. The quality of the farms, i.e. the background reward rate, was manipulated by varying the distribution of high and low yield patches (Fig 7.2B). Thus, for the rich, gold farm, 70% of the patches were high yield, and 30% low yield. The opposite was true for the poor, green farm, with 30% of high yield patches and 70% of low yield ones. In both blocks, the travel time was fixed at six seconds. The order of the patches within a farm was pseudorandomised – for instance, in a rich farm, in every series of 10 patches, seven were high and three low yield patches. Furthermore, the low yield patches were distributed so that no more than two appeared in a series of five patches. Patches were then randomised under these constraints.

7.3.3.4 Study 2 – Manipulation of environmental richness by travel time

Participants completed three blocks of the experiment, with short breaks after the first and second blocks. Unlike study one, patches were equally distributed across farms in a pseudorandomised order, such that in every series of 10 patches, half were high, and half were low yield patches in a random sequence. The background reward rate (the quality of the farm) was defined by the duration of the travel time associated with moving from one patch to another. Thus, in the rich, gold farm, the time needed to find another field was 2.8 seconds, while in the poor, green farm the travel time was fixed at 5 seconds. Higher

travel times make the background reward rate decreases, as there are longer periods without receiving rewards when leaving a patch. Differences in travel time were framed as distance between fields, i.e. fields in gold farms were closer to each other compared with those in green farms. Note that this reduction in travel times (5 s and 2.8 s in study 2 vs 6 s in study 1) increases the background reward rate of study 2 relative to study 1 farms. This predicts, according to MVT principles (see below), lower patch residency times for study 2 (Fig. 7.2A), increasing therefore trial sampling across all conditions.

7.3.4 Optimal predictions using the Marginal Value Theorem

The MVT formalises the patch-leaving problem of an agent foraging for rewards in a patchy environment, considering foreground and background reward rates, which depend on the energetic costs of foraging. Thus, MVT provides a solution that maximises the energy intake of the agent. MVT assumes that patches deplete over time, which is a natural property of many resources (the more that is consumed the less there is available and thus the slower your intake will be). Figure 7.2A illustrates how the high (grey line) and low (black line) yield patches decreases with time following the patch reward function provided above (*g(T)*). To calculate optimal leaving times, MVT uses the following formula for a patch type *i* to compute the background reward rate of the environment *R*:

$$R = \frac{P_i g_i(T_i) - k_i}{d_i + P_i T_i}$$

Where *P* is the probability to encounter patch *i*, that has a patch reward function of *g(T)*, *k* is the energetic cost of foraging, *d* is the travel time, and *T* is the patch residency time. Differentiating this formula relative to time and rearranging its elements, shows that R is maximised when leaving times meet the following:

$$g'_i(T_i) = \frac{P_i g_i(T_i)}{d_i + P_i T_i}$$

Where the left part of the equation corresponds to the foreground reward rate, while the right part is the background reward rate. Notice that the equation above does not contain the parameter k, as in these experiments foraging did not have energetic costs associated. Therefore, according to MVT, the optimal leaving time is when instantaneous foreground reward rate equals the average background reward rate of the environment.

Using MVT, studies 1 and 2 were designed considering optimal leaving times *a priori*, according to the parameters used in each experiment. In the Figure 7.2A, green dots correspond to the leaving times for studies 1 (left) and 2 (right) in the poor environment, while gold dots are for the gold environment. Green and gold dashed lines correspond to the reward rate in which is optimal to leave the poor and the gold environments respectively. Thus, longer patch residency time is predicted for high than low yield patches regardless of the environment, as the former reach the background reward rate later than the latter. This effect (difference in optimal leaving times between high and low yield patches) was set at 5 seconds across studies 1 and 2. Furthermore, MVT predicts a longer residency time in the poor, green environment than in the rich, gold environment irrespective of the patch. The environment effect (difference in optimal leaving times between poor and rich environment) was set in 2 seconds across studies. Importantly, notice that the reward rate at the time of leaving is the same for high and low yield patches within an environment, as at the optimal time of leaving a patch its reward rate equals the reward rate of the environment. Therefore, for statistical analyses, the main indexes used were leaving times and reward rate at time of leaving, as they reflect the main behavioural outcomes in a patch-leaving problem.

7.3.5 Statistical Analyses

Linear mixed-effects models (MMs) were applied as the primary analysis tool. In this model, for a trial *i*, the continuous variable patch leaving time *LT* is predicted by the fixed effects of type of patch *P*, environment *E* and beneficiary *B*, such that:

$$LT_i = \beta_{0j[i]} + \beta_{1j[i]}P_i + \beta_{2j[i]}E_i + \beta_3 B_i + \beta_4 P_i B_i + \beta_5 E_i B_i + + \beta_6 E_i P_i + \beta_7 P_i E_i B_i$$

This model has a random intercept clustered in participants *j*, crucial for this kind of data as participants have different number of trials per condition depending on how long they stay in each patch. Previous studies have shown that participants vary in their sensitivities to foreground (patch) and background (environment) reward rate (Constantino & Daw, 2015; Kolling, Behrens, Mars, & Rushworth, 2012; Campbell Le Heron et al., 2020; Wittmann et al., 2016), although both types of reward information do not interact. Thus, random slopes for patch and environment were added to the model. Comparing the AIC of a model with only random-intercept and one with random-slopes added reveals that the latter outperforms the former (AIC for study 1, random-intercept only = 29747, random-slopes = 29393; study 2, random-intercept only = 38408, random-slopes = 37408). Furthermore, a similar model to that outlined above was also used to predict reward rate at the time of leaving instead of leaving times, aiming to clarify how participants made decisions in relation with the optimal solution predicted by MVT. Finally, to contrast participants' average leaving times with MVT predictions, Wilcoxon signed-rank test were used as this data was not normally distributed.

For the above analyses, I excluded those trials per participant with leaving times above 3 standard deviations from the mean of that subject, to avoid biases effects of outlying data (< 0.8% of the total trials on average across studies), particularly important as the number of trials differs per participant in each condition. Furthermore, trials with leaving times below 500 milliseconds were also excluded, as these were unlikely to be intentional choices as the patches yield was unlikely to be inferred by that point.

Sample size for study 2 was estimated considering the mean and the standard deviation of the number of trials of each participant per condition in study 1, together with the random intercept and the fixed effects of the MM on leaving times. With these parameters, I simulated leaving times 2000 times, with half (i.e.1000) simulating data for 25 participants, and the other half simulating data for 30. For each simulation, a MM similar to the one described above was run for main and interaction effects. Thus, power was defined as the number of times the effect was significant ($p < 0.05$) divided by the total number of simulations. Using these parameters, a sample between 25 and 30 participants was defined sufficient to capture main and interaction effects of study 1 with a power of > 0.8.

7.4 Results

7.4.1 Study 1

The main behavioural indexes in the social foraging task were leaving times and reward rate at time of leaving a patch. However, total reward accrued in self and other conditions can also shed light on behavioural differences in reward collection. Yet, differences in leaving times do not necessarily translate into significant differences in total reward accrued in this experiment. If participants overstay in patches, the reward gain over time is minimal, i.e. higher differences in reward occur early in the patch as it is evident in Figure 7.2A. Thus, people in study 1 collected slightly more rewards for others (arbitrary units a.u. M = 18,712.5, SEM = 418.2) than self (M = 18,501.8, SEM = 418.7), but this difference was not significant (Wilcoxon Signed-Rank Test, z = -1.48, p = 0.14).

7.4.1.1 Foreground but not background reward rate sensitivity is different between collecting rewards for self and other.

According to MVT, foreground and background reward rates should have independent effects on participants' leaving times. Thus, participants should leave earlier in a rich than a poor environment across patches, and they should leave later high than low yield patches across environments. I tested whether participants were equally sensitive to foreground and background reward rates when they collected rewards for others compared with themselves. I used MMs predicting participants' leaving times by type of patch, environment, beneficiary, and their interaction. This analysis revealed a main effect of patch ($b = 4.9$, $SEM = 0.45$, $t = 10.85$, $df = 73$, $p < 0.001$) and environment ($b = 1.44$, $SEM = 0.46$, $t = 3.1$, $df = 69.6$, $p < 0.003$), but not an interaction between reward rates, in line with MVT predictions. Crucially, there was a patch x beneficiary interaction ($b = -1.37$, $SEM = 0.39$, $t = -3.53$, $df = 4898.8$, $p < 0.001$, Fig. 7.3A), indicating that the patch sensitivity changed depending on who was the beneficiary of the reward. Post-hoc analysis revealed that the differences between high and low yield patches, although significant on both self ($t = -11.76$, $df = 49.9$, $p < 0.0001$) and other ($t = -8.24$, $df = 48.8$, $p < 0.0001$) trials, were significantly stronger when collecting rewards for participants themselves than for the receiver ($t = -5.35$, $df = 4901$, $p < 0.001$). Furthermore, leaving times for high ($t = -10.66$, $df = 4895.4$, $p < 0.001$) and low ($t = -3.12$, $df = 4899.7$, $p < 0.01$) yield patches were significantly higher for self than other. Importantly, there was no significant interaction between beneficiary and environment, meaning that background reward rate shaped decisions across conditions. This was confirmed by a post hoc analysis showing that leaving times were significantly different between rich and poor environments for self ($t = -3.12$, $df = 49$, $p < 0.004$) and other ($t = -2.39$, $df = 48$, $p < 0.03$).

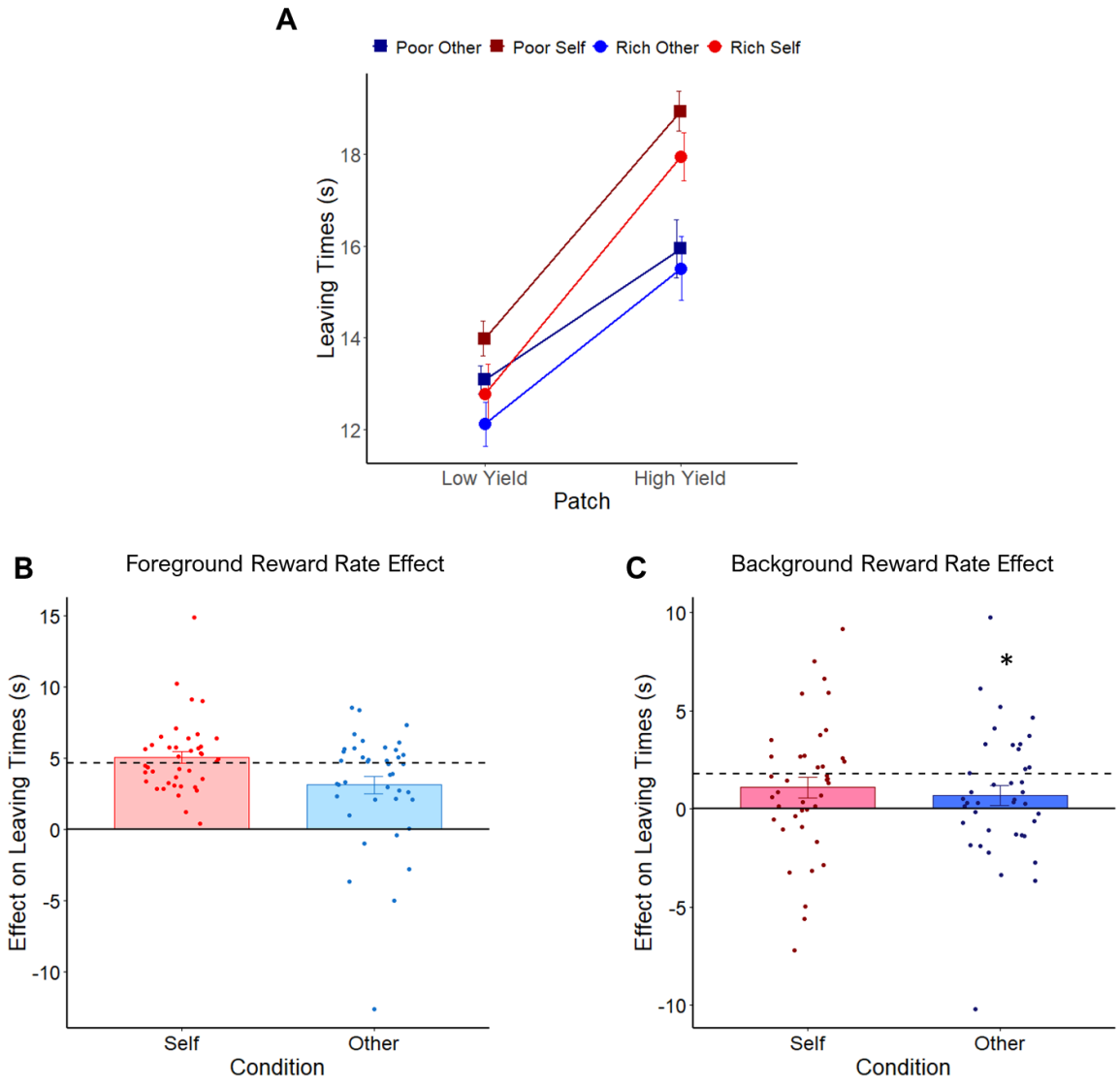


Figure 7.3. Leaving times results of study 1. A. Average leaving times (y-axis) for each condition in the experiment. Red lines correspond to *self* environments while blue lines are *other* environments, including rich (lighter circles) and poor (darker squares) qualities. Participants stayed longer in high than low yield patches (x-axis), following MVT predictions. However, this effect was stronger when collecting rewards for self than other. Furthermore, participants left the rich environment earlier than the poor one. Crucially, there was no environment x beneficiary interaction, suggesting that this effect was not modulated by the beneficiary of the reward. B. Foreground reward rate effects on leaving times. Mean leaving times were subtracted between high and low yield patches and compared with optimal predictions. Participants showed no significant differences between sensitivity to foreground reward rate (y-axis) and MVT predictions (dotted line) for both self and other. C. Background reward rate effects on leaving times. Mean leaving times were subtracted between poor and rich environments and compared with optimal predictions. Participants showed significant differences between sensitivity to background reward rate and MVT predictions for other but not self. * $p < 0.05$. Error bars depict SEM

7.4.1.2 Participants conform broadly to optimal foraging when collecting rewards for self but no for other

Participants showed foreground and background reward rate sensitivity in the direction predicted by MVT, i.e. longer leaving times for high relative to low yield patches, and for poor relative to the rich environment. However, participants overstayed in all patches and environments - leaving times for high and low yield patches in rich and poor environments, for both self and other, deviated significantly from what the optimal solution predicted (Wilcoxon signed-rank test, $p < 0.001$ for all contrasts). For self, participants deviated about 7 seconds from the optimal solutions for patches and environments, while 6 seconds was the deviation when they collected rewards for others.

This 'stickiness' in decisions to leave is consistent with previous results in humans and other animals (Constantino & Daw, 2015; Hayden, Pearson, & Platt, 2011; Campbell Le Heron et al., 2020; Nonacs, 2001; Stephens, 2008). Importantly, previous results have also shown that once controlling for this overstaying, by subtracting mean leaving times and calculating the magnitude of the effects between different types of patches and environments, people are close to optimal. Therefore, I calculated the mean leaving time for each condition, and subtracted the patches (high – low yield patches) and environments (poor – rich) mean leaving times separately for self and other to estimate the magnitude of the effects of the changes in foreground and the background reward rates, respectively. Next, I tested whether these magnitudes conformed with MVT predictions, using one-sample Wilcoxon signed-ranked tests against the optimal effect (dotted line in Fig. 7.3B-C). These analyses revealed that foreground reward rate effect was not significantly different from the optimal solution for self ($z = 0.51$, $p > 0.6$) and other ($z = -1.88$, $p < 0.06$, Fig 7.3B), although the latter deviated more from MVT predictions than the former, showing a non-significant trend. However, the background reward rate effect conformed to optimal solution only when collecting rewards for self ($z =$

-1.47, $p > 0.1$) but not for other ($z = -2.38$, $p < 0.02$, Fig. 7.3C). Thus, people conformed to MVT predictions mainly when collecting rewards for themselves but not for others.

So far, I have focused mainly on leaving times. However, MVT also makes predictions about the foreground reward rate at the time of leaving a patch. Within an environment, the optimal solution predicts that participants should leave the patch when its foreground reward rate reaches the background reward rate of the environment. This implies that, for a given environment, foragers should leave any patch at the same reward rate. Thus, I performed a MM similar to that outlined above but instead of predicting leaving times, the outcome was the foreground reward rate at the time when participants left the patch. Results of this analysis revealed an interaction between patch and beneficiary ($b = 1.93$, SEM = 0.41, $t = 4.7$, $df = 4894.5$, $p < 0.001$, Fig. 7.4A). A post-hoc analysis showed that high and low yield patches had significantly different foreground reward rates at time of leaving when collecting rewards for other ($t = 3.42$, $df = 44.2$, $p < 0.005$), but not for self. The model also reveals a main effect of environment ($b = -1.88$, SEM = 0.59, $t = -3.21$, $df = 59.1$, $p < 0.003$), suggesting that participants adjusted their leaving times considering the background reward rate across conditions. Finally, an environment x patch interaction was also found ($b = 0.9$, SEM = 0.41, $t = 2.2$, $df = 4888.7$, $p < 0.03$), but these results are likely to be driven by the great modulation that beneficiary had on patch effects.

It is important to note that the analysis above assumes that people were perfectly accurate in estimating the background reward rate. However, in this experiment participants influence the reward rate they experience, because they are free to leave one patch for another and thus two participants with different leaving times will actually experience different background reward rates over time. Thus, participants could have conformed to an optimal solution predicted from their own history with the environment, and not with the objective optimal overall given the environment. Therefore, I calculated each participant's long-run obtained reward rate per environment for each beneficiary.

Next, I estimated the optimal leaving times according to these values. Participants' average obtained background reward rate was significantly lower than the maximum available according to MVT. This was true for both self and other in rich (both: $z = -5.51$, $p < 0.001$; Fig. 7.4B), and poor environments (both: $z = -5.5$, $p < 0.001$; Fig. 7.4C). Adjusted optimal leaving times in the rich environment were, on average, 1.6 seconds later for self, and 1.4 for other. In the poor environments, adjusted optimal leaving times were 1.4 seconds later for self and other. Despite these adjustments, participants showed significantly later leaving times for all conditions (Wilcoxon sign-rank test, $p < 0.001$). Therefore, the overstaying bias on leaving times cannot be simply attributed to a difference in the actual background reward rate obtained by participants.

In summary, in study 1 participants distinguished foreground and background reward rates in the direction predicted by MVT for both self and other. However, the effect of foreground reward rate was stronger when collecting rewards for self than others. Furthermore, when participants collected rewards for themselves, they broadly conformed with MVT predictions when controlling for overstay bias. However, this seems not to be the case when collecting rewards for others. Despite participants having the same directionality in the effects predicted by MVT, they deviate significantly from the optimal solution in the social condition.

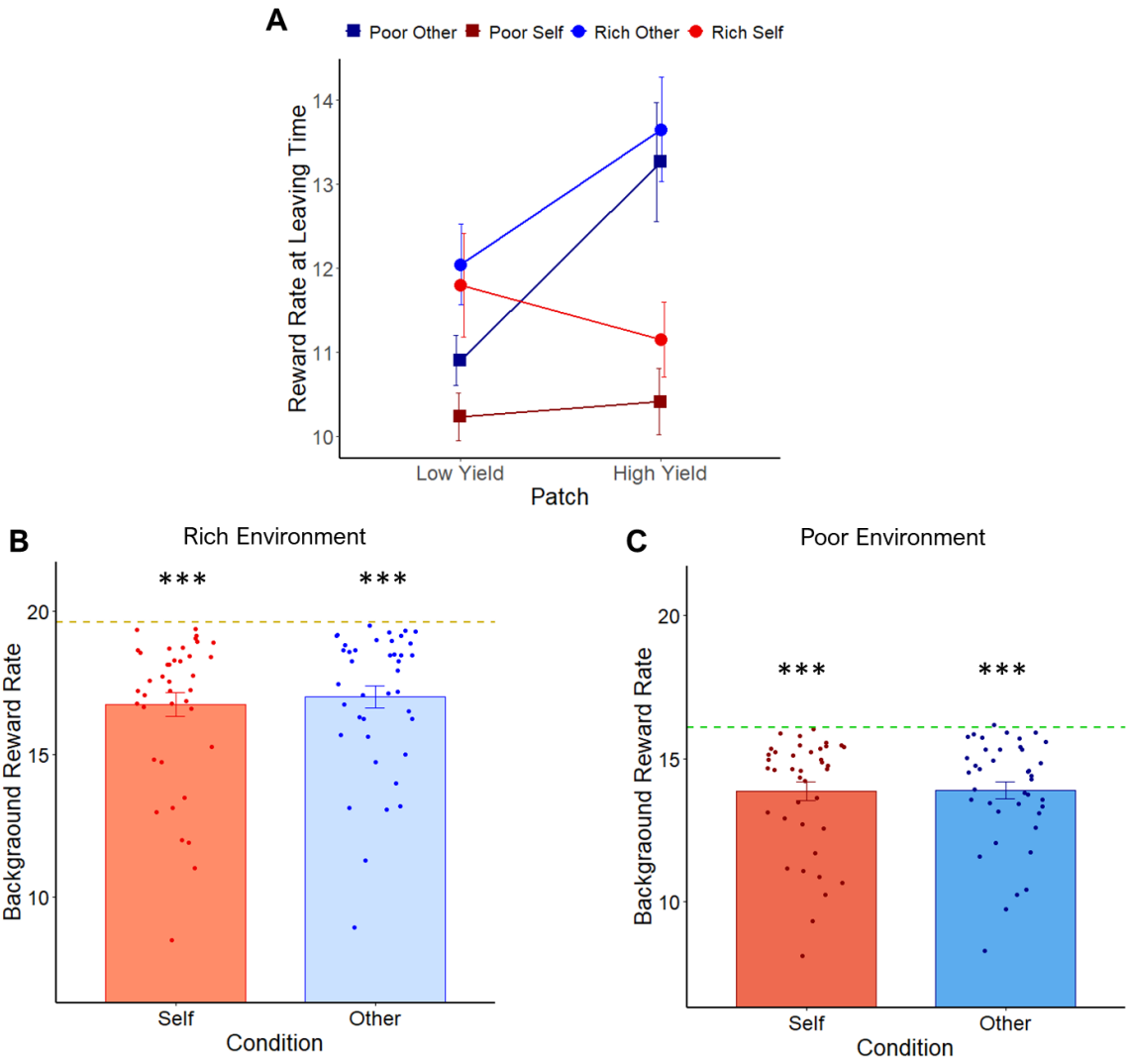


Figure 7.4. Study 1 results on obtained reward rate. A. MVT predicts that the foreground reward rate (y-axis) at which participants choose to leave each patch should not vary between patches, only changing as a function of the environment. Participants showed this effect only for self, but not for other. When collecting rewards for another person, the reward rate at leaving time was influenced not only by the environment, but also by the type of patch. Red lines correspond to *self* environments while blue lines are *other* environments, including rich (lighter circles) and poor (darker squares) qualities. B. Obtained long-run background reward rates in the rich environment for self and other were significantly lower than the actual background reward rate in the rich environment (gold dotted line). C. Obtained long-run background reward rates in the poor environment for self and other were significantly lower than the actual background reward rate in the poor environment (green dotted line). *** $p < 0.001$. Error bars depict SEM.

7.4.2 Study 2

In study 2, I aimed to conceptually replicate the results of study 1 but introduced a different manipulation of the environment. In study 2, the rich environment is the one where berry bushes are close to each other, while in the poor environment bushes are spread further apart. In this way, the forager spends less time travelling from patch to patch in the rich than in the poor environment, increasing the background reward rate of the former relative to the latter. In contrast to study 1, the distribution of high and low yield patches was now the same across environments, with 50% of each. All the other features of the experiment remained the same to those in study 1. Importantly, and in contrast with study 1, participants collected significantly more rewards in total for self (a.u. M = 34,695.2, SEM = 519.8) than others (M = 33,998.1, SEM = 558.3; z = - 2.2, $p < 0.03$) albeit weakly, suggesting that differences in optimality in this study impacted the reward accrued by participants for each condition

7.4.2.1 Sensitivity to foreground, but not background, reward rate is lower when foraging for other vs self rewards

A MM was performed on leaving times with types of patch, environment, beneficiary and their interaction as predictors. This analysis revealed similar results to those found in study 1. Thus, main effects of patch (b = 4.64, SEM = 0.36, t = 12.99, df = 34.7, $p < 0.001$) and environment (b = 2.12, SEM = 0.41, t = 5.15, df = 33.7, $p < 0.001$), but not their interaction, were found, following the same direction of what is predicted by the optimal solution. Crucially, the model also showed a beneficiary x patch interaction (b = -1.26, SEM = 0.19, t = -6.69, df = 7236, $p < 0.001$, Fig. 7.5A). A post-hoc analysis confirmed that the difference on leaving times between patches was significantly stronger when collecting reward for self than other (t = -8.48, df = 7238, $p < 0.001$), although on both beneficiaries the patch effect was present. Importantly, participants stayed significantly longer in high yield patches for self than other (t = -10.93, df = 7236.5, p <

0.001), but not in low yield patches. These results suggest that lower sensitivity to foreground reward rate when collecting rewards for others is triggered by differences in how participants perceived instantaneous reward in high yield patches.

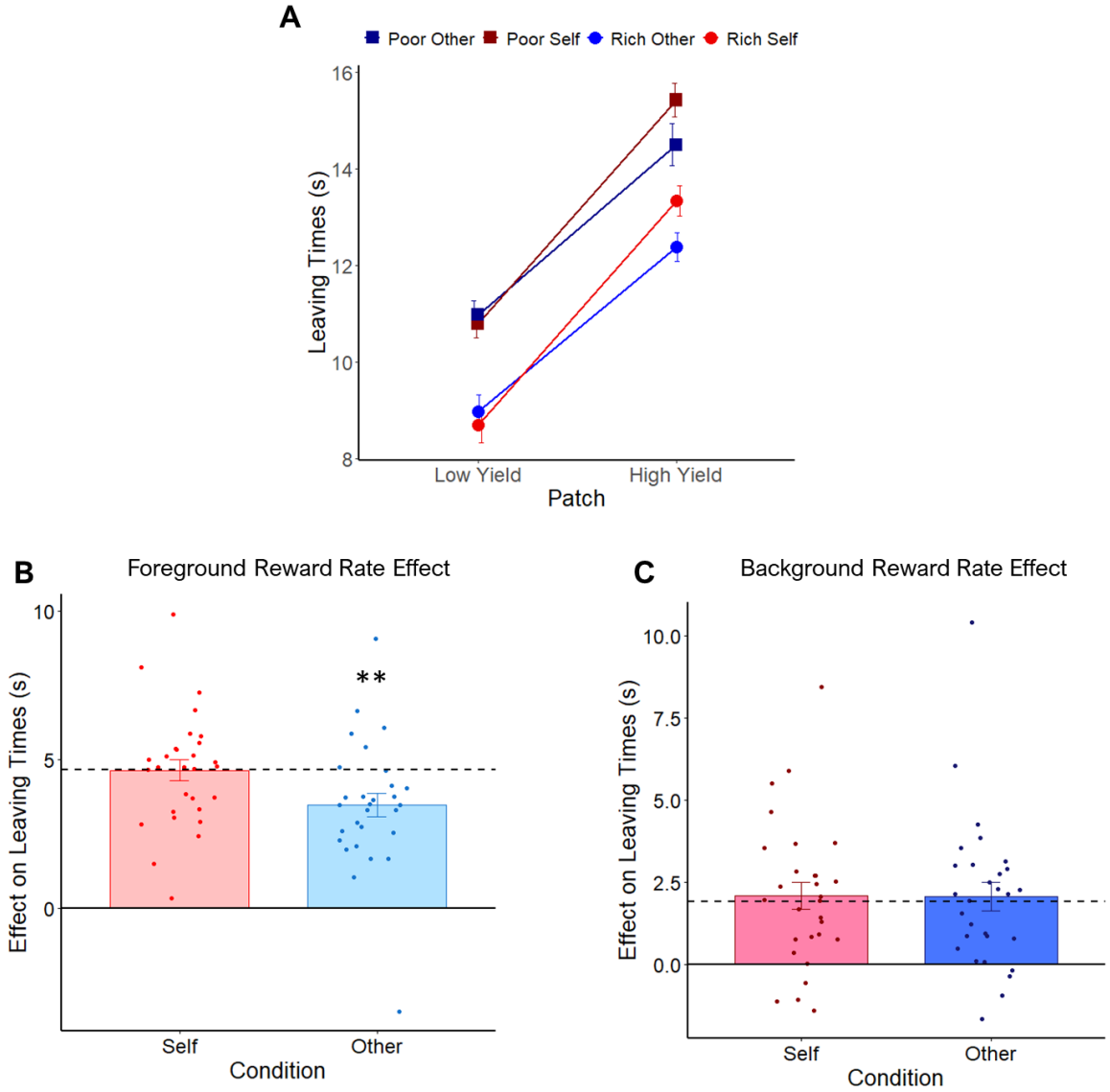


Figure 7.5. Leaving times results of study 2. A. Leaving times (y-axis) for each condition in study 2. Replicating results of study 1, participants stayed longer in high than low yield patches (x-axis), effect that was stronger when collecting rewards for self than other. These results seem to be driven by a significant difference on leaving times for high yield patches. Participants left sooner the rich than the poor environment across beneficiaries. Red lines correspond to *self* environments while blue lines are *other* environments, including rich (lighter circles) and poor (darker squares) qualities. B. Foreground reward rate effects on leaving times. Participants showed significant deviation from optimality (dotted line) when collecting rewards for others but not self. C. Background reward rate effects on leaving times. Participants showed no significant difference between sensitivity to background reward rate and MVT predictions for both self and other. ** $p < 0.01$. Error bars depict SEM.

7.4.2.2 Participants conform broadly to optimal foraging for self but not other.

Participants stayed significantly longer than the optimal solution across conditions ($p < 0.001$), displaying a similar overstay bias to study 1. Participants stayed 7.5 seconds more for each environment and patch than optimal prediction when collecting rewards for self, and 6 seconds when collecting rewards for others. I controlled for the 'stickiness' in leaving decisions looking at the magnitude of the differences between foreground and background reward rates, and tested for optimality. Following this approach, participants were optimally sensitive to variation in foreground reward rates when collecting rewards for self ($z = -0.03$, $p > 0.9$), but not for others ($z = -2.9$, $p < 0.004$, Fig. 7.5B). Differently from study 1, in study 2 the deviation from optimality in foreground reward rate when collecting rewards for others reached statistical significance. Finally, participants also showed an optimal adjustment in their leaving times to variations in the background reward rate for both self and other ($p > 0.8$, Fig. 7.5C). The latter differed from study 1 as the previous results revealed a suboptimal environment adaptation when collecting rewards for others.

Next, I looked at the reward rate at time of leaving to examine whether participants left optimally each patch at the same rate within and environment for self and other. A MM with type of patch, environment, beneficiary and their interactions as predictors revealed a main effect of environment ($b = -2.59$, $SEM = 0.47$, $t = -5.52$, $df = 42.4$, $p < 0.001$), but not patch ($p > 0.5$), in line with MVT predictions. However, and consistent with study 1, an interaction between beneficiary and patch was found ($b = 2.53$, $SEM = 0.3$, $t = 8.4$, $df = 7238$, $p < 0.001$, Fig. 7.6A). A post-hoc analysis revealed that, when collecting rewards for others, participants left high and low yield patches at a significantly different foreground reward rate ($t = 4.58$, $df = 30.1$, $p < 0.001$). This was not the case for self trials. This adds evidence that participants deviated from the optimal solution when collecting rewards for others.

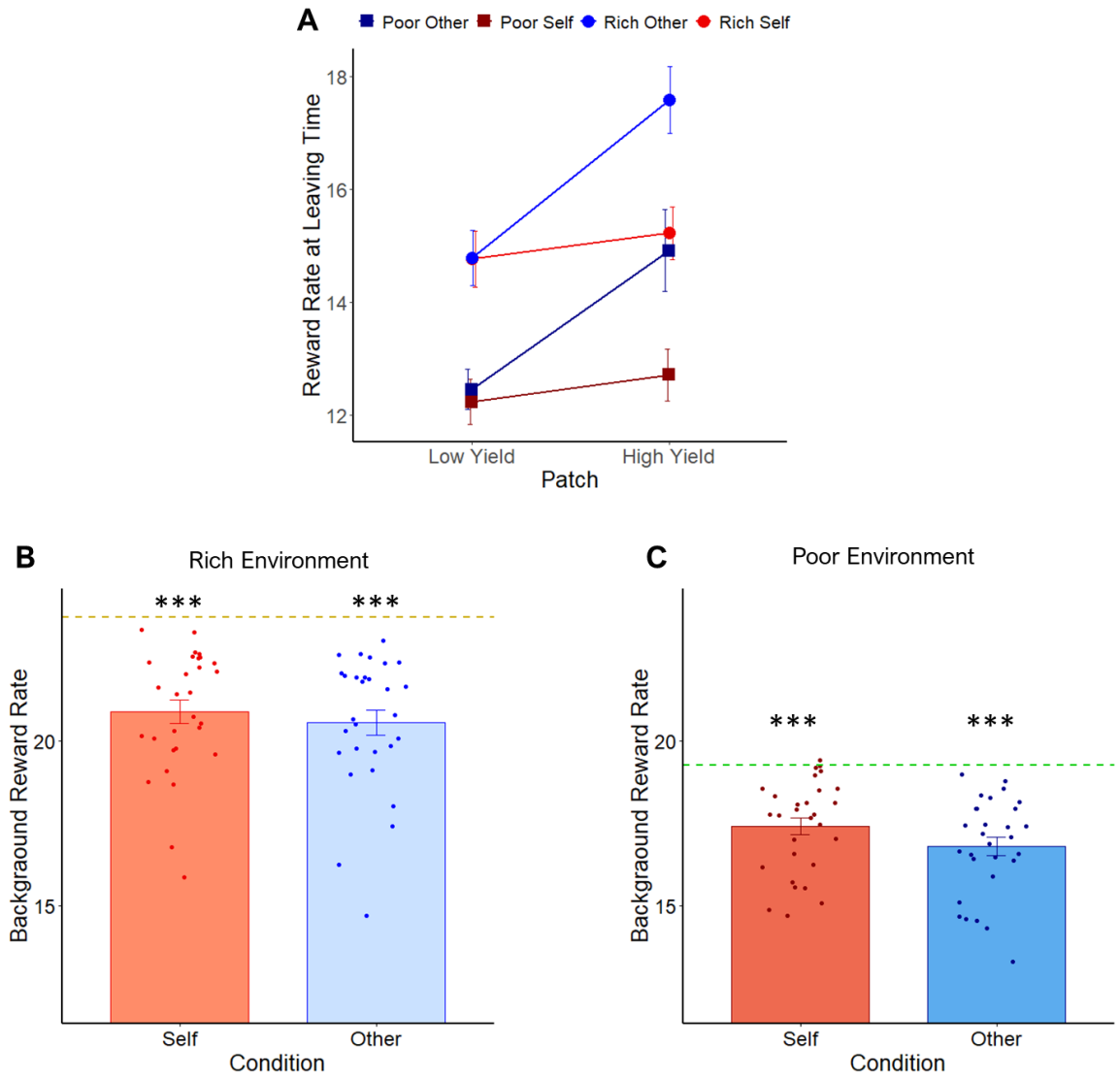


Figure 7.6. Study 2 results on obtained reward rate. A. Fully replicating study 1 results, the reward rate at leaving time was influenced by both the environment and type of patch when collecting rewards for others. For self, only the environment had an effect on patch leaving reward rates. Red lines correspond to *self* environments while blue lines are *other* environments, including rich (lighter circles) and poor (darker squares) qualities. B, C. Replicating study 1, obtained long-run background reward rates in the rich (B) and poor (C) environments for self and other were significantly lower than the actual background reward (gold dotted line, green dotted line, respectively). *** $p < 0.001$. Error bars depict SEM.

Similar to study 1, the background reward rate obtained was significantly lower than the maximal available according to MVT for both environments in both beneficiary conditions (Wilcoxon sign-rank tests, $p < 0.001$). Next, I calculated MVT leaving times predictions for each participant according to their specific average background reward rate obtained in each environment per beneficiary. These calculations adjusted optimal leaving times, on average, in 1.2 second for self and 1.4 for other in the rich environment, and 1 second for self and 1.3 for other in the poor environment. However, despite these adjustments, participants still stayed significantly longer than predicted (Wilcoxon sign-rank tests, $p < 0.001$).

Therefore, results in study 2 partially replicate results of study 1, with participants broadly conforming with MVT predictions for self but not for other. Furthermore, participants were more sensitive to their own instantaneous reward compared with others', although they seem to be similarly sensitive to the average reward in the environment for both self and other.

7.5 Discussion

Previous studies have shown that people manifest prosocial tendencies. Yet, people value their own rewards more than others' (Camerer, 2011; Engel, 2011; Fehr & Camerer, 2007). Many of the studies showing these effects use decision-making tasks where participants must decide between discrete options to reward others at some cost. However, these settings do not capture the dynamics of the decision-making process from which prosocial behaviour likely evolved (Gabay & Apps, 2020; Mobbs et al., 2018). Here, I designed a task based on foraging theories from behavioural ecology, aiming to study prosocial decisions in a dynamic context. Participants made a series of decisions where they continuously had to contrast the instantaneous (foreground) reward rate, with the average (background) reward rate of the environment to maximise their reward

intake. Critically, participants collected rewards for both themselves and for another unknown person. Using this approach, I found that participants showed more sensitivity to the foreground reward rate when collecting rewards for themselves than for others. Intriguingly, however, participants similarly adjusted their behaviour to the average reward of the environment when collecting rewards for both self and other, being similarly sensitive to the background reward rate across beneficiaries. Thus, participants dynamically changed their decisions according to different sources of information, following the ecological principles proposed by MVT (Charnov, 1976). However, participants were suboptimal in their choices, especially when collecting rewards for another person. These results suggest that people are not biased towards their own relative to others' rewards in all aspects of reward-based decisions in a dynamic, ecological setting.

Participants were more sensitive to instantaneous, foreground reward rate when collecting rewards for self than for others. This is consistent with a considerable amount of evidence from cognitive neuroscience showing that people value more their own rewards compared with others' rewards (Camerer, 2011; Declerck et al., 2013; Engel, 2011; Fehr & Camerer, 2007; Strombach et al., 2015; Winking & Mizer, 2013). The current results also suggest that this effect is mainly driven by differences at high reward rates. Participants stayed longer in high yield patches for self rewards relative to others', an effect that was not clearly seen in low yield patches. This opens the possibility that the self-bias manifested in previous studies could occur especially in those decisions that involved high net intake. (Côté, House, & Willer, 2015; Hassall, Silver, Turk, & Krigolson, 2016; Seaman et al., 2018; Strombach et al., 2015; Uhlmann & Barnes, 2014). This fits with results reported in previous chapters, and in the original Lockwood, Hamonet et al. (2017)'s study, where people are more incentivised by higher rewards for *self* than *other* when deciding to put in effort. Thus, depending on the intake, instantaneous rewards are treated differently in a prosocial context.

Crucially, though, participants did not show significant differences in how the background reward rate of the environment affected their foraging decisions when collecting rewards for self and other. Neuroscientific studies have shown that the encoding of current value (foreground) and the comparison with the environment (background) may be underpinned by distinct brain systems (Kolling et al., 2012; Wittmann et al., 2016). Similarly, Le Heron et al. (2020) found that the delivery of a dopamine agonist modulated background, but not foreground reward sensitivity. Thus, different neural and cognitive mechanisms might underlie foreground and background reward sensitivity. Importantly, previous studies have shown that some portions of areas such as the ACC specifically track rewards for others compared with self (Apps et al., 2016). It might be possible then that, even though background reward influences similarly behaviour for self and other in foraging decisions, they are implemented in specialised regions of the brain (Lockwood, Apps, et al., 2020). Future neuroimaging studies could shed light on how foreground and background rewards are implemented in the brain, and how socially-specific or domain general these processes are.

Participants were sensitive to foreground and background reward rates when collecting rewards for both self and other in the direction predicted by MVT, leaving poor vs rich environments, and high vs low yield patches later. However, participants were suboptimal in all the conditions. Participants stayed much longer than what predicted by the optimal solution, consistent with previous results (Constantino & Daw, 2015; Hayden et al., 2011; Campbell Le Heron et al., 2020; Nonacs, 2001; Stephens, 2008). This suboptimal foraging behaviour is starting to be understood by examining the neural (Constantino & Daw, 2015; Hayden et al., 2011; Kolling et al., 2012; Wittmann et al., 2016) and the computational (Mobbs et al., 2018; Wittmann et al., 2016) mechanisms underlying foraging decisions. It is important to note that MVT is a model aiming to understand energy intake for the survival of the organism. Artificial tasks cannot emulate

completely these settings and, as such, some suboptimality in participants behaviour is expected.

Once the overstay bias is controlled for, though, participants in the current study broadly conformed with optimal solutions, but only when collecting rewards for self, and not for other. This suboptimality in the social context is given by participants' decreased sensitivity to foreground reward. This contrasts with evidence showing no differences between collecting rewards for self and other in MVT-like foraging behaviour (Zacharopoulos, Shenhav, Constantino, Maio, & Linden, 2018), and other social-related processes such us social information seeking, which has been linked to MVT-like predictions (Turrin, Fagan, Monte, & Chang, 2017; Yoon, Geary, Ahmed, & Shadmehr, 2018). Nevertheless, these experiments are essentially different from the current task. The former used discrete decisions for when to leave a patch, and focused on total reward intake as the main behavioural index. The latter used social information in a foraging setting, but there were no rewards allocated for others. Here, my main focus was the independent effects of foreground and background reward rates for self and other on leaving times of continuously depleting patches. Thus, the suboptimal behaviour when collecting rewards for others is likely due to different mechanisms. Despite this prosocial suboptimality, though, participants adjusted their behaviour according to the opportunity cost of the environment following MVT principles. This suggests that the low sensitivity on foreground reward rate observed when collecting others' rewards is strong enough to alter optimality even under the same ecological principles that underlie self behaviour.

This chapter comprised two studies where the background reward rate was manipulated differently. Similar results were found across experiments. However, the effects were more consistent in study 2 than 1, even though it had a smaller sample size. Two reasons, among others, could explain these differences. First, despite having a smaller sample, study 2 had more observations per condition, as one entire block was added, and the leaving times of participants were shorter than in study 1, sampling more

cases per condition. Second, even though the manipulation of the environment produced similar results, with participants adjusting their behaviour according to the opportunity of the environment, the processing of this information could have been different for each experiment. While in study 2 travel costs were different but constant, and high and low yield patches were evenly distributed per environment, in study 1 the background reward rate was manipulated by proportions of high and low yield, with 70% and 30% of each type of patch depending on which environment the participant was in. Thus, study 1 could have been perceived as more uncertain compared with study 2. Even though participants knew in advance the probability to encounter a certain patch in each environment in study 1, previous studies have indicated that people can have difficulties in estimating probabilities (Kahneman & Tversky, 1972), and are averse to uncertainty, trying to avoid it or diminish it when they find it (FeldmanHall & Shenhav, 2019). This might have had an impact on the computations underlying performance in the task. Importantly, delay and probability discounting of rewards can be disentangled neurally and behaviourally, suggesting that different mechanisms might underlie background reward estimation in studies 1 and 2 (Białaszek, Ostaszewski, Green, & Myerson, 2019; Holt, Green, & Myerson, 2003; Olson, Hooper, Collins, & Luciana, 2007; Peters & Büchel, 2009; Seaman et al., 2018; Weber & Huettel, 2008). Therefore, these results suggest that properties of the environment beyond its background reward rate could have influenced foraging behaviour. Given two environments with the same background reward rate, it is possible that people vary their foraging decisional process depending on how this average reward is calculated.

The results presented in this chapter suggest that this foraging framework could be powerful to research prosocial behaviour in different settings. For instance, future experiments could introduce energy consumption, such as physical effort, as a parameter that varies foreground and background reward rate, making more salient the self-cost in reward-based prosocial decisions. Moreover, moral decision-making could also be seen

from a different perspective if foraging settings were used to examine, for example, harm aversion for self and other. Furthermore, having independent effects of foreground and background rewards rates on decisions could be useful to distinguish the effects of traits such as empathy and apathy on prosocial reward processing. Sample sizes in studies 1 and 2, taken independently, were not large enough to look at individual differences in this paradigm, and due to the differences in designs between the two studies, they could not be simply collapsed together to test for correlations. Future work could examine the traits and mechanisms underlying variability between people in background and foreground reward sensitivities for self and other.

When to leave a rewarding activity for another potential rewarding location is a key decision problem in multiple situations. Here, using foraging theories, I designed a task where participants made this kind of decisions to reward either themselves or prosocially another person, reporting differences on how participants behaved depending on the beneficiary of the reward. Importantly, the current results revealed that, even though self bias exists in sensitivity to foreground reward rates, the opportunity cost of the environment shapes similarly self and prosocial decisions, suggesting that foraging problems can be powerful when applied to prosocial behaviour, unveiling novel aspects of its decisional process.

Chapter 8

General Discussion

8.1 Introduction

Every day, we perform simple acts to help others, from assisting a colleague in their work, to giving a gift to a loved one. In these actions, people demonstrate they are willing to incur some personal costs, big or small, to provide a direct benefit to others. However, these prosocial actions do not occur with the same frequency across individuals and are more likely to happen under certain circumstances than others. Why are some people more prosocially motivated than others? What are the psychological and neural processes involved in these differences? Across five empirical chapters, I aimed to provide some answers to these questions, unveiling novel aspects underlying individual differences in prosocial behaviour. In this chapter, I will offer a general overview of how these results complement the existing literature on prosocial behaviour, aiming to explain why people vary in their motivations for helping. I also consider the limitations of the current findings, and propose future lines of research to complement them.

8.2 Mechanisms underlying prosocial motivation

In this thesis, I proposed a framework for prosocial behaviour based on goal-directed behaviour, where idiosyncratic sensitivities to personal costs and social benefits underlie individual differences in motivation across different prosocial domains. This framework aimed to unify distinctive theories of prosocial behaviour around one of the core definition of prosociality, i.e. self costs for others' benefits, through a methodological approach that disentangles and manipulates parametrically costs and benefits, separates self-

benefitting from prosocial motivation, and uses effort as a main cost. In the following, I will summarise the novel findings revealed using this model.

Chapters 3 to 5 focused on the relationship between aversion to harm others and the willingness to exert effort to benefit others. Previous studies have shown that, in a harm aversion context, people usually are more averse to harm others than themselves for self-benefit, i.e. hyperaltruism (Crockett et al., 2014, 2015, 2017). In contrast, other studies have shown that people are less willing to exert effort to get rewards for others compared to self, i.e prosocial apathy (Lockwood, Hamonet, et al., 2017; Lockwood et al., 2021). However, it was unknown whether these two prosocial behaviours shared variance i.e. a person who is more averse to harming others is also more willing to exert effort to benefit others. Furthermore, it was unclear whether there was common set of traits and neural mechanisms that might underlie this potential relationship.

In Chapter 3, harm aversion and prosocial effort decisions were positively correlated, such that people who were more prosocial in one context were also more prosocial in the other. People who showed this profile were more incentivised by the benefits obtained by others regardless of their valence. That is, how sensitive someone was to the harm another would receive, was linked to how incentivised they were by the rewards they could obtain for another. The costs for self in the two tasks – the effort to be incurred or the money to be given up – did not link people's behaviour together in these two contexts. This representation of the others' benefits was also reflected in neural responses. Individual differences in ACCg, vAI and TPJp responses to others' harm in harm aversion decisions were associated with how incentivised people were to work for others' rewards in prosocial effort. In contrast, sensitivity to profit when harming others relative to self also influenced prosocial effort, but this effect was associated with the anterior portion of the TPJ.

In Chapter 4, I examined what traits might be predictive of this common disposition in prosocial behaviour. To do this, I took a transdiagnostic approach, showing that people

who display higher prosocial motivation across harm aversion and prosocial effort have greater affective reactivity. Across different affective, psychiatric and subclinical traits, people with a profile characterised by traits such as high empathic concern and low emotional apathy and alexithymia, were more prosocial in reducing others' harm and collecting others' rewards. These results thus suggest that affective traits are, over and above psychiatric disorders, the features more strongly associated with prosociality.

These chapters therefore identified that being prosocial may have a common underlying source. However, people sometimes are hypocritical - they do not always meet their own declared moral standards. Could this be linked to a reduced motivation to help others? In Chapter 5, I found that moral hypocrisy in a harm aversion context, i.e. the discrepancy between moral judgements of others and one's harm aversion decisions, was associated with reduced motivation to exert effort for others. Strikingly, while others' benefits were what linked the harm aversion and prosocial effort behaviour together, hypocrisy was related specifically to a higher sensitivity to the effort costs when working for another person's rewards. This suggests that, if someone has high moral standards, but is reluctant to exert effort for others, it is likely that their moral standards only get manifested through judgements and not through actions. Taken together, Chapters 3 to 5 demonstrate strong links between prosocial behaviour in different contexts, with heightened sensitivity to others benefits creating a common prosocial disposition, but heightened sensitivity to prosocial effort costs linked to a tendency to be morally hypocritical (Fig. 8.1A).

In the last two chapters of this thesis, I aimed to examine in more detail a key element in goal-directed behaviour – reward. In a prosocial action, an agent must be sufficiently incentivised by others' rewards to overcome costs such as effort. However, it is unclear which factors impact sensitivity to vicarious rewards, how variability in neural representation of others' rewards affect prosocial behaviour, and how the rewards in the environment can shape prosocial decisions.

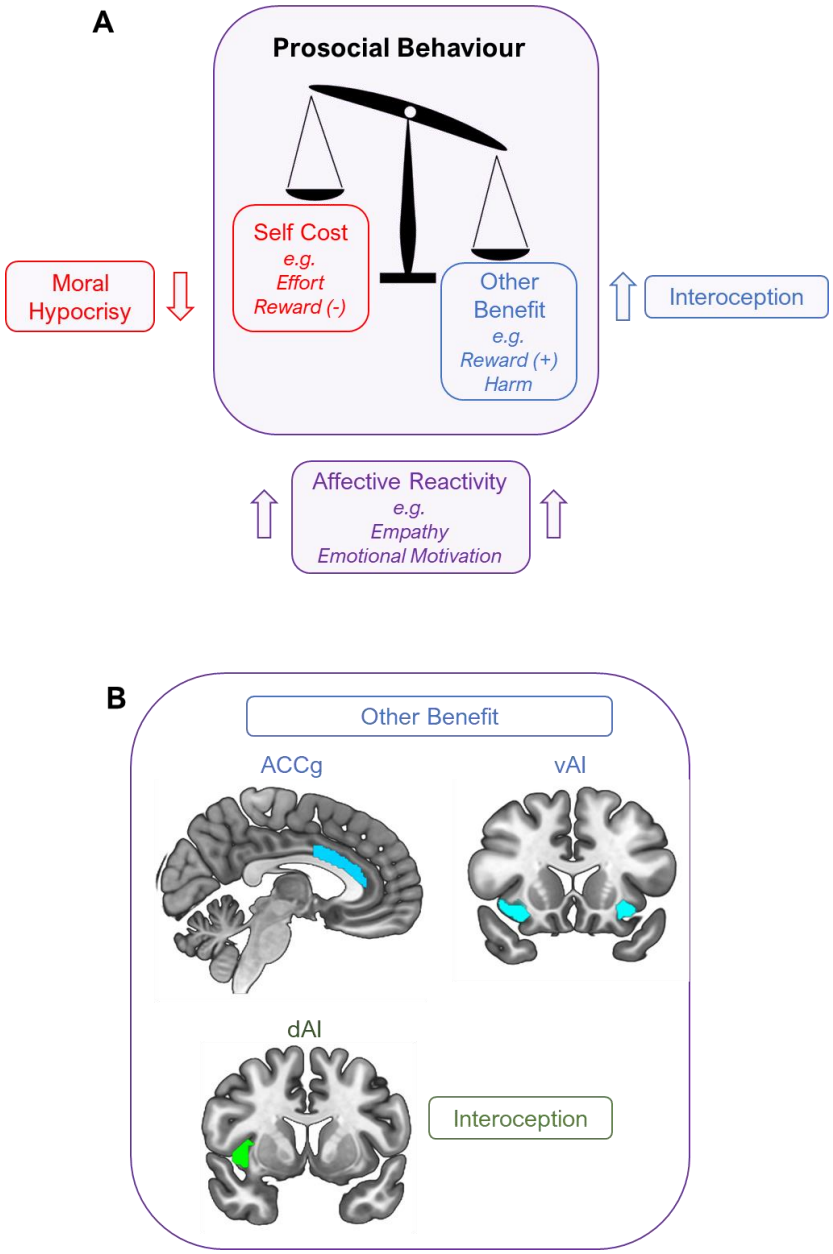


Figure 8.1. Underlying factors of prosocial motivation variability. A. Under a goal-directed framework, prosocial motivation is defined as the willingness to overcome personal costs for the benefit of another person. Two manifestations of prosocial motivation are harm aversion (sacrificing profit to prevent harming others) and prosocial effort (exerting effort to reward another person). Importantly, people who are prosocial in one context are more prosocial in the other, a prosocial disposition characterised by high affective reactivity. Furthermore, people who have high interoceptive ability are more prosocial, as they are more incentivised by other's benefits, while people who are hypocritical are less prosocial, as they are more reluctant to incur costs to help others. B. Anterior cingulate cortex (ACC) and anterior insula (AI) are two key brain structures underlying individual differences in prosocial motivation. Differences in ACC gyri (ACCg) and ventral AI (vAI) responses to others' benefits can predict prosocial behaviour in a different domain. In contrast, the right dorsal AI (dAI) is associated with interoceptive aspects of others' benefits.

In Chapter 6, I found that one of the factors that is involved in others' reward processing is interoception, i.e. the perception of internal bodily signals. Interoception predicted being more incentivised by rewards for others in prosocial effort decisions. Importantly, reward representations in dAI and the ACCg had distinctive functional roles in the association between interoception and prosocial behaviour. On the one hand, similarity between self and other reward signals in the ACCg was linked with the willingness to work for others' rewards. On the other hand, dissimilarity between self and other reward representation in the right dAI, but not in the ACC, predicted how much a person relied on their internal bodily signals. Thus, while ACCg reward representations are tied to motivation, dAI representations are anchored to bodily signals.

Yet, could sensitivity to others' rewards be modulated by the availability of rewards in the environment? In Chapter 7, I proposed a new framework to examine prosocial behaviour, with a focus on dynamic costs in a changing environment. Using foraging theories from behavioural ecology, I designed a task where participants' decisions were based on contrasting the instantaneous foreground reward, with the average background reward in the environment. This approach revealed that the environment shaped both prosocial and self-benefitting decisions, although participants were more sensitive to foreground reward rate when deciding for themselves rather than for others.

Taken together, these chapters shed light on the internal and environmental factors that affect prosocial motivation, unveiling some of the cognitive and neural mechanisms underlying cost-benefit prosocial behaviour.

8.3 Towards an integrated model of individual differences in prosocial motivation

8.3.1 Previous approaches do not focus on individual differences

The current cost-benefit approach is different from previous work in social psychology and behavioural economics in crucial aspects. For instance, findings in economic games have suggested that people value more their own rewards than others' (Camerer, 2011; Engel, 2011). In contrast, theories of morality have proposed that people are moral following action-based decisions, where immoral actions get devalued regardless of their outcomes (Cushman, 2013; Greene, 2009). However, these approaches, among others, fail in separating specific sensitivities to cost and benefits, and likewise, they do not separate self from prosocial motivation. Therefore, an action-based interpretation, or self-bias in reward representations, are problematic deductions without accounting for specific variability in sensitivities to cost and benefits for self and other.

Behaviourally, I found that people are hyperaltruistic because they need more profit to be willing to harm others compared to themselves, consistent with an action-based moral hypothesis, and following previous fMRI results using a similar cost-benefit approach (Crockett et al., 2017). On the other hand, apathetic people in the prosocial effort domain are less sensitive to others' rewards compared with self, consistent with a self-bias valuation of rewards. However, they were also more sensitive to effort costs when benefitting others compared with self, suggesting that cost weighting also triggers apathy, consistent with previous studies that adopted a similar goal-directed approach (Lockwood, Hamonet, et al., 2017). Therefore, this cost-benefit model allows us to support previous interpretations on the mechanisms of prosocial behaviour with well-controlled paradigms where cost and benefits for self and other are parametrically manipulated and disentangled.

Crucially, though, I also found a high variability between people in both prosocial effort and harm aversion tasks that often are not accounted for or have been less well explored, especially given that the patterns of behaviour are somewhat opposite between both domains. Thus, here I focused on the mechanisms underlying individual differences within and between prosocial domains. Using this approach proved highly fruitful - I discovered, for example, that people who are prosocial in the harm aversion task are also prosocial in the prosocial effort context, and they have higher sensitivity to benefits regardless of their valences and across domains. This is consistent with a goal-directed conceptualisation of prosocial behaviour, where people need to be sufficiently incentivised by others' benefits to overcome the cost required to perform a prosocial act (Le Heron et al., 2018; Salamone & Correa, 2018). Thus, and crucially, even though harm aversion and prosocial effort seem to be driven by different mechanisms on the surface, the results provided in this thesis suggest that some individuals display consistency in their prosociality. But, what are the key elements that make a person 'do good' in different contexts?

8.3.2 Individual differences in cost-benefit processing shape prosocial motivation

The shared variance between prosocial effort and harm aversion were thus mainly driven by behavioural and neural sensitivity to others' benefits, suggesting other-regarding mechanisms. This does not mean that moral behaviour is intrinsically driven by social incentives, and does not provide evidence against an action-based model of harm aversion. Rather, it suggests that those people who are sensitive to others' outcomes in a harm aversion context can translate that prosocial interest behaviourally to other prosocial domains.

This suggestion is not trivial. The modality of each benefit in the harm aversion and prosocial effort tasks is different and has opposite valences. Typically, harm is avoided while rewards are approached (Carver & Scheier, 2001; Gable et al., 2003; Gray, 1990). Thus, other-regarding mechanisms seem to facilitate prosocial behaviour across modalities, by putting the focus on the welfare of other people rather than on the self-costs that must be incurred. This perspective is consistent with an empathy-based account on prosocial behaviour (de Waal, 2008; Batson, 2010, 2017). Empathy for pain theories and social rewards approaches to prosocial behaviour outlined in Chapter 1 have emphasised the role of others' affective states in helping actions toward others. However, these perspectives have focused mainly on how affective reactions to observing others' receiving either painful or rewarding outcomes respectively elicit prosocial behaviour, without building a bridge between domains (e.g. Hein et al., 2010, Fukuda et al., 2019). Importantly, sensitivities to others' benefits across contexts might be relatively stable in people having high levels of social and affective traits, suggested by studies showing that these traits are key facilitators of prosocial behaviours, with empathic concern having a fundamental role among them (e.g. FeldmanHall et al., 2015; Hu et al., 2021; Lockwood, et al, 2014). If others' welfare becomes an incentive for a person to act prosocially, adequately representing the affective states of others must be crucial for acting in the benefit of another person regardless of its modality.

Results in Chapter 4 support this claim, revealing that empathic concern, over and above psychiatric traits, predicts higher prosocial motivation across domains. Results in this chapter also suggest that empathetic feelings might be tied to general affective reactions, i.e. the degree to which people engaged emotionally is also a key factor in prosociality. Empathic feelings have been closely related to affective motivation (Lockwood, Ang, et al., 2017), and it is likely that together they facilitate people's understanding of the emotional experiences that others are feeling, acting accordingly to increase their well-being. This affective sensitivity of others' welfare is also consistent with

hypotheses proposing a link between interoception and prosocial behaviour (Palmer & Tsakiris, 2018). Interoception has been identified as a crucial ability to experience and recognise emotions, including empathic feelings, which could guide goal-directed behaviour (Craig, 2009; Critchley & Garfinkel, 2017; Grynberg & Pollatos, 2015; Paulus, 2007). In Chapter 6, I found that higher interoceptive ability was linked to higher sensitivity to others' benefits, and not self-costs, increasing prosocial behaviour. Thus, people who are highly prosocial might be so because they are more motivated and reactive to affective experiences, informed by their bodily internal signals, allowing empathetic feeling towards others' benefits irrespective of their valences. Thus, other-regarding mechanisms are supported by higher affective reactivity and interoceptive ability, which work as facilitator factors for prosocial behaviour (Fig. 8.1A).

But, how do moral principles and cost sensitivity fit in this proposal? In Chapter 3, people who also showed less sensitivity to ill-gotten relative to self rewards were more willing to work for others' rewards in the prosocial effort task. Some theories propose that when moral norms are internalised they can make a person be driven by others' welfare and not mainly by the value of an action (Batson et al., 2002; Eisenberg, 2000; Eisenberg et al., 2016). Results in Chapter 5 add more evidence to the latter interpretation. Besides decisions, moral principles guide the blame assigned to others' harmful actions. Still, most people fail to live-up to their moral standards and fall into hypocritical behaviour – doing those same actions that they blame others for doing (Alicke et al., 2013; Dover, 2019). Results in this chapter showed that great effort sensitivity while deciding and performing prosocial actions was associated with high levels of hypocrisy, i.e. higher gaps between judgement and actions. This suggests that failure to live up moral standards might relate to cost aversion and prosocial apathy (Batson & Thompson, 2001). However, people who may have more internalised moral rules, could overcome costs across domains to benefit others. Thus, some hypocrites might not be inherently immoral and

Machiavellian, as some philosophical traditions suggest (Dover, 2019; Howe & Monin, 2017; Szabados & Soifer, 2004), but too sensitive to the costs of being moral.

In sum, higher prosocial motivation is associated with other-regarding mechanisms, empathy, affective reactivity and interoception. Likewise, being sensitive to immoral actions also predict prosociality in other contexts. Strikingly, this evidence aligns with multiple different theories of prosocial behaviour, such as interceptive hypotheses, empathy-based prosociality, and social reward processing. As such, this thesis highlights that they are not mutually exclusive accounts, but can be unified by how sensitive a person is to another's benefit. However, transgressing moral standards is associated with cost sensitivity and prosocial apathy. The notion that effort matters in social behaviour and cognition is growing (Cameron, Spring, & Todd, 2017; Contreras-Huerta et al., 2020; Lockwood, Hamonet, et al., 2017), adding complexity to traditional perspectives of hypocrisy as fundamentally immoral. Thus, a model of prosocial behaviour with focus on individual differences in cost and benefit sensitivities can unveil common mechanisms across contexts, which would have been lost at the group level.

8.3.3 Neural structures at the base of common prosocial disposition

ACC and AI have been proposed as crucial in goal-directed behaviour – they signal costs and benefits ascribing value to actions (Naqvi & Bechara, 2009; Rushworth et al., 2008). Importantly, besides being involved in prosocial behaviour, vicarious outcomes processing, and empathy (Cutler & Campbell-Meiklejohn, 2019; Lamm et al., 2011; Lockwood, 2016), individual differences in their responses to others' outcomes can predict prosocial preferences (Greening et al., 2014; Tusche et al., 2016). Because of these reasons, these areas are good candidates for being a hub for motivation to help others in different contexts. Importantly, not many other regions have been associated with similar characteristics, apart from TPJ, which has also been linked to complex social

behaviour and cognition (Behrens et al., 2008; Decety & Lamm, 2007; Le Bouc & Pessiglione, 2013; Saxe, 2010). Thus, the scope of this thesis was on individual differences in ACC and AI activation (and TPJ in Chapter 3), and their relationship with prosocial behaviour.

Chapter 3 results showed that the parametric responses in ACCg, vAI and TPJp to others' harm was linked to prosocial effort behaviour. Strikingly, I was able to predict how incentivised a person was by the rewards another person would receive in the prosocial effort task, from individual differences in neural responses in these regions during the harm aversion task. This is despite the fact that the other person's benefit was in opposite valences – pain in the harm aversion, fMRI moral task, and reward in the behavioural, effort task. Moreover, this link occurred only for signalling others' outcomes in moral decisions but not self. These results indicate that individual differences in how a person processes others' harm from a potential immoral action, is associated with prosocial behaviour in a different domain, complementing results showing that empathic responses to others' harm in these areas can elicit helping actions (e.g. Hein, Silani, Preuschoff, Batson, & Singer, 2010; Hu, Hu, Li, & Zhou, 2021).

These results suggest that individual differences in these regions may underlie a common disposition to act prosocially across contexts. This social outcome signal might be relatively stable in each individual, such that people whose neural activity is strong to others' harm in one context, could be also strong to others' rewards in another context, suggested by studies showing that empathy traits positively modulate neural responses to others' outcomes in these areas (Hein et al., 2010; Singer et al., 2004). Chapter 4 supports this interpretation, as the shared variance between prosocial effort and harm aversion was predicted by empathic and affective traits. Therefore, and consistent with other-regarding mechanisms and empathy-based theories of prosocial behaviour, neural sensitivity to others' benefits in one prosocial domain can predict prosocial behaviour in another domain.

8.3.3.1 Distinctive functional roles in neural prosocial variability

Results from Chapter 3 and 6 also suggest that these areas have distinctive roles in prosocial behaviour depending on their anatomical portion. For instance, while TPJp signalled other-regarding mechanisms in Chapter 3, TPJa played a moral-conflict role, such that people who had lower TPJa responses to rewards earned by immoral relative to non-moral means, made more prosocial effort decisions. This suggests a functional dissociation between portions of TPJ. Previous studies have shown that TPJ is involved in representing others' welfare in prosocial behaviour (Hare et al., 2010; Hutcherson et al., 2015; Park et al., 2017; Strombach et al., 2015), whereas other studies have suggested that this area signals the conflict between self and others' interests in a prosocial action, (FeldmanHall et al., 2014; Morishima et al., 2012; Obeso et al., 2018). Chapter 3 results indicate that both theories could be partially true, only that the focus changes depending on the anatomical area within this region. The classic differentiation between a socio-affective role of TPJp and a cognitive-attentional role of TPJa is consistent with these results (Mars et al., 2012). Importantly, the effect of TPJa predicting prosocial behaviour through moral-conflict also highlights the importance of action-based moral decisions in an integrated model of prosociality.

Intriguingly, different portions of the ACC and AI seem to process distinctive information related to prosocial behaviour, suggesting certain specificity in their functional roles. In Chapter 3, ACCg and vAI were associated with other-regarding mechanisms in prosocial behaviour. ACCg has been associated to signalling socially specific information in learning and decision-making processes (Apps et al., 2016), consistent with the current results. Furthermore, vAI has been linked with more affective processing of stimuli, relative to dAI, being involved in empathic feelings and emotional processing (Corradi-Dell'Acqua et al., 2016; Deen et al., 2011; Uddin et al., 2017). Importantly, stronger connectivity between vAI and ACCg has been associated with enhanced affective experiences (Touroutoglou et al., 2012), suggesting that vAI and ACCg might be a circuit

processing affective and motivational aspects of others' benefits, which may facilitate prosocial behaviour.

What would the role of dAI be then? In contrast with the affective role of vAI, and the proposed ACCg function in motivation, dAI has been associated with body ownership and awareness, integrating interoceptive signal with cognitive information (Craig, 2002, 2009; Critchley, Wiens, Rotshtein, Öhman, & Dolan, 2004). This contrasts with the proposed other-regarding mechanisms underlying prosocial behaviour variability. The traditional view from action-perception theories of empathy (de Waal, 2008; Preston & de Waal, 2002) is that the representation of vicarious outcomes should be isomorphic to first-hand experience. From this perspective, vicarious rewards should be represented similarly for self and other in ACC and AI if they act in an empathy-based manner. However, other work (Lockwood 2016; Apps et al., 2016; Lockwood, Apps, et al., 2020) has argued that in fact social specificity in processing others' rewards might be more important in prosocial behaviour. This is consistent with dAI role in body awareness, and interoception - representation of others' outcomes might be anchored to, and distinguished from, self outcomes (Palmer & Tsakiris, 2018).

Results in Chapter 6 shed light on this issue, showing that how dissimilar the representation of rewards obtained by self and others is in the right dAI predicts how much a person is sensitive to their internal bodily signals in a completely different task. Therefore, dAI reward processing is not isomorphic between self and other in relation to interoceptive processes, and it is likely that it has a function in disentangling the beneficiary of a reward considering information from the body. Therefore, even though behavioural results across different chapters suggest an other-regarding mechanism underlying differences in prosociality, with affective reactivity and interoceptive ability facilitating sensitivity to others' outcomes, the implementation in the brain seems to be specialised, with dAI distinctive representations for self and other outcomes predicting interoceptive sensibility but not prosocial behaviour.

On the other hand, increased similarity between self and other reward representations in ACCg predicted more willingness to work for others' rewards, consistent with a motivational role of ACC in prosocial behaviour. Thus, results in Chapter 3 and 6 on ACCg suggest that this area signals vicarious outcomes in an other-regarding manner, such that the more isomorphic the representation of self and others' outcomes is in ACCg, the more prosocial people are. Taken together, processing outcomes for others probably shares mechanisms with self representations, but it is likely that it has its own functional structure depending on their neural implementation.

8.3.4 Dynamic cost-benefit decisions in prosocial behaviour

In light of the results reported in this thesis, within a cost-benefit framework for prosocial behaviour, sensitivity to benefits, such as rewards, seems to be crucial to incentivise prosocial actions. However, participants in the experiments reported here and in most of previous work only consider other's instantaneous benefits – what other people receive after a prosocial decision, without accounting for changes in the environment. But, how do people make decisions in a dynamic environment, where the value of a choice depends on the options available? Chapter 7 provides evidence that, even though people are less sensitive to others' relative to self instantaneous rewards, they are equally sensitive to the opportunity-cost of the environment for self and other, i.e. people adjust their prosocial behaviour according to the dynamic changes in average background reward available.

These results open the opportunity to research cost-benefit decisions in a dynamic context, more similar to the evolutionary roots of prosocial behaviour – ecological foraging decisions that had consequences on the survival of the community (Kelly, 2013). How would other-regarding mechanisms, action-based decisions, interoceptive processing of outcomes and empathic feelings influence these kinds of decisions? As most of these

interpretations are only based on instantaneous benefits, it is unclear how they would impact individual differences in dynamic cost-benefit processing in prosocial behaviour, where opportunity-cost is crucial. In Chapter 7, I designed one of the first experiments, to my knowledge, of this type applied to prosocial behaviour, and I focused on main effects at the group level. Future research endeavours should focus on individual differences to elucidate how idiosyncratic features influence prosocial motivation.

8.4 Concluding Remarks

8.4.1 Limitations and Future Directions

Throughout the different chapters of this thesis, I revealed novel findings in prosocial behaviour using an approach based on individual differences in cost-benefit evaluations. With this aim, I used mainly correlational designs, testing whether variability on one aspect of cognition, behaviour or neural activation can predict another one, revealing, from hypothesis driven analyses, properties of the phenomena to be explained. Correlational designs of course have limitations and thus the strength of all conclusions must be dampened by the absence of causal evidence. However, this approach proved to be fruitful, as I could shed light on some characteristics of prosocial motivation that could have been lost when embedded within group effects. Therefore, future studies should further research the hypotheses generated in this thesis within a controlled manipulation in the experimental design that suggests causal links. For instance, other-regarding mechanisms underlying prosocial motivation could be addressed in a task where different modalities of costs and benefits (e.g. preventing harm through effort) are manipulated. The prosocial effort-sensitivity account of hypocrisy could be tested through a judgement task that includes effort-based prosocial decisions. The distinctive roles of AI and ACC in representing vicarious rewards could be tested though an fMRI task that integrates interoceptive signals into effortful prosocial decisions or with brain stimulation techniques

to disrupt processing. Therefore, even though powerful, the correlational approach must be complemented with well-controlled experimental settings.

It is important to highlight that the results reported here fit into a specific definition of prosocial behaviour within a cost-benefit framework, which emulates a common conceptualisation in biological theories of altruism (de Waal, 2008; Kurzban et al., 2015), and that is based on goal-directed behaviours (Contreras-Huerta et al., 2020). It was not the aim of this thesis to propose this conceptualisation as the only valid definition of the wide spectrum of actions that have been called *prosocial behaviour*. Rather, framing prosocial behaviour under a cost-benefit approach allowed this thesis to (i) operationalise behaviourally prosocial motivation, (ii) unify different theoretical and methodological traditions on prosocial behaviour, (iii) give a clear methodological program to research prosociality in different domains such as harm aversion and prosocial effort, and (iv) focus on the factors that modulate individual differences in prosocial motivation. Despite these advantages, and acknowledging that, indeed, there are many prosocial manifestations in everyday life that fit into a cost-benefit model, there is a still an open debate on the extant in which this approach captures the *essence* of prosociality. Other traditions have emphasised the intentional goal of the agent when acting prosocially – prosocial behaviours are those actions that have the *purpose* of maximising others' welfare (Batson et al., 2015; Padilla-Walker & Carlo, 2015; Schroeder & Graziano, 2015). Notice that in this definition, the costs of the agent, and even the outcome itself (the *actual* social benefit) have secondary roles, as the essence of prosociality would be the agent's motive. Future theoretical and empirical work could elucidate whether the cost-benefit mechanisms proposed here can be expanded to wider and different approaches, and whether agent's motives can be linked to sensitivities to cost and benefits. This thesis is agnostic to those questions, and its findings are limited to the cost-benefit framework in which prosocial behaviour was conceptualised here.

Following this goal-directed framework, results and interpretations of prosocial behaviour given in this thesis have been mainly derived from two specific domains and paradigms – prosocial effort and harm aversion. The former represents the archetype of the cost-benefit approach proposed here, with harm aversion being a moral extension of this conceptual and methodological proposal. However, prosocial behaviour covers a high variety of behaviours depending on the cost and the benefit involved, including charity donations, cooperation between individuals, alleviating pain of others, time-consuming help, reciprocity, and so on (Camerer, 2011; de Waal, 2008; Hutcherson et al., 2015; Rameson et al., 2012). Here, I have revealed that people who show prosocial preferences in prosocial effort and harm aversion contexts have a profile characterised by other-regarding mechanisms. However, it still an open question whether these mechanisms can be translated to other prosocial domains. Future studies should test these hypotheses in a variety of other tasks, following the same methodological considerations such as manipulating parametrically cost and benefits for self and others.

This extension of the current results should also reach the characteristics of the beneficiaries. In all prosocial experiments across chapters, prosocial decisions benefitted an unknown, anonymous person. However, different studies have shown that people are more willing to help, and empathise, to a person increasingly close to the helper (Bernhard et al., 2006; Hein et al., 2010; Strombach et al., 2015). Likewise, outgroups in general are either ignored by people when deciding to help, or are source of opposite, antisocial feelings and actions. Yet, these group biases also vary greatly between people. Are prosocial people across domains also prosocial across beneficiaries? If this answer is yes, are the same mechanisms involved? These questions are very relevant in the field, and they have not been researched in detail, especially under the framework proposed in this thesis. Future research therefore could expand our knowledge on these matters.

Finally, the neuroimaging approach taken in the present thesis also present some limitations. First, I based my analyses on specific hypothesis-driven brain structures

aiming to connect individual differences in neural activation to prosocial decisions. Therefore, I cannot conclusively say that these, and only these, are the core areas underlying prosocial variability. Since the focus of these experiments were on individual differences and not group effects, what it can be deduced from present results is that ACC and AI have predictive power on relevant behaviour associated to prosocial motivation. Other experiments could further elucidate the role of other areas in prosocial behaviour at group and individual levels.

Secondly, and connected with the previous point, the current focus allowed to predict behaviour in one task from neural activation in another completely different task. Even though this is striking and unveil relevant mechanisms in prosocial behaviour, it is a rather indirect measure of the neural mechanisms underlying that specific behaviour. Following research should measure, for instance, dAI activity directly from interoceptive respiratory tasks, or ACCg and vAI responses in a prosocial task that integrate multiple valences and modalities to confirm their involvement in general prosocial disposition. Finally, there are general limitations associated with fMRI as a technique (Huettel, Song, & McCarthy, 2004). It is important to note that fMRI measures indirectly neuronal activity through the BOLD signal, that its temporal resolution is limited, and that is mainly a correlational methodological approach. Thus, future studies could test some of the hypotheses derived from this thesis using more direct neural measures such as single-neuron recording in monkeys or rodents, or techniques with better temporal resolution such as electroencephalography to investigate the temporal dynamics of these effects, or more causal approaches such as lesions studies. All these further investigations will enrich the findings revealed throughout this thesis. Although I was cautious in avoiding reversed-inferences hypothesis and interpretations, and neural activations always were considered in relation to behaviour, the nature of my approach bring some limitations that future experiments could address, complementing the current results.

8.4.2 General Conclusion

What makes a prosocial individual? Across my thesis, my results show that different types of prosocial behaviour are driven by cost-benefit evaluations. There is a prosocial disposition focused on an interoceptive, instantaneous, affective responses to the benefits another will receive from one's help. Prosocial people focus on others' benefits, with greater sensitivity to them in one type of prosocial act predictive of greater sensitivity in another prosocial domain. This prosocial disposition is driven by individual differences in how others' benefits are processed in the ACCg and AI, both when making decisions about avoiding another's harm, and when simply observing rewards being delivered to others. These representations are partially interoceptive and affective, with higher sensitivity to internal states and higher levels of affective traits also linked to stronger processing of others benefits and greater incentivisation by them when choosing whether to exert effort to help others. However, not all aspects of others' benefits underpin prosocial variability, which appears to relate to instantaneous rewards and not to the background average of the environment one is in. Moreover, people who focus on the personal costs of being prosocial are also more hypocritical, suggesting that the independent sensitivity to costs and benefits is a balancing act that leads some people to prosociality but others to more selfish tendencies. Therefore, disentangling the costs and benefits of prosocial acts is vital for understanding individual differences in helping, its neural basis, and thus, complex social interactions.

References

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *NeuroImage*, *35*(4), 1480–1494. https://doi.org/10.1016/j.neuroimage.2007.02.005

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, *26*(5), 673–701. https://doi.org/10.1080/09515089.2012.677397

Amano, T., Ushiyama, K., Moriguchi, S., Fujita, G., & Higuchi, H. (2006). Decision-making in group foragers with incomplete information: test of individual-based model in geese. *Ecological Monographs*, *76*(4), 601–616.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Ang, Y. S., Lockwood, P., Apps, M. A. J., Muhammed, K., & Husain, M. (2017). Distinct subtypes of apathy revealed by the apathy motivation index. *PLoS ONE*, *12*(1), 1–15. https://doi.org/10.1371/journal.pone.0169938

Ang, Y. S., Lockwood, P. L., Kienast, A., Plant, O., Drew, D., Slavkova, E., … Husain, M. (2018). Differential impact of behavioral, social, and emotional apathy on Parkinson's disease. *Annals of Clinical and Translational Neurology*. https://doi.org/10.1002/acn3.626

Apps, M. A. J., & Ramnani, N. (2014). The Anterior Cingulate Gyrus Signals the Net Value of Others' Rewards. *Journal of Neuroscience*, *34*(18), 6190–6200. https://doi.org/10.1523/JNEUROSCI.2701-13.2014

Apps, M.A.J., Green, R., & Ramnani, N. (2013). Reinforcement learning signals in the anterior cingulate cortex code for others' false beliefs. *NeuroImage*, *64*(1). https://doi.org/10.1016/j.neuroimage.2012.09.010

Apps, Matthew A.J., Grima, L. L., Manohar, S., & Husain, M. (2015). The role of cognitive

effort in subjective reward devaluation and risky decision-making. *Scientific Reports*, *5*, 1–11. https://doi.org/10.1038/srep16880

Apps, Matthew A.J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, *90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Apps, Matthew A.J., & Sallet, J. (2017). Social Learning in the Medial Prefrontal Cortex. *Trends in Cognitive Sciences*, *21*(3), 151–152. https://doi.org/10.1016/j.tics.2017.01.008

Apps, Matthew A J. (2018). Stimulating cingulate: distinct behaviours arise from discrete zones. *Brain*, *141*(10), 2827–2830.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, *54*(3), 2033–2044.

Bagby, R. M., Parker, J. D. A., & 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. https://doi.org/10.1016/0022-3999(94)90005-1

Barden, J., Rucker, D. D., & Petty, R. E. (2005). “Saying one thing and doing another”: Examining the impact of event order on hypocrisy judgments of others. *Personality and Social Psychology Bulletin*, *31*(11), 1463–1474. https://doi.org/10.1177/0146167205276430

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing : Keep it maximal. *Journal of Memory and Language*,

68(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, L. F., Quigley, K. S., Bliss-Moreau, E., & Aronson, K. R. (2004). Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*, *87*(5), 684.

Bartel, C. (2019). Hypocrisy as either deception or akrasia. *Philosophical Forum*, *50*(2), 269–281. https://doi.org/10.1111/phil.12220

Bartz, J. A., & Hollander, E. (2006). The neuroscience of affiliation: Forging links between basic and clinical research on neuropeptides and social behavior. *Hormones and Behavior*, *50*(4), 518–528. https://doi.org/10.1016/j.yhbeh.2006.06.018

Basile, B. M., Schafroth, J. L., Karaskiewicz, C. L., Chang, S. W. C., & Murray, E. A. (2020). The anterior cingulate cortex is necessary for forming prosocial preferences from vicarious reinforcement in monkeys. *PLoS Biology*, *18*(6). https://doi.org/10.1371/journal.pbio.3000677

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv Preprint ArXiv:1506.04967*. Retrieved from http://arxiv.org/abs/1506.04967

Batson, C. D. (2010). Empathy-induced altruistic motivation. In M. Mikulincer & P. R. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 15–34). American Psychological Association.

Batson, C. Daniel. (2008a). Empathy-Induced Altruistic Motivation. *Herzliya Symposium on Prosocial Motives, Emotions, and Behavior*, 1–32. https://doi.org/http://dx.doi.org/10.1037/12061-001

Batson, C. Daniel. (2008b). Moral masquerades: Experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences*, *7*(1), 51–66. https://doi.org/10.1007/s11097-007-9058-y

Batson, C. Daniel, Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D.

(1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, *72*(6), 1335–1348. https://doi.org/10.1037/0022-3514.72.6.1335

Batson, C. Daniel, Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., … Sampat, B. (2003). ". . . As you would have them do unto you": Does imagining yourself in the other's place stimulate moral action? *Personality and Social Psychology Bulletin*, *29*(9), 1190–1201. https://doi.org/10.1177/0146167203254600

Batson, C. Daniel, & Thompson, E. R. (2001). Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science*, *10*(2), 54–57. https://doi.org/10.1111/1467-8721.00114

Batson, C. Daniel, Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330–339. https://doi.org/10.1037/0022-3514.83.2.330

Batson, C Daniel. (2017). The empathy altruism hypothesis what and so what? *The Oxford Handbook of Compassion Science*, 27–40.

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

Batson, C Daniel, Eklund, J. H., Chermok, V. L., Hoyt, J. L., & Ortiz, B. G. (2007). An additional antecedent of empathic concern: valuing the welfare of the person in need. *Journal of Personality and Social Psychology*, *93*(1), 65–74. https://doi.org/10.1037/0022-3514.93.1.65

Batson, C Daniel, Lishner, D. A., & Stocks, E. L. (2015). 13 The Empathy–Altruism Hypothesis. *The Oxford Handbook of Prosocial Behavior*, 259.

Batson, C Daniel, Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: appearing moral to oneself without being so. *Journal of*

Personality and Social Psychology, *77*(3), 525.

Beckmann, M., Johansen-Berg, H., & Rushworth, M. F. S. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. *Journal of Neuroscience*, *29*(4), 1175–1190.

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

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, *37*(1), 90–101.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. https://doi.org/10.1038/nature04981

Berridge, K. C. (1996). Food Reward: Brain Substrates of Wanting and Liking KENT. *Neuroscience and Biobehavioral Reviews*, *20*(1), 1–25. https://doi.org/10.1016/0149-7634(95)00033-B

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: “liking”, “wanting”, and learning. *Current Opinion in Pharmacology*, *9*(1), 65–73. https://doi.org/10.1016/j.coph.2008.12.014

Białaszek, W., Ostaszewski, P., Green, L., & Myerson, J. (2019). On Four Types of Devaluation of Outcomes Due to Their Costs: Delay, Probability, Effort, and Social Discounting. *Psychological Record*, *69*(3), 415–424. https://doi.org/10.1007/s40732-

019-00340-x

Bickel, W. K., Jarmolowicz, D. P., Mueller, E. T., Franck, C. T., Carrin, C., & Gatchalian, K. M. (2012). Altruism in time: Social temporal discounting differentiates smokers from problem drinkers. *Psychopharmacology*, *224*(1), 109–120. https://doi.org/10.1007/s00213-012-2745-6

Bird, G., Silani, G., Brindley, R., White, S., Frith, U., & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*, *133*(5), 1515–1525. https://doi.org/10.1093/brain/awq060

Bird, G., & Viding, E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience and Biobehavioral Reviews*, *47*, 520–532. https://doi.org/10.1016/j.neubiorev.2014.09.021

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, *11*(9), 387–392. https://doi.org/10.1016/j.tics.2007.07.003

Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences*, *21*(1), 24–31.

Boccia, M., Dacquino, C., Piccardi, L., Cordellieri, P., Guariglia, C., Ferlazzo, F., … Giannini, A. M. (2017). Neural foundation of human moral reasoning: an ALE meta-analysis about the role of personal perspective. *Brain Imaging and Behavior*, *11*(1), 278–292. https://doi.org/10.1007/s11682-016-9505-x

Bolker, B. M. (2015). Linear and generalized linear mixed models. In *Ecological Statistics: Contemporary theory and application* (pp. 309–333).

Bolton, D. (2008). *What is mental disorder?: an essay in philosophy, science, and values*. Oxford University Press.

Bonnelle, V., Veromann, K. R., Burnett Heyes, S., Lo Sterzo, E., Manohar, S., & Husain, M. (2015). Characterization of reward and effort mechanisms in apathy. *Journal of Physiology Paris*, *109*(1–3), 16–26. https://doi.org/10.1016/j.jphysparis.2014.04.002

Bora, E., Yucel, M., & Allen, N. B. (2009). Neurobiology of human affiliative behaviour: Implications for psychiatric disorders. *Current Opinion in Psychiatry*, *22*(3), 320–325. https://doi.org/10.1097/YCO.0b013e328329e970

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). https://doi.org/10.1038/s41598-017-15385-7

Braams, B. R., Güroğlu, B., de water, E., Meuwese, R., Koolschijn, P. C., Peper, J. S., & Crone, E. A. (2014). Reward-related neural responses are dependent on the beneficiary. *Social Cognitive and Affective Neuroscience*, *9*(7), 1030–1037. https://doi.org/10.1093/scan/nst077

Braun, U. (2018). A Network Perspective on the Search for Common Transdiagnostic Brain Mechanisms. *Biological Psychiatry*, *84*(6), e47–e48. https://doi.org/10.1016/j.biopsych.2018.07.017

Brethel-Haurwitz, K. M., Cardinale, E. M., Vekaria, K. M., Robertson, E. L., Walitt, B., VanMeter, J. W., & Marsh, A. A. (2018). Extraordinary Altruists Exhibit Enhanced Self–Other Overlap in Neural Responses to Distress. *Psychological Science*, *29*(10), 1631–1641. https://doi.org/10.1177/0956797618779590

Brewer, R., Cook, R., & Bird, G. (2016). Alexithymia: a general deficit of interoception. *Royal Society Open Science*, *3*(10), 150664.

Brinkmann, K., Franzen, J., Rossier, C., & Gendolla, G. H. E. (2014). I don't care about others' approval: Dysphoric individuals show reduced effort mobilization for obtaining a social reward. *Motivation and Emotion*, *38*(6), 790–801. https://doi.org/10.1007/s11031-

014-9437-y

Brown, S. L., & Brown, R. M. (2015). Connecting prosocial behavior to improved physical health: Contributions from the neurobiology of parenting. *Neuroscience & Biobehavioral Reviews*, *55*, 1–17.

Bruneau, E. G., Kteily, N. S., & Urbiola, A. (2020). A collective blame hypocrisy intervention enduringly reduces hostility towards Muslims. *Nature Human Behaviour*, *4*(1), 45–54. https://doi.org/10.1038/s41562-019-0747-7

Burgos-Robles, A., Gothard, K. M., Monfils, M. H., Morozov, A., & Vicentic, A. (2019). Conserved features of anterior cingulate networks support observational learning across species. *Neuroscience and Biobehavioral Reviews*, *107*, 215–228. https://doi.org/10.1016/j.neubiorev.2019.09.009

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

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

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

Cameron, C. D., Spring, V. L., & Todd, A. R. (2017). The empathy impulse: A multinomial model of intentional and unintentional empathy for pain. *Emotion*. https://doi.org/10.1037/emo0000266

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology*, *39*(1/4), 106–124.

Capraro, R. M., & Capraro, M. M. (2001). Commonality analysis: Understanding variance contributions to overall canonical correlation effects of attitude toward mathematics on geometry achievement. *Multiple Linear Regression Viewpoints*, *27*(2), 16–23. Retrieved from http://mlrv.ua.edu/2001/VOL_27_2/vol27_n2_a3.pdf

Carnes, N., & Janoff-Bulman, R. (2012). Harm, Help, and the Nature of (Im)Moral (In)Action. *Psychological Inquiry*, *23*(2), 137–142. https://doi.org/10.1080/1047840X.2012.667768

Carpenter, T. P., & Marshall, M. A. (2009). An examination of religious priming and intrinsic Religious motivation in the moral hypocrisy paradigm. *Journal for the Scientific Study of Religion*, *48*(2), 386–393. https://doi.org/10.1111/j.1468-5906.2009.01454.x

Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., & Keysers, C. (2019). Emotional mirror neurons in the rat's anterior cingulate cortex. *Current Biology*, *29*(8), 1301–1312.

Carver, C. S., & Scheier, M. F. (2001). *On the self-regulation of behavior*. Cambridge University Press.

Cascio, C. J., Moore, D., & McGlone, F. (2019). Social touch and human development. *Developmental Cognitive Neuroscience*, *35*, 5–11. https://doi.org/10.1016/j.dcn.2018.04.009

Cassini, M. H., Lichtenstein, G., Ongay, J. P., & Kacelnik, A. (1993). Foraging behaviour in guinea pigs: further tests of the marginal value theorem. *Behavioural Processes*, *29*(1–2), 99–112.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron*, *70*(3), 560–572. https://doi.org/10.1016/j.neuron.2011.02.056

Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: Functional parcellation and large-scale reverse inference. *Cerebral Cortex*, *23*(3), 739–749. https://doi.org/10.1093/cercor/bhs065

Chang, S. W. C., Gariepy, J.-F., & Platt, M. L. (2013). Neuronal reference frames for social decisions in primate frontal cortex. *Nature Neuroscience*, *16*(2), 243–250. https://doi.org/10.1038/nn.3287

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218–1221.

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), 1–28. https://doi.org/10.1371/journal.pbio.1002598

Chow, T. W. (2000). Personality in frontal lobe disorders. *Current Psychiatry Reports*, *2*(5), 446–451. https://doi.org/10.1007/s11920-000-0031-5

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Chudek, M., & Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–226.

Chung, J., & Monroe, G. S. (2003). Exploring social desirability bias. *Journal of Business Ethics*, *44*(4), 291–302.

Cibula, D. A., & Zimmerman, M. (1984). The effect of plant density on departure decisions: testing the marginal value theorem using bumblebees and Delphinium nelsonii. *Oikos*, 154–158.

Coates, D. J., & Tognazzini, N. A. (2013). *Blame: Its nature and norms*. Oxford University Press.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective and Behavioral Neuroscience*, *15*(4), 837–853. https://doi.org/10.3758/s13415-015-0350-y

Contreras-Huerta, L. S., Baker, K. S., Reynolds, K. J., Batalha, L., & Cunnington, R. (2013). Racial bias in neural empathic responses to pain. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0084001

Contreras-Huerta, L. S., Pisauro, M. A., & Apps, M. A. J. (2020). Effort shapes social cognition and behaviour: A neuro-cognitive framework. *Neuroscience and Biobehavioral Reviews*, *118*, 426–439. https://doi.org/10.1016/j.neubiorev.2020.08.003

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, *58*(3), 306–324. https://doi.org/10.1016/j.neuron.2008.04.017

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2013). Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Social Cognitive and Affective Neuroscience*, *8*(4), 424–431. https://doi.org/10.1093/scan/nss014

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., & Singer, T. (2016). Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms10904

Côté, S., House, J., & Willer, R. (2015). High economic inequality leads higher-income individuals to be less generous. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(52), 15838–15843. https://doi.org/10.1073/pnas.1511536112

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, *10*(4-5), 171–178.

Craig, A. D. B. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, *10*(1), 59–70.

https://doi.org/10.1038/nrn2555

Craig, A. D. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4), 500–505. https://doi.org/10.1016/S0959-4388(03)00090-4

Crego, C., & Widiger, T. A. (2015). Psychopathy and the DSM. *Journal of Personality*, *83*(6), 665–677. https://doi.org/10.1111/jopy.12115

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. https://doi.org/10.1901/jaba.2001.34-101

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective, and cognitive integration. *Journal of Comparative Neurology*, *493*(1), 154–166. https://doi.org/10.1002/cne.20749

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. https://doi.org/10.1016/j.copsyc.2017.04.020

Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189–195. https://doi.org/10.1038/nn1176

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Crockett, M. J. (2016). How Formal Models Can Illuminate Mechanisms of Moral Judgment and Decision Making. *Current Directions in Psychological Science*, *25*(2), 85–90. https://doi.org/10.1177/0963721415624012

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), 173201–17325. https://doi.org/10.1073/pnas.1424572112

Crockett, M. J., & Lockwood, P. L. (2018). Extraordinary Altruism and Transcending the Self. *Trends in Cognitive Sciences*, *22*(12), 1071–1073. https://doi.org/10.1016/j.tics.2018.09.003

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. https://doi.org/10.1038/nn.4557

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

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0057410

Cushman, F. (2013). Action, Outcome, and Value: A Dual-System Framework for Morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

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

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

Cutler, J., & Campbell-Meiklejohn, D. (2019). A comparative fMRI meta-analysis of altruistic and strategic decisions to give. *NeuroImage*, *184*, 227–241. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.09.009

Dal Monte, O., Chu, C. C. J., Fagan, N. A., & Chang, S. W. C. (2020). Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. *Nature Neuroscience*, *23*(4), 565–574. https://doi.org/10.1038/s41593-020-0593-y

Dale, A., & Anderson, D. (1978). Information variables in voluntary control and classical conditioning of heart rate: Field dependence and heart-rate perception. *Perceptual and Motor Skills*, *47*(1), 79–85.

Daniel Batson, C., Ahmad, N., & Tsang, J. A. (2002). Four motives for community involvement. *Journal of Social Issues*, *58*(3), 429–445. https://doi.org/10.1111/1540-4560.00269

Danziger, N., Faillenot, I., & Peyron, R. (2009). Can we share a pain we never felt? Neural correlates of empathy in patients with congenital insensitivity to pain. *Neuron*, *61*(2), 203–212. https://doi.org/10.1016/j.neuron.2008.11.023

David, B., Hu, Y., Krüger, F., & Weber, B. (2017). Other-regarding attention focus modulates third-party altruistic choice: An fMRI study. *Scientific Reports*, *7*, 1–12. https://doi.org/10.1038/srep43024

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalogue of Selected Documents in Psychology*, *10*(85).

De Berardis, D., Fornaro, M., Orsolini, L., Valchera, A., Carano, A., Vellante, F., … Di Giannantonio, M. (2017). Alexithymia and suicide risk in psychiatric disorders: A mini-review. *Frontiers in Psychiatry*, *8*(AUG), 148. https://doi.org/10.3389/fpsyt.2017.00148

de Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences*, *10*(10), 435–441. https://doi.org/10.1016/j.tics.2006.08.008

de Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, *59*, 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

de Waal, F. B. M., & Preston, S. D. (2017). Mammalian empathy: Behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, *18*(8), 498–509. https://doi.org/10.1038/nrn.2017.72

Decety, J. (2011). Dissecting the neural mechanisms mediating empathy. *Emotion Review*, *3*(1), 92–108.

Decety, J., Bartal, I. B.-A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150077. https://doi.org/10.1098/rstb.2015.0077

Decety, J., & Cowell, J. M. (2014a). Friends or Foes: Is Empathy Necessary for Moral Behavior? *Perspectives on Psychological Science*, *9*(5), 525–537. https://doi.org/10.1177/1745691614545130

Decety, J., & Cowell, J. M. (2014b). The complex relation between morality and empathy. *Trends in Cognitive Sciences*, *18*(7), 337–339. https://doi.org/10.1016/j.tics.2014.04.008

Decety, J., & Cowell, J. M. (2017). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology*, *30*, 1–12. https://doi.org/10.1017/S0954579417000530

Decety, J., & Grèzes, J. (2006). The power of simulation: imagining one's own and other's behavior. *Brain Research*, *1079*(1), 4–14. https://doi.org/10.1016/j.brainres.2005.12.115

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The*

neuroscientist, 13(6), 580-593.

Declerck, C. H., Boone, C., & Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition*, *81*(1), 95–117. https://doi.org/10.1016/j.bandc.2012.09.009

Deen, B., Pitskel, N. B., & Pelphrey, K. A. (2011). Three systems of insular functional connectivity identified with cluster analysis. *Cerebral Cortex*, *21*(7), 1498–1506. https://doi.org/10.1093/cercor/bhq186

Demers, L. A., & Koven, N. S. (2015). The relation of alexithymic traits to affective theory of mind. *American Journal of Psychology*, *128*(1), 31–42. https://doi.org/10.5406/amerjpsyc.128.1.0031

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of Anterior Cingulate Cortex to Behaviour. *Brain*, *118*, 279–306. Retrieved from %3CGo

Dichter, G. S., Richey, J. A., Rittenberg, A. M., Sabatino, A., & Bodfish, J. W. (2012). Reward circuitry function in autism during face anticipation and outcomes. *Journal of Autism and Developmental Disorders*, *42*(2), 147–160. https://doi.org/10.1007/s10803-011-1221-1

Dillon, D. G., Holmes, A. J., Jahn, A. L., Bogdan, R., Wald, L. L., & Pizzagalli, D. A. (2008). Dissociation of neural regions associated with anticipatory versus consummatory phases of incentive processing. *Psychophysiology*, *45*(1), 36–49. https://doi.org/10.1111/j.1469-8986.2007.00594.x

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2012.01.073

Dondi, M., Simion, F., & Caltran, G. (1999). Can newborns discriminate between their

own cry and the cry of another newborn infant? *Developmental Psychology*, *35*(2), 418.

Dong, M., van Prooijen, J. W., & van Lange, P. A. M. (2019). Self-enhancement in moral hypocrisy: Moral superiority and moral identity are about better appearances. *PLoS ONE*, *14*(7), 1–17. https://doi.org/10.1371/journal.pone.0219382

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., … Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(26), 11073–11078. https://doi.org/10.1073/pnas.0704320104

Dover, D. (2019). The walk and the talk. *Philosophical Review*, *128*(4), 387–422. https://doi.org/10.1215/00318108-7697850

Droutman, V., Bechara, A., & Read, S. J. (2015). Roles of the Different Sub-Regions of the Insular Cortex in Various Phases of the Decision-Making Process. *Frontiers in Behavioral Neuroscience*, *9*, 1–14. https://doi.org/10.3389/fnbeh.2015.00309

Drucker, D. M., & Aguirre, G. K. (2009). Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cerebral Cortex*, *19*(10), 2269–2280. https://doi.org/10.1093/cercor/bhn244

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., … Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21*(12), 1835–1844.

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

Effron, D. A., Markus, H. R., Jackman, L. M., Muramoto, Y., & Muluk, H. (2018). Hypocrisy and culture: Failing to practice what you preach receives harsher interpersonal reactions in independent (vs. interdependent) cultures. *Journal of Experimental Social Psychology*, *76*, 371–384.

Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does "saying one thing but doing another" invite condemnation? *Research in Organizational Behavior*, *38*, 61–75.

Eisenberg, N. (2000). Emotion , Regulation , and Moral. *Annual Review of Psychology*, *51*, 665–697. https://doi.org/10.1146/annurev.psych.51.1.665

Eisenberg, N., VanSchyndel, S. K., & Spinrad, T. L. (2016). Prosocial Motivation: Inferences From an Opaque Body of Work. *Child Development*, *87*(6), 1668–1678. https://doi.org/10.1111/cdev.12638

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, *1*(1), 171–193.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

Ernst, J., Northoff, G., Böker, H., Seifritz, E., & Grimm, S. (2013). Interoceptive awareness enhances neural activity during empathy. *Human Brain Mapping*, *34*(7), 1615–1624. https://doi.org/10.1002/hbm.22014

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *Neuroimage*, *62*(2), 911–922.

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience and*

Biobehavioral Reviews, *35*(3), 903–911. https://doi.org/10.1016/j.neubiorev.2010.10.009

Fareri, D. S., & Delgado, M. R. (2014). Social rewards and social networks in the human brain. *Neuroscientist*, *20*(4), 387–402. https://doi.org/10.1177/1073858414521869

Fauth-Bühler, M., Zois, E., Vollstädt-Klein, S., Lemenager, T., Beutel, M., & Mann, K. (2014). Insula and striatum activity in effort-related monetary reward processing in gambling disorder: the role of depressive symptomatology. *NeuroImage: Clinical*, *6*, 243–251.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427. https://doi.org/10.1016/j.tics.2007.09.002

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

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

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage*, *105*, 347–356. https://doi.org/10.1016/j.neuroimage.2014.10.043

FeldmanHall, O., Dalgleish, T., & Mobbs, D. (2013). Alexithymia decreases altruism in real social decisions. *Cortex*, *49*(3), 899–904. https://doi.org/10.1016/j.cortex.2012.10.015

FeldmanHall, O., Mobbs, D., & Dalgleish, T. (2014). Deconstructing the brain's moral network: Dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *9*(3), 297–306. https://doi.org/10.1093/scan/nss139

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

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Ferri, F., Ardizzi, M., Ambrosecchia, M., & Gallese, V. (2013). Closing the gap between the inside and the outside: interoceptive sensitivity and social distances. *PLoS One*, *8*(10), e75758.

Field, A. P., & Wright, D. B. (2011). A Primer on Using Multilevel Models in Clinical and Experimental Psychopathology Research. *Journal of Experimental Psychopathology*, *2*(2), 271–293. https://doi.org/10.5127/jep.013711

Fritz, K. G., & Miller, D. (2018). Hypocrisy and the Standing to Blame. *Pacific Philosophical Quarterly*, *99*(1), 118–139. https://doi.org/10.1111/papq.12104

Fukuda, H., Ma, N., Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., … Nakahara, H. (2019). Computing social value conversion in the human brain. *Journal of Neuroscience*, *39*(26), 5153–5172. https://doi.org/10.1523/JNEUROSCI.3117-18.2019

Gabay, A. S., & Apps, M. A. J. (2020). Foraging optimally in social neuroscience: computations and methodological considerations. *Social Cognitive and Affective Neuroscience*, 1–13. https://doi.org/10.1093/scan/nsaa037

Gable, S. L., Reis, H. T., & Elliot, A. J. (2003). Evidence for bivariate systems: An empirical test of appetition and aversion across domains. *Journal of Research in Personality*, *37*(5), 349–372.

Gao, Q., Ping, X., & Chen, W. (2019). Body influences on social cognition through interoception. *Frontiers in Psychology*, *10*, 1–5. https://doi.org/10.3389/fpsyg.2019.02066

Gert, B. (2004). *Common Morality: Deciding What to Do*. Oxford University Press.

Gibson, J. (2019). Mindfulness, Interoception, and the Body: A Contemporary Perspective. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02012

Gilead, A. (1999). How is Akrasia Possible After All? *Ratio*, *12*, 257–270.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … Polimeni, J. R. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, *80*, 105–124.

Gleichgerrcht, E., Tomashitis, B., & Sinay, V. (2015). The relationship between alexithymia, empathy and moral judgment in patients with multiple sclerosis. *European Journal of Neurology*, *22*(9), 1295–1303. https://doi.org/10.1111/ene.12745

Gleichgerrcht, E., & Young, L. (2013). Low Levels of Empathic Concern Predict Utilitarian Moral Judgment. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0060418

Glenn, A., Raine, A., & Schug, R. A. (2009). The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry*, *14*(1), 6–9. https://doi.org/10.1038/mp.2008.83

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Gospic, K., Sundberg, M., Maeder, J., Fransson, P., Petrovic, P., Isacsson, G., … Ingvar, M. (2013). Altruism costs-the cheap signal from amygdala. *Social Cognitive and Affective Neuroscience*, *9*(9), 1325–1332. https://doi.org/10.1093/scan/nst118

Goulson, D. (2000). Why do pollinators visit proportionally fewer flowers in large patches? *Oikos*, *91*(3), 485–492.

Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When Values and Behavior Conflict: Moral Pluralism and Intrapersonal Moral Hypocrisy. *Social and Personality Psychology Compass*, *9*(3), 158–170. https://doi.org/10.1111/spc3.12158

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition & Emotion*, *4*(3), 269–288.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Greene, J.D. (2009). The cognitive neuroscience of moral judgment. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed., pp. 987–999). MIT Press.

Greene, Joshua D. (2015). The rise of moral cognition. *Cognition*, *135*, 39–42. https://doi.org/10.1016/j.cognition.2014.11.018

Greene, Joshua D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greening, S., Norton, L., Virani, K., Ty, A., Mitchell, D., & Finger, E. (2014). Individual differences in the anterior insula are associated with the likelihood of financially helping versus harming others. *Cognitive, Affective and Behavioral Neuroscience*, *14*(1), 266–277. https://doi.org/10.3758/s13415-013-0213-3

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63–72.

Grynberg, D., Luminet, O., Corneille, O., Grèzes, J., & Berthoz, S. (2010). Alexithymia in the interpersonal domain: A general deficit of empathy? *Personality and Individual Differences*, *49*(8), 845–850. https://doi.org/10.1016/j.paid.2010.07.013

Grynberg, D., & Pollatos, O. (2015). Perceiving one's body shapes empathy. *Physiology and Behavior*, *140*, 54–60. https://doi.org/10.1016/j.physbeh.2014.12.026

Guttman, H., & Laporte, L. (2002). Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive Psychiatry*, *43*(6), 448–455. https://doi.org/10.1053/comp.2002.35905

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. https://doi.org/10.1093/scan/nsx045

Hadland, K. A., Rushworth, M. F. S., Gaffan, D., & Passingham, R. E. (2003). The effect of cingulate lesions on social behaviour and emotion. *Neuropsychologia*, *41*(8), 919–931.

Haidt, J. (2007). The new synthesis in moral psychology. *Science (New York, N.Y.)*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Upper saddle River, NJ: Prentice Hall.

Hall-McMaster, S., & Luyckx, F. (2019). Revisiting foraging approaches in neuroscience. *Cognitive, Affective and Behavioral Neuroscience*, *19*(2), 225–230. https://doi.org/10.3758/s13415-018-00682-z

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, *97*(896), 354–356.

Han, S., Fan, Y., Xu, X., Qin, J., Wu, B., Wang, X., … Mao, L. (2009). Empathic neural responses to others' pain are modulated by emotional contexts. *Human Brain Mapping*,

30(10), 3227–3237. https://doi.org/10.1002/hbm.20742

Han, Y., Bruls, R., Soyman, E., Thomas, R. M., Pentaraki, V., Jelinek, N., … Pruis, I. (2019). Bidirectional cingulate-dependent danger information transfer across rats. *PLoS Biology*, *17*(12).

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*(5831), 1622–1625.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583–590. https://doi.org/10.1523/JNEUROSCI.4089-09.2010

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., … Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *2018*(5), 1–32. https://doi.org/10.7717/peerj.4794

Hartmann, M. N., Hager, O. M., Tobler, P. N., & Kaiser, S. (2013). Parabolic discounting of monetary rewards by physical effort. *Behavioural Processes*, *100*, 192–196. https://doi.org/10.1016/j.beproc.2013.09.014

Hassall, C. D., Silver, A., Turk, D. J., & Krigolson, O. E. (2016). We are more selfish than we think: The endowment effect and reward processing within the human medial-frontal cortex. *Quarterly Journal of Experimental Psychology*, *69*(9), 1676–1686. https://doi.org/10.1080/17470218.2015.1091849

Hassell, M. P., & Southwood, T. R. E. (1978). Foraging strategies of insects. *Annual Review of Ecology and Systematics*, *9*(1), 75–98.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–

21.

Hauser, M. D., Tonnaer, F., & Cima, M. (2009). When moral intuitions are immune to the law: A case study of euthanasia and the act-omission distinction in the Netherlands. *Journal of Cognition and Culture*, *9*(3), 149–169.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939. https://doi.org/10.1038/nn.2856

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149–160. https://doi.org/10.1016/j.neuron.2010.09.003

Hendryx, M. S., Haviland, M. G., & Shaw, D. G. (1991). Dimensions of Alexithymia and their relationships to Anxiety and Depression. *Journal of Personality Assessment*, *56*(2), 227–237. https://doi.org/10.1207/s15327752jpa5602

Herbert, B. M., Ulbrich, P., & Schandry, R. (2007). Interoceptive sensitivity and physical effort: Implications for the self-control of physical load in everyday life. *Psychophysiology*, *44*(2), 194–202. https://doi.org/10.1111/j.1469-8986.2007.00493.x

Hernandez-Lallement, J., Attah, A. T., Soyman, E., Pinhal, C. M., Gazzola, V., & Keysers, C. (2020). Harm to Others Acts as a Negative Reinforcer in Rats. *Current Biology*, *30*(6), 949-961.e7. https://doi.org/10.1016/j.cub.2020.01.017

Herpertz, S. C., & Bertsch, K. (2014). The social-cognitive basis of personality disorders. *Current Opinion in Psychiatry*, *27*(1), 73–77. https://doi.org/10.1097/YCO.0000000000000026

Hill, K. (2002). Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Human Nature*, *13*(1), 105–128. https://doi.org/10.1007/s12110-002-1016-3

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms12722

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American Economic Review*, *86*(3), 653–660.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 122–128. https://doi.org/10.1016/j.tics.2011.12.008

Holt, D. D., Green, L., & Myerson, J. (2003). Is discounting impulsive?: Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behavioural Processes*, *64*(3), 355–367.

House, B. R. (2018). How do social norms influence prosocial development? *Current Opinion in Psychology*, *20*, 87–91. https://doi.org/10.1016/j.copsyc.2017.08.011

Howard, R. (2015). Personality disorders and violence: what is the link? *Criminology & Criminal Justice: An International Journal*, *14*(1), 2–12. https://doi.org/10.1186/s40479-015-0033-x

Howe, L. C., & Monin, B. (2017). Healthier Than Thou? Practicing What You Preach Backfires by Increasing Anticipated Devaluation Superior Behavior Triggers Concern for Devaluation, *112*(5), 718–735. Retrieved from http://dx.doi.org/10.1037/pspi0000085.supp

Hu, J., Hu, Y., Li, Y., & Zhou, X. (2021). Computational and Neurobiological Substrates of Cost-Benefit Integration in Altruistic Helping Decision. *The Journal of Neuroscience*, *41*(15), 3545–3561. https://doi.org/10.1523/jneurosci.1939-20.2021

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1). Sinauer Associates Sunderland, MA.

Hull, C. L. (1943). *Principles of behaviour*. New York, NY: Appleton-Century-Crofts.

Husain, M., & Roiser, J. P. (2018). Neuroscience of apathy and anhedonia: A transdiagnostic approach. *Nature Reviews Neuroscience*, *19*(8), 470–484. https://doi.org/10.1038/s41583-018-0029-9

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, *87*(2), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031

Ibáñez, A., Hurtado, E., Lobos, A., Escobar, J., Trujillo, N., Baez, S., … Decety, J. (2011). Subliminal presentation of other faces (but not own face) primes behavioral and evoked cortical processing of empathy for pain. *Brain Research*, *1398*, 72–85. https://doi.org/10.1016/j.brainres.2011.05.014

Imas, A. (2014). Working for the warm glow: On the benefits and limits of prosocial incentives. *Journal of Public Economics*. https://doi.org/10.1016/j.jpubeco.2013.11.006

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D., Quinn, K., … Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry Online*, 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2018.01.007

Iodice, P., Porciello, G., Bufalari, I., Barca, L., & Pezzulo, G. (2019). An interoceptive illusion of effort induced by false heart-rate feedback. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 13897–13902. https://doi.org/10.1073/pnas.1821032116

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive*

Neuroscience, *22*(4), 621–631. https://doi.org/10.1162/jocn.2009.21228

Jackson, P. L., Rainville, P., & Decety, J. (2006). To what extent do we share the pain of others? Insight from the neural bases of pain empathy. *Pain*, *125*(1–2), 5–9. https://doi.org/10.1016/j.pain.2006.09.013

James R, & Blair, R. (2013). The neurobiology of psychopathic traits in youths. *Nature Reviews. Neuroscience*, *14*(11), 786–799. https://doi.org/10.1038/nrn3577

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537. https://doi.org/10.1037/a0013779

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Jeon, D., Kim, S., Chetana, M., Jo, D., Ruley, H. E., Lin, S.-Y., … Shin, H.-S. (2010). Observational fear learning involves affective pain system and Ca v 1.2 Ca 2+ channels in ACC. *Nature Neuroscience*, *13*(4), 482.

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., … Casey, B. J. (2011). Behavioral and neural properties of social reinforcement learning. *Journal of Neuroscience*, *31*(37), 13039–13045. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating Stimuli as a Random Factor in Social Psychology : A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under riske. *Econometrica*, *47*, 263–292.

Karns, C. M., Moore, W. E., & Mayr, U. (2017). The cultivation of pure altruism via gratitude: A functional MRI study of change with gratitude practice. *Frontiers in Human Neuroscience*, *11*, 1–14. https://doi.org/10.3389/fnhum.2017.00599

Kelly, R. L. (2013). *The lifeways of hunter-gatherers: the foraging spectrum*. Cambridge University Press.

Kennerley, S. W., Dahmubed, A. F., Lara, A. H., & Wallis, J. D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of Cognitive Neuroscience*, *21*(6), 1162–1178. https://doi.org/10.1162/jocn.2009.21100

Khalsa, S. S., & Lapidus, R. C. (2016). Can interoception improve the pragmatic search for biomarkers in psychiatry? *Frontiers in Psychiatry*, *7*(JUL), 1–19. https://doi.org/10.3389/fpsyt.2016.00121

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science*, *321*(5890), 806–810. https://doi.org/10.1126/science.1156902

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Kirk, U., Gu, X., Sharp, C., Hula, A., Fonagy, P., & Montague, P. R. (2016). Mindfulness training increases cooperative decision making in economic exchanges: Evidence from fMRI. *NeuroImage*, *138*, 274–283. https://doi.org/10.1016/j.neuroimage.2016.05.075

Kittay, E. F. (1982). On hypocrisy. *Metaphilosophy*, *13*(3-4), 277–289.

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.

Klein-Flügge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D., & Bestmann, S. (2015). Behavioral Modeling of Human Choices Reveals Dissociable Effects of Physical Effort and Temporal Delay on Reward Devaluation. *PLoS Computational Biology*, *11*(3), 1–31. https://doi.org/10.1371/journal.pcbi.1004116

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*, *42*(2), 998–1031.

Koenigs, M., Kruepke, M., & Newman, J. P. (2010). Economic decision-making in psychopathy: A comparison with ventromedial prefrontal lesion patients. *Neuropsychologia*, *48*(7), 2198–2204. https://doi.org/10.1016/j.neuropsychologia.2010.04.012

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714. https://doi.org/10.1093/scan/nsr048

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *335*(6077), 95–98. https://doi.org/10.1126/science.1216930

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-018-0401-9

Koyama, T., Kato, K., & Mikami, A. (2000). During pain-avoidance neurons activated in the macaque anterior cingulate and caudate. *Neuroscience Letters*, *283*(1), 17–20.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal

foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953-IN3.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*, 661–679.

Kurzban, Robert, Burton-Chellew, M. N., & West, S. A. (2015). The Evolution of Altruism in Humans. *Annual Review of Psychology*, *66*(1), 575–599. https://doi.org/10.1146/annurev-psych-010814-015355

Kuylen, A. A. A., & Verhallen, T. M. M. (1981). The use of canonical analysis. *Journal of Economic Psychology*, *1*(3), 217–237. https://doi.org/10.1016/0167-4870(81)90039-8

Lambert, Z. V., & Durand, R. M. (1975). Some Precautions in Using Canonical Analysis, *12*(4), 468–475.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *Journal of cognitive neuroscience*, 19(1), 42-58.

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lamm, C., Meltzoff, A. N., & Decety, J. (2010). How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of cognitive neuroscience*, 22(2), 362-376.

Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure and Function*, 214(5-6), 579-591.

Lammers, J., Stapel, D. A., & Galinsky, A. D. (2010). Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. *Psychological Science*, *21*(5), 737–744. https://doi.org/10.1177/0956797610368810

Lander, G. C., Lutz-Zois, C. J., Rye, M. S., & Goodnight, J. A. (2012). The differential association between alexithymia and primary versus secondary psychopathy. *Personality and Individual Differences*, *52*(1), 45–50. https://doi.org/10.1016/j.paid.2011.08.027

Langford, D. J., Crager, S. E., Shehzad, Z., Smith, S. B., Sotocinal, S. G., Levenstadt, J. S., … Mogil, J. S. (2006). Social modulation of pain as evidence for empathy in mice. *Science*, *312*(5782), 1967–1970.

Laurent, S. M., & Clark, B. A. M. (2019). What Makes Hypocrisy? Folk Definitions, Attitude/Behavior Combinations, Attitude Strength, and Private/Public Distinctions. *Basic and Applied Social Psychology*, *41*(2), 104–121. https://doi.org/10.1080/01973533.2018.1556160

Le Bouc, R., & Pessiglione, M. (2013). Imaging social motivation: Distinct brain mechanisms drive effort production during collaboration versus competition. *Journal of Neuroscience*, *33*(40), 15894–15902. https://doi.org/10.1523/JNEUROSCI.0143-13.2013

Le Heron, C., Apps., M. A. J., & Husain, M. (2018). The anatomy of apathy: A neurocognitive framework for amotivated behaviour. *Neuropsychologia*, *118*(May 2017), 54–67. https://doi.org/10.1016/j.neuropsychologia.2017.07.003

Le Heron, Campbell, Kolling, N., Plant, O., Kienast, A., Janska, R., Ang, Y. S., … Apps, M. A. J. (2020). Dopamine Modulates Dynamic Decision-Making during Foraging. *Journal of Neuroscience*, *40*(27), 5273–5282. https://doi.org/10.1523/JNEUROSCI.2586-19.2020

Leach, L. F., & Henson, R. K. (2014). Bias and precision of the squared canonical correlation coefficient under nonnormal data conditions. *Journal of Modern Applied Statistical Methods*, *13*(1), 110–139. https://doi.org/10.22237/jmasm/1398917220

Lemanski, N. J., Cook, C. N., Smith, B. H., & Pinter-Wollman, N. (2019). A multiscale review of behavioral variation in collective foraging behavior in honey bees. *Insects*, *10*(11), 370.

Lewis, A. R. (1982). Selection of nuts by gray squirrels and optimal foraging theory. *American Midland Naturalist*, 250–257.

Li, Z., Yu, H., Zhou, Y., Kalenscher, T., & Zhou, X. (2020). Guilty by association: How group-based (collective) guilt arises in the brain. *NeuroImage*, *209*(December 2019), 116488. https://doi.org/10.1016/j.neuroimage.2019.116488

Lindenberg, S., Steg, L., Milovanovic, M., & Schipper, A. (2018). Moral hypocrisy and the hedonic shift: A goal-framing approach. *Rationality and Society*, *30*(4), 393–419. https://doi.org/10.1177/1043463118795719

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., & Barrett, L. F. (2016). The Brain Basis of Positive and Negative Affect: Evidence from a Meta-Analysis of the Human Neuroimaging Literature. *Cerebral Cortex*, *26*(5), 1910–1922. https://doi.org/10.1093/cercor/bhv001

Lischke, A., Weippert, M., Mau-Moeller, A., Jacksteit, R., & Pahnke, R. (2020). Interoceptive accuracy is associated with emotional contagion in a valence- and sex-dependent manner. *Social Neuroscience*, *15*(2), 227–233. https://doi.org/10.1080/17470919.2019.1690573

Lockwood, P. L. (2016). The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behavioural Brain Research*, *311*, 255–266. https://doi.org/10.1016/j.bbr.2016.05.048

Lockwood, P. L., Abdurahman, A., Gabay, A. S., Drew, D., Tamm, M., Husain, M., & Apps, M. A. J. (2021). Aging Increases Prosocial Motivation for Effort. *Psychological Science*, *32*(5), 668–681. https://doi.org/10.1177/0956797620975781

Lockwood, P. L., Ang, Y.-S., Husain, M., & Crockett, M. J. (2017). Individual differences in empathy are associated with apathy-motivation. *Scientific Reports*, *7*(1), 17293. https://doi.org/10.1038/s41598-017-17415-w

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is there a "social" brain? Implementations and algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813.

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of Vicarious Reward Prediction in Anterior Cingulate Cortex and Relationship with Trait Empathy. *The Journal of Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763–9768. https://doi.org/10.1073/pnas.1603198113

Lockwood, P. L., Bird, G., Bridge, M., & Viding, E. (2013). Dissecting empathy : high levels of psychopathic and autistic traits are characterized by difficulties in different social information processing domains, *7*, 1–6. https://doi.org/10.3389/fnhum.2013.00760

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. J. (2017). Prosocial apathy for helping others when effort is required. *Nature Human Behaviour*, *1*(7), 0131. https://doi.org/10.1038/s41562-017-0131

Lockwood, P. L., & Klein-Flügge, M. C. (2020). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and*

Affective Neuroscience, 1–11. https://doi.org/10.1093/scan/nsaa040

Lockwood, P. L., Klein-Flügge, M. C., Abdurahman, A., & Crockett, M. J. (2020). Model-free decision making is prioritized when learning to avoid harming others. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27719–27730. https://doi.org/10.1073/pnas.2010890117

Lockwood, P. L., Seara-Cardoso, A., & Viding, E. (2014). Emotion regulation moderates the association between empathy and prosocial behavior. *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0096555

Lockwood, P. L., Sebastian, C. L., McCrory, E. J., Hyde, Z. H., Gu, X., De Brito, S. A., & Viding, E. (2013). Association of callous traits with reduced neural response to others' pain in children with conduct problems. *Current Biology*, 23(10), 901-905https://doi.org/10.1016/j.cub.2013.04.018

Löffler, A., Foell, J., & Bekrater-Bodmann, R. (2018). Interoception and Its Interaction with Self, Other, and Emotion Processing: Implications for the Understanding of Psychosocial Deficits in Borderline Personality Disorder. *Current Psychiatry Reports*, *20*(4). https://doi.org/10.1007/s11920-018-0890-2

Lönnqvist, J. E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology*, *55*, 53–62. https://doi.org/10.1016/j.jesp.2014.06.004

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Machado, D. G. da S., Farias Junior, L. F. de, Nascimento, P. H. D. do, Tavares, M. P. M., Anselmo da Silva, S. K., Agrícola, P. M. D., … Okano, A. H. (2019). Can

interoceptive accuracy influence maximal performance, physiological and perceptual responses to exercise? *Physiology and Behavior*, *204*, 234–240. https://doi.org/10.1016/j.physbeh.2019.02.038

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*. https://doi.org/10.1016/j.dcn.2014.01.002

Mannarini, S., Balottin, L., Toldo, I., & Gatta, M. (2016). Alexithymia and psychosocial problems among Italian preadolescents. A latent class analysis approach. *Scandinavian Journal of Psychology*, *57*(5), 473–481. https://doi.org/10.1111/sjop.12300

Manohar, S. G., Chong, T. T. J., Apps, M. A. J., 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. https://doi.org/10.1016/j.cub.2015.05.038

Marin, R. S. (1990). Differential diagnosis and classification of apathy. *American Journal of Psychiatry*, *147*(1), 22–30. https://doi.org/10.1176/ajp.147.1.22

Mars, R. B., Sallet, J., Schueffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. S. (2012). Connectivity-Based Subdivisions of the Human Right "Temporoparietal Junction Area": Evidence for Different Areas Participating in Different Cortical Networks. *Cerebral Cortex*, *22*(8), 1894–1903.

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences*, *111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for "social pain" and subsequent prosocial behavior. *NeuroImage*, *55*(1), 381–

388. https://doi.org/10.1016/j.neuroimage.2010.11.060

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McGorry, P., & Nelson, B. (2016). Why we need a transdiagnostic staging approach to emerging psychopathology, early diagnosis, and treatment. *JAMA Psychiatry*, *73*(3), 191–192. https://doi.org/10.1001/jamapsychiatry.2015.2868

McMorris, T., Barwood, M., & Corbett, J. (2018). Central fatigue theory and endurance exercise: Toward an interoceptive model. *Neuroscience and Biobehavioral Reviews*, *93*(February), 93–107. https://doi.org/10.1016/j.neubiorev.2018.03.024

Mele, A. R. (1989). Akratic feelings. *Philosophy and Phenomenological Research*, *50*(2), 277–288.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Merritt, A., & Monin, B. (2012). Moral Hypocrisy, Moral Inconsistency, and the Struggle for Moral Integrity. *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, 167–184.

Metz, M. C., Vucetich, J. A., Smith, D. W., Stahler, D. R., & Peterson, R. O. (2011). Effect of sociality and season on gray wolf (Canis lupus) foraging behavior: implications for estimating summer kill rate. *PloS One*, *6*(3), e17332.

Mier, D., Lis, S., Esslinger, C., Sauer, C., Hagenhoff, M., Ulferts, J., … Kirsch, P. (2013). Neuronal correlates of social cognition in borderline personality disorder. *Social Cognitive and Affective Neuroscience*, *8*(5), 531–537. https://doi.org/10.1093/scan/nss028

Miller, J. D., Lamkin, J., Maples-Keller, J. L., Sleep, C. E., & Lynam, D. R. (2018). A test of the empirical profile and coherence of the DSM-5 psychopathy specifier. *Psychological Assessment*, *30*(7), 870–881. https://doi.org/10.1037/pas0000536

Miller, R., & Cushman, F. (2013). Aversive for me, wrong for you: First-person behavioral aversions underlie the moral condemnation of harm. *Social and Personality Psychology Compass*, *7*(10), 707–718. https://doi.org/10.1111/spc3.12066

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: Insights from ethology. *Nature Reviews Neuroscience*, *19*(7), 419–427. https://doi.org/10.1038/s41583-018-0010-7

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. https://doi.org/10.1126/science.1170539

Moll, J., Krueger, F., Zahn, R., Pardini, M., De Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15623–15628. https://doi.org/10.1073/pnas.0604475103

Monfardini, E., Gazzola, V., Boussaoud, D., Brovelli, A., Keysers, C., & Wicker, B. (2013). Vicarious Neural Processing of Outcomes during Observational Learning. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0073879

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., Mann, A., & T, A. T. (2003). Standardised Assessment of Personality − Abbreviated Scale (SAPAS): preliminary validation of a brief screen for personality disorder Standardised Assessment of Personality - Abbreviated Scale (SAPAS): preliminary validation of a brief screen for pers. *British Journal of Psychiatry*, *183*, 228–232. https://doi.org/10.1192/03-32

Morelli, S. A., Rameson, L. T., & Lieberman, M. D. (2014). The neural components of

empathy: Predicting daily prosocial behavior. *Social Cognitive and Affective Neuroscience*, *9*(1), 39–47. https://doi.org/10.1093/scan/nss088

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, *112*, 244–253. https://doi.org/10.1016/j.neuroimage.2014.12.056

Moriguchi, Y., Decety, J., Ohnishi, T., Maeda, M., Mori, T., Nemoto, K., … Komaki, G. (2007). Empathy and judging other's pain: an fMRI study of alexithymia. *Cerebral Cortex*, *17*(9), 2223–2234. https://doi.org/10.1093/cercor/bhl130

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, *75*(1), 73–79. https://doi.org/10.1016/j.neuron.2012.05.021

Mosner, M. G., Kinard, J. L., McWeeny, S., Shah, J. S., Markiewitz, N. D., Damiano-Goodwin, C. R., … Dichter, G. S. (2017). Vicarious Effort-Based Decision-Making in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *47*(10), 2992–3006. https://doi.org/10.1007/s10803-017-3220-3

Murphy, J., Brewer, R., Hobson, H., Catmur, C., & Bird, G. (2018). Is alexithymia characterised by impaired interoception? Further evidence, the importance of control variables, and the problems with the Heartbeat Counting Task. *Biological Psychology*, *136*(May), 189–197. https://doi.org/10.1016/j.biopsycho.2018.05.010

Murphy, J., Catmur, C., & Bird, G. (2018). Alexithymia is associated with a multidomain, multidimensional failure of interoception: Evidence from novel tests. *Journal of Experimental Psychology: General*, *147*(3), 398–408. https://doi.org/10.1037/xge0000366

Naqvi, N. H., & Bechara, A. (2009). The hidden island of addiction: the insula. *Trends in Neurosciences*, *32*(1), 56–67. https://doi.org/10.1016/j.tins.2008.09.009

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional “trolley problem”. *Emotion*, *12*(2), 364.

Neubert, F. X., Mars, R. B., Sallet, J., & Rushworth, M. F. S. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), E2695–E2704. https://doi.org/10.1073/pnas.1410767112

New, A. S., aan het Rot, M., Ripoll, L. H., Perez-Rodriguez, M. M., Lazarus, S., Zipursky, E., … Siever, L. J. (2012). Empathy and Alexithymia in Borderline Personality Disorder: Clinical and Laboratory Measures. *Journal of Personality Disorders*, *26*(5), 660–675. https://doi.org/10.1521/pedi_2012_26_037

Newton-Howes, G., & Clark, L. A. (2015). Personality disorder across the life course. *The Lancet*, *385*, 727–734.

Nimon, K., Henson, R. K., & Gates, M. S. (2010). Revisiting interpretation of canonical correlation analysis: A tutorial and demonstration of canonical commonality analysis. *Multivariate Behavioral Research*, *45*(4), 702–724. https://doi.org/10.1080/00273171.2010.498293

Niv, Y. (2007). Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Annals of the New York Academy of Sciences*, *1104*, 357–376. https://doi.org/10.1196/annals.1390.018

Nomi, J. S., Schettini, E., Broce, I., Dick, A. S., & Uddin, L. Q. (2018). Structural connections of functionally defined human insular subdivisions. *Cerebral Cortex*, *28*(10), 3445–3456. https://doi.org/10.1093/cercor/bhx211

Nomi, Jason S., Farrant, K., Damaraju, E., Rachakonda, S., Calhoun, V. D., & Uddin, L.

Q. (2016). Dynamic functional network connectivity reveals unique and overlapping profiles of insula subdivisions. *Human Brain Mapping*, *37*(5), 1770–1787. https://doi.org/10.1002/hbm.23135

Nonacs, P. (2001). State dependent behavior and the marginal value theorem. *Behavioral Ecology*, *12*(1), 71–83.

O'Connor, K., Effron, D. A., & Lucas, B. J. (2020). Moral cleansing as hypocrisy: When private acts of charity make you feel better than you deserve. *Journal of Personality and Social Psychology*.

Obeso, I., Moisa, M., Ruff, C. C., & Dreher, J. C. (2018). A causal role for right temporo-parietal junction in signaling moral conflict. *ELife*, *7*, 1–16. https://doi.org/10.7554/eLife.40671

Olson, E. A., Hooper, C. J., Collins, P., & Luciana, M. (2007). Adolescents' performance on delay and probability discounting tasks: contributions of age, intelligence, executive functioning, and self-reported externalizing behavior. *Personality and Individual Differences*, *43*(7), 1886–1897.

Ondobaka, S., Kilner, J., & Friston, K. (2017). The role of interoceptive inference in theory of mind. *Brain and Cognition*, *112*, 64–68. https://doi.org/10.1016/j.bandc.2015.08.002

Padilla-Walker, L. M., & Carlo, G. (Eds.). (2015). *Prosocial development: A multidimensional approach*. Oxford University Press.

Palmer, C. E., & Tsakiris, M. (2018). Going at the heart of social cognition: is there a role for interoception in self-other distinction? *Current Opinion in Psychology*, *24*, 21–26. https://doi.org/10.1016/j.copsyc.2018.04.008

Palomero-Gallagher, N., Mohlberg, H., Zilles, K., & Vogt, B. (2008). Cytology and receptor architecture of human anterior cingulate cortex. *Journal of Comparative Neurology*, *508*(6), 906–926.

Park, S. Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., & Tobler, P. N. (2017). A neural link between generosity and happiness. *Nature Communications*, *8*, 1–10. https://doi.org/10.1038/ncomms15964

Patel, G. H., Sestieri, C., & Corbetta, M. (2019). The evolution of the temporoparietal junction and posterior superior temporal sulcus. *Cortex*, *118*, 38–50. https://doi.org/10.1016/j.cortex.2019.01.026

Patel, G. H., Yang, D., Jamerson, E. C., Snyder, L. H., Corbetta, M., & Ferrera, V. P. (2015). Functional evolution of new and expanded attention networks in humans. *Proceedings of the National Academy of Sciences*, *112*(30), 9454–9459. https://doi.org/10.1073/pnas.1516559112

Patil, I., & Silani, G. (2014a). Alexithymia increases moral acceptability of accidental harms. *Journal of Cognitive Psychology*, *26*(5), 597–614. https://doi.org/10.1080/20445911.2014.929137

Patil, I., & Silani, G. (2014b). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology*, *5*(MAY), 1–12. https://doi.org/10.3389/fpsyg.2014.00501

Paulhus, D. L., Neumann, C. S., & Hare, R. D. (2009). *Manual for the self-report psychopathy scale*. Toronto, Canada: Multi-Health Systems.

Paulus, M. P. (2007). Neural basis of reward and craving-a homeostatic point of view. *Dialogues in Clinical Neuroscience*, *9*(4), 379–387.

Paulus, M. P., & Stewart, J. L. (2014). Interoception and drug addiction. *Neuropharmacology*, *76*, 342–350. https://doi.org/10.1016/j.neuropharm.2013.07.002

Pessiglione, M., Vinckier, F., Bouret, S., Daunizeau, J., & Le Bouc, R. (2018). Why not try harder? Computational approach to motivation deficits in neuro-psychiatric diseases. *Brain*. https://doi.org/10.1093/brain/awx278

Peters, J., & Büchel, C. (2009). Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. *Journal of Neuroscience*, *29*(50), 15727–15734. https://doi.org/10.1523/JNEUROSCI.3489-09.2009

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 13099–13104. https://doi.org/10.1073/pnas.1008137107

Picard, N., & Strick, P. L. (1996). Motor areas of the medial wall: a review of their location and functional activation. *Cerebral Cortex*, *6*(3), 342–353.

Piech, R. M., Strelchuk, D., Knights, J., Hjälmheden, J. V., Olofsson, J. K., & Aspell, J. E. (2017). People with higher interoceptive sensitivity are more altruistic, but improving interoception does not increase altruism. *Scientific Reports*, *7*(1), 1–5. https://doi.org/10.1038/s41598-017-14318-8

Pollatos, O., Traut-Mattausch, E., Schroeder, H., & Schandry, R. (2007). Interoceptive awareness mediates the relationship between anxiety and the intensity of unpleasant feelings. *Journal of Anxiety Disorders*, *21*(7), 931–943.

Polman, E., & Ruttan, R. L. (2012). Effects of anger, guilt, and envy on moral hypocrisy. *Personality and Social Psychology Bulletin*, *38*(1), 129–139. https://doi.org/10.1177/0146167211422365

Porreca, F., & Navratilova, E. (2017). Reward, motivation and emotion of pain and its relief. *Pain*, *158*, S43.

Post, S. G. (2005). Altruism, happiness, and health: It's good to be good. *International Journal of Behavioral Medicine*, *12*(2), 66–77. https://doi.org/10.1207/s15327558ijbm1202_4

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, *84*, 320–341.

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, *25*(1), 1–20; discussion 20-71. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12625087

Prévost, C., Pessiglione, M., Météreau, E., Cléry-Melin, M. L., & Dreher, J. C. (2010). Separate valuation subsystems for delay and effort decision costs. *Journal of Neuroscience*, *30*(42), 14080–14090. https://doi.org/10.1523/JNEUROSCI.2752-10.2010

Prims, J., & Motyl, M. (2018). A tool for detecting low quality data in internet research. Retrieved from https://github.com/SICLab/detecting-bots

Rachlin, H., & Jones, B. A. (2008). Altruism among relatives and non-relatives. *Behavioural Processes*, *79*(2), 120–123.

Rameson, L. T., Morelli, S. A., & Lieberman, M. D. (2012). The neural correlates of empathy: Experience, automaticity, and prosocial behavior. *Journal of Cognitive Neuroscience*, *24*(1), 235–245. https://doi.org/10.1162/jocn_a_00130

Rand, D. G., & Epstein, Z. G. (2014). Risking your life without a second thought: Intuitive decision-making and extreme altruism. *PLoS ONE*, *9*(10), 1–6. https://doi.org/10.1371/journal.pone.0109687

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current Research and Theory*, 64–99.

Rilling, J. K., Glenn, A. L., Jairam, M. R., Pagnoni, G., Goldsmith, D. R., Elfenbein, H. A., & Lilienfeld, S. O. (2007). Neural Correlates of Social Cooperation and Non-Cooperation as a Function of Psychopathy. *Biological Psychiatry*, *61*(11), 1260–1271. https://doi.org/10.1016/j.biopsych.2006.07.021

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, *18*(2), 159–165. https://doi.org/10.1016/j.conb.2008.06.003

Rilling, J. K., & Sanfey, A. G. (2011). The neurobiology of social decision-making. *Annual Review of Psychology*, *62*(1), 23–48. https://doi.org/10.1016/j.conb.2008.06.003

Ring, C., & Brener, J. (2018). Heartbeat counting is unrelated to heartbeat detection: A comparison of methods to quantify interoception. *Psychophysiology*, *55*(9), 1–10. https://doi.org/10.1111/psyp.13084

Rita, H., Ranta, E., & Peuhkuri, N. (1997). Group foraging, patch exploitation time and the finder's advantage. *Behavioral Ecology and Sociobiology*, *40*(1), 35–39.

Robson, S. E., Repetto, L., Gountouna, V.-E., & Nicodemus, K. K. (2019). A review of neuroeconomic gameplay in psychiatric disorders. *Molecular Psychiatry*, 1. https://doi.org/10.1038/s41380-019-0405-5

Rolls, E. T. (2016). Functions of the anterior insula in taste, autonomic, and related functions. *Brain and Cognition*, *110*, 4–19. https://doi.org/10.1016/j.bandc.2015.07.002

Rosati, A. G. (2017). Foraging Cognition: Reviving the Ecological Intelligence Hypothesis. *Trends in Cognitive Sciences*, *21*(9), 691–702. https://doi.org/10.1016/j.tics.2017.05.011

Rudebeck, P. H., Buckley, M. J., Walton, M. E., & Rushworth, M. F. S. (2006). A role for the macaque anterior cingulate gyrus in social valuation. *Science*, *313*(5791), 1310–1312. https://doi.org/10.1126/science.1128197

Rushworth, M. F.S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417. https://doi.org/10.1016/j.tics.2004.07.009

Rushworth, Matthew F.S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. https://doi.org/10.1038/nn2066

Rütgen, M., Seidel, E. M., Silani, G., Riečanský, I., Hummer, A., Windischberger, C., … Lamm, C. (2015). Placebo analgesia and its opioidergic regulation suggest that empathy for pain is grounded in self pain. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), E5638–E5646. https://doi.org/10.1073/pnas.1511269112

Salamone, J. D., & Correa, M. (2018). Neurobiology and pharmacology of activational and effort-related aspects of motivation: rodent studies. *Current Opinion in Behavioral Sciences*, *22*, 114–120. https://doi.org/https://doi.org/10.1016/j.cobeha.2018.01.026

Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., … Rushworth, M. F. S. (2011). Social Network Size Affects Neural Circuits in Macaques. *Science*, *334*(6056).

Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. In *Handbook of Theory of Mind*.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology*, *18*(4), 483–488.

Schielzeth, H., & Forstmeier, W. (2008). Conclusions beyond support : overconfident

estimates in mixed models. *Behavioral ecology,* 20(2), 416-420. https://doi.org/10.1093/beheco/arn145

Schroeder, D. A., & Graziano, W. G. (Eds.). (2015). *The Oxford handbook of prosocial behavior*.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Seaman, K. L., Brooks, N., Karrer, T. M., Castrellon, J. J., Perkins, S. F., Dang, L. C., … Samanez-Larkin, G. R. (2018). Subjective value representations during effort, probability and time discounting across adulthood. *Social Cognitive and Affective Neuroscience*, *13*(5), 449–459. https://doi.org/10.1093/scan/nsy021

Seedorff, M., Oleson, J., & McMurray, B. (2019). Maybe maximal: Good enough mixed models optimize power while controlling Type I error. https://doi.org/10.31234/osf.io/xmhfr

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., … Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, *27*(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Shah, P., Catmur, C., & Bird, G. (2017). From heart to mind: Linking interoception, emotion, and theory of mind. *Cortex*, *93*, 220–223. https://doi.org/10.1016/j.cortex.2017.02.010

Shah, P., Hall, R., Catmur, C., & Bird, G. (2016). Alexithymia, not autism, is associated with impaired interoception. *Cortex*, *81*, 215–220.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *17*(1), 18–24. https://doi.org/10.1177/1073858410379268

Sherry, A., & Henson, R. K. (2005). Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. *Journal of Personality Assessment*, *84*(1), 37–48. https://doi.org/10.1207/s15327752jpa8401

Showers, M. J. C. (1959). The cingulate gyrus: additional motor area and cortical autonomic regulator. *Journal of Comparative Neurology, 112,* 231– 287.

Silani, G., Bird, G., Brindley, R., Singer, T., Frith, C., & Frith, U. (2008). Levels of emotional awareness and autism: An fMRI study. *Social Neuroscience*, *3*(2), 97–112. https://doi.org/10.1080/17470910701577020

Sims, T. B., Van Reekum, C. M., Johnstone, T., & Chakrabarti, B. (2012). How reward modulates mimicry: EMG evidence of greater facial mimicry of more rewarding happy faces. *Psychophysiology*, *49*(7), 998–1004.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*(8), 334–340. https://doi.org/10.1016/j.tics.2009.05.001

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. *New Methods in Cognitive Psychology*, 4–31. https://doi.org/10.4324/9780429318405-2

Smith, A. (2010). *The theory of moral sentiments*. Cambridge, United Kingdom: Cambridge University Press. (Original work published 1790).

Smith, A. R., Steinberg, L., & Chein, J. (2014). The role of the anterior insula in adolescent decision making. *Developmental Neuroscience*, *36*(3–4), 196–209.

Smith, E. A., Bettinger, R. L., Bishop, C. A., Blundell, V., Cashdan, E., Casimir, M. J., … Hayden, B. (1983). Anthropological applications of optimal foraging theory: a critical review [and comments and reply]. *Current Anthropology*, *24*(5), 625–651.

Smith, D. V., Hayden, B. Y., Truong, T. K., Song, A. W., Platt, M. L., & Huettel, S. A. (2010). Distinct value signals in anterior and posterior ventromedial prefrontal cortex. *Journal of Neuroscience*, *30*(7), 2490–2495. https://doi.org/10.1523/JNEUROSCI.3319-09.2010

Starkstein, S. E., & Leentjens, A. F. G. (2008). The nosological position of apathy in clinical practice. *Journal of Neurology, Neurosurgery and Psychiatry*, *79*(10), 1088–1092. https://doi.org/10.1136/jnnp.2007.136895

Stellar, J. E., & Willer, R. (2014). The Corruption of Value: Negative Moral Associations Diminish the Value of Money. *Social Psychological and Personality Science*, *5*(1), 60–66. https://doi.org/10.1177/1948550613484770

Stephens, D. W. (2008). Decision ecology: foraging and the ecology of animal decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 475–484.

Stern, G. R. (2019). Altruism: Giving for Mental Well-Being. *Journal of the American Psychiatric Nurses Association*, *25*(4), 314–315. https://doi.org/10.1177/1078390319857080

Stibbard-Hawkes, D. N. E. (2019). Costly signaling and the handicap principle in hunter-gatherer research: A critical review. *Evolutionary Anthropology*, *28*(3), 144–157. https://doi.org/10.1002/evan.21767

Story, G. W., Kurth-Nelson, Z., Crockett, M., Vlaev, I., Darzi, A., & Dolan, R. J. (2020). Social discounting of pain. *Journal of the Experimental Analysis of Behavior*, *114*(3),

308–325. https://doi.org/10.1002/jeab.631

Strauss, G. P., & Cohen, A. S. (2017). A Transdiagnostic Review of Negative Symptom Phenomenology and Etiology. *Schizophrenia Bulletin*, *43*(4), 712–729. https://doi.org/10.1093/schbul/sbx066

Strickland, C. M., Drislane, L. E., Lucy, M., Krueger, R. F., & Patrick, C. J. (2013). Characterizing Psychopathy Using DSM-5 Personality Traits. *Assessment*, *20*(3), 327–338. https://doi.org/10.1177/1073191113486691

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *Neuroimage*, *54*(1), 671–680.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications.* CRC press.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135–170. https://doi.org/10.1037/0033-295X.88.2.135

Szabados, B., & Soifer, E. (2004). *Hypocrisy: ethical investigations*. Broadview Press.

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. https://doi.org/10.1177/1745691618776263

Tamnes, C. K., Overbye, K., Ferschmann, L., Fjell, A. M., & Walhovd, K. B. (2018). Social

Perspective Taking Is Associated With Self-Reported Prosocial Behavior and Regional Cortical Thickness Across Adolescence. *Developmental Psychology*, *54*(9), 1745–1757. https://doi.org/10.1037/dev0000541.supp

Tang, H., Wang, S., Liang, Z., Sinnott-Armstrong, W., Su, S., & Liu, C. (2018). Are proselfs more deceptive and hypocritical? Social image concerns in appearing fair. *Frontiers in Psychology*, *9*, 1–9. https://doi.org/10.3389/fpsyg.2018.02268

Tang, H., Ye, P., Wang, S., Zhu, R., Su, S., Tong, L., & Liu, C. (2017). Stimulating the right temporoparietal junction with tDCS decreases deception in moral hypocrisy and unfairness. *Frontiers in Psychology*, *8*, 1–7. https://doi.org/10.3389/fpsyg.2017.02033

Taylor, G. J., & Bagby, R. M. (2004). New Trends in Alexithymia Research. *Psychotherapy and Psychosomatics*, *73*(2), 68–77. https://doi.org/10.1159/000075537

Terasawa, Y., Fukushima, H., & Umeda, S. (2013). How does interoceptive awareness interact with the subjective experience of emotion? An fMRI Study. *Human Brain Mapping*, *34*(3), 598–612. https://doi.org/10.1002/hbm.21458

Terasawa, Y., Motomura, K., Natsume, A., Iijima, K., Chalise, L., Sugiura, J., … Umeda, S. (2021). Effects of insular resection on interactions between cardiac interoception and emotion recognition. *Cortex*, *137*, 271–281. https://doi.org/10.1016/j.cortex.2021.01.011

Thobois, S., Prange, S., Sgambato-Faure, V., Tremblay, L., & Broussolle, E. (2017). Imaging the Etiology of Apathy, Anxiety, and Depression in Parkinson's Disease: Implication for Treatment. *Current Neurology and Neuroscience Reports*, *17*(10). https://doi.org/10.1007/s11910-017-0788-0

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Todd, P. (2019). A Unified Account of the Moral Standing to Blame. *Nous*, *53*(2), 347–374. https://doi.org/10.1111/nous.12215

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., & Feldman Barrett, L. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage*, *60*(4), 1947–1958. https://doi.org/10.1016/j.neuroimage.2012.02.012

Tremblay, R. E. (2000). The development of aggressive behaviour during childhood: What have we learned in the past century? *International Journal of Behavioral Development*, *24*(2), 129–141. https://doi.org/10.1080/016502500383232

Tricomi, E., Rangel, A., Camerer, C. F., & Doherty, J. P. O. (2010). Neural evidence for inequality-averse social preferences. *Nature*, *463*(7284), 1089–1091. https://doi.org/10.1038/nature08785

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge University Press.

Turrin, C., Fagan, N. A., Monte, O. D., & Chang, S. W. C. (2017). Social resource foraging is guided by the principles of the Marginal Value Theorem. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-11763-3

Tusche, A., Bockler, A., Kanske, P., Trautwein, F.-M., & Singer, T. (2016). Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. *Journal of Neuroscience*, *36*(17), 4719–4732. https://doi.org/10.1523/JNEUROSCI.3392-15.2016

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320.

Uddin, L., Kinnison, J., Pessoa, L., & Anderson, M. L. (2014). Beyond the tripartite cognition–emotion–interoception model of the human insular cortex. *Journal of Cognitive Neuroscience*, *26*(1), 16-27. https://doi.org/10.1162/jocn

Uddin, L. Q., Nomi, J. S., Hébert-Seropian, B., Ghaziri, J., & Boucher, O. (2017). Structure and function of the human insula. *Journal of Clinical Neurophysiology*, *34*(4), 300–306. https://doi.org/10.1097/WNP.0000000000000377.Structure

Uhlmann, E. L., & Barnes, C. M. (2014). Selfish play increases during high-stakes NBA games and is rewarded with more lucrative contracts. *PLoS ONE*, *9*(4), 1–5. https://doi.org/10.1371/journal.pone.0095745

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

Vahl, P., Colins, O. F., Lodewijks, H. P. B., Markus, M. T., Doreleijers, T. A. H., & Vermeiren, R. R. J. M. (2014). Psychopathic-like traits in detained adolescents: Clinical usefulness of self-report. *European Child and Adolescent Psychiatry*, *23*(8), 691–699. https://doi.org/10.1007/s00787-013-0497-4

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, *44*(5), 1334–1338. https://doi.org/10.1016/j.jesp.2008.03.010

Valdespino, A., Antezana, L., Ghane, M., & Richey, J. A. (2017). Alexithymia as a transdiagnostic precursor to empathy abnormalities: The functional role of the insula. *Frontiers in Psychology*, *8*, 1–7. https://doi.org/10.3389/fpsyg.2017.02234

Van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A. R. B., & Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Social Cognitive and Affective Neuroscience*, *4*(3), 294–304.

Van Der Meulen, M., Van IJzendoorn, M. H., & Crone, E. A. (2016). Neural correlates of prosocial behavior: Compensating social exclusion in a four-player cyberball game.

PLoS ONE, *11*(7), 1–13. https://doi.org/10.1371/journal.pone.0159045

van Gurp, S., Hoog, J., Kalenscher, T., & van Wingerden, M. (2020). Vicarious reward unblocks associative learning about novel cues in male rats. *Elife*, *9*, e60755.

Van Reekum, R. ., Stuss, D. T. ., & Ostrander, L. . (2005). Apathy: Why care? *Journal of Neuropsychiatry and Clinical Neurosciences*, *17*(1), 7–19. https://doi.org/10.1176/appi.neuropsych.17.1.7

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2014). Model comparison and the principle of parsimony. *Oxford Handbook Of*, 1–29. Retrieved from http://www.ejwagenmakers.com/inpress/VandekerckhoveEtAlinpress.pdf

Venkataraman, V. V, Kraft, T. S., Dominy, N. J., & Endicott, K. M. (2017). Hunter-gatherer residential mobility and the marginal value of rainforest patches. *Proceedings of the National Academy of Sciences*, *114*(12), 3097–3102.

Vogt, B. (2009). *Cingulate Neurobiology and Disease*. New York: Oxford University Press.

Vogt, B. A., Nimchinsky, E. A., Vogt, L. J., & Hof, P. R. (1995). Human cingulate cortex: surface features, flat maps, and cytoarchitecture. *Journal of Comparative Neurology*, *359*(3), 490–506.

Vogt, B. A., & Paxinos, G. (2014). Cytoarchitecture of mouse and rat cingulate cortex with human homologies. *Brain Structure and Function*, *219*(1), 185–192. https://doi.org/10.1007/s00429-012-0493-3

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, *114*(30), 7963–7968. https://doi.org/10.1073/pnas.1706693114

Vucetich, J. A., Peterson, R. O., & Waite, T. A. (2004). Raven scavenging favours group

foraging in wolves. *Animal Behaviour*, *67*(6), 1117–1126.

Wallace, R. J. (2010). Hypocrisy, moral address, and the equal standing of persons. *Philosophy & Public Affairs*, *38*(4), 307–341.

Wallis, J. D., & Kennerley, S. W. (2011). Contrasting reward signals in the orbitofrontal cortex and anterior cingulate cortex. *Annals of the New York Academy of Sciences*, *1239*(1), 33–42.

Walton, M. E., Croxson, P. L., Behrens, T. E. J., Kennerley, S. W., & Rushworth, M. F. S. (2007). Adaptive decision making and value in the anterior cingulate cortex. *Neuroimage*, *36*, T142–T154.

Walton, M. E., Rudebeck, P. H., Bannerman, D. M., & Rushworth, M. F. S. (2007). Calculating the cost of acting in frontal cortex. *Annals of the New York Academy of Sciences*, *1104*, 340–356. https://doi.org/10.1196/annals.1390.009

Wang, H. T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S., & Bzdok, D. (2020). Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, *216*(March), 116745. https://doi.org/10.1016/j.neuroimage.2020.116745

Wang, H., Zhang, J., & Jia, H. (2019). Separate Neural Systems Value Prosocial Behaviors and Reward: An ALE Meta-Analysis. *Frontiers in Human Neuroscience*, *13*(August), 1–6. https://doi.org/10.3389/fnhum.2019.00276

Wang, Y., Shima, K., Sawamura, H., & Tanji, J. (2001). Spatial distribution of cingulate cells projecting to the primary, supplementary, and pre-supplementary motor areas: a retrograde multiple labeling study in the macaque monkey. *Neuroscience Research*, *39*(1), 39–49.

Warach, B., Josephs, L., & Gorman, B. S. (2019). Are Cheaters Sexual Hypocrites? Sexual Hypocrisy, the Self-Serving Bias, and Personality Style. *Personality and Social*

Psychology Bulletin, *45*(10), 1499–1511. https://doi.org/10.1177/0146167219833392

Warneken, F. (2016). Insights into the biological foundation of human altruistic sentiments. *Current Opinion in Psychology*, *7*, 51–56.

Watson, G. W., & Sheikh, F. (2008). Normative self-interest or moral hypocrisy?: The importance of context. *Journal of Business Ethics*, *77*(3), 259–269. https://doi.org/10.1007/s10551-006-9348-2

Weber, B. J., & Huettel, S. A. (2008). The neural substrates of probabilistic and intertemporal decision making. *Brain Research*, *1234*, 104–115. https://doi.org/10.1016/j.brainres.2008.07.105

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045.

White, B. A. (2014). Who cares when nobody is watching? Psychopathic traits and empathy in prosocial behaviors. *Personality and Individual Differences*, *56*(1), 116–121. https://doi.org/10.1016/j.paid.2013.08.033

White, S. F., Zhao, H., Leong, K. K., Smetana, J. G., Nucci, L. P., & Blair, R. J. R. (2017). Neural correlates of conventional and harm/welfare-based moral decision-making. *Cognitive, Affective and Behavioral Neuroscience*, *17*(6), 1114–1128. https://doi.org/10.3758/s13415-017-0536-6

Wiebe, A., Kersting, A., & Suslow, T. (2017). Deployment of attention to emotional pictures varies as a function of externally-oriented thinking: An eye tracking investigation. *Journal of Behavior Therapy and Experimental Psychiatry*, *55*, 1–5. https://doi.org/10.1016/j.jbtep.2016.11.001

Wiens, S. (2005). Interoception in emotional experience. *Current Opinion in Neurology*, *18*(4), 442–447.

Williams, Z. M., Bush, G., Rauch, S. L., Cosgrove, G. R., & Eskandar, E. N. (2004). Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nature Neuroscience*, *7*(12), 1370–1375. https://doi.org/10.1038/nn1354

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, *34*(4), 288–293.

Wittmann, M. K., Kolling, N., Akaishi, R., Chau, B. K. H., Brown, J. W., Nelissen, N., & Rushworth, M. F. S. (2016). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nature Communications*, *7*(1), 1–13. https://doi.org/10.1038/ncomms12327

Woodard, S. H., & Jha, S. (2017). Wild bee nutritional ecology: predicting pollinator population dynamics, movement, and services from floral resources. *Current Opinion in Insect Science*, *21*, 83–90.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108. https://doi.org/10.1523/JNEUROSCI.1642-12.2013

Xiao, X., & Zhang, Y.-Q. (2018). A new perspective on the anterior cingulate cortex and affective pain. *Neuroscience & Biobehavioral Reviews*, *90*, 200–211.

Yamada, M., & Decety, J. (2009). Unconscious affective processing and empathy: an investigation of subliminal priming on the detection of painful facial expressions. *Pain*, *143*(1–2), 71–75.

Yang, T. T., Simmons, A. N., Matthews, S. C., Tapert, S. F., Frank, G. K., Bischoff-Grethe, A., … Paulus, M. P. (2009). Adolescent subgenual anterior cingulate activity is related to harm avoidance. *Neuroreport*, *20*(1), 19.

Yoon, T., Geary, R. B., Ahmed, A. A., & Shadmehr, R. (2018). Control of movement vigor and decision making during foraging. *Proceedings of the National Academy of Sciences*

of the United States of America 115(44), E10476–E10485.

Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, *9*(8), 1150–1158. https://doi.org/10.1093/scan/nst090

Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modeling Morality in 3-D: Decision-Making, Judgment, and Inference. *Topics in Cognitive Science*, *11*(2), 409–432. https://doi.org/10.1111/tops.12382

Yu, R., Geddes, J. R., & Fazel, S. (2012). Personality Disorders, Violence, and Antisocial Behavior: A Systematic Review and Meta-Regression Analysis. *Journal of Personality Disorders*, *26*(5), 775–792. https://doi.org/10.1521/pedi.2012.26.5.775

Zacharopoulos, G., Shenhav, A., Constantino, S., Maio, G. R., & Linden, D. E. J. (2018). The effect of self-focus on personal and social foraging behaviour. *Social Cognitive and Affective Neuroscience*, *13*(9), 967–975. https://doi.org/10.1093/scan/nsy057

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/a0037679

Zaki, J., Davis, J. I., & Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *NeuroImage*, *62*(1), 493–499. https://doi.org/10.1016/j.neuroimage.2012.05.012

Zanarini, M. C., Vujanovic, A. A., Parachini, E. A., Boulanger, J. L., Frankenburg, F. R., & Hennen, J. (2003). A Screening Measure for BPD: The McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD). *Journal of Personality Disorders*, *17*(6), 568–573. https://doi.org/10.1521/pedi.17.6.568.25355

Zhan, Y., Xiao, X., Tan, Q., Li, J., Fan, W., Chen, J., & Zhong, Y. (2020). Neural correlations of the influence of self-relevance on moral decision-making involving a trade-off between harm and reward. *Psychophysiology*, *57*(9), 1–15.

https://doi.org/10.1111/psyp.13590

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.

Appendix A

Supplementary tables for whole brain analyses in Chapter 3.

Tables A.1 – A.4. Uncorrected ($p < 0.001$, $k > 30$) whole brain results for differences in money and shocks between the harmful and the helpful options for *self* and *other* trials in the harm aversion task.

Table A.1 Differences in shocks in self trials

Regions	*x*	*y*	*z*	*F*	*k*
Lingual Gyrus	12	-79	-7	79.99	343
Superior Frontal Gyrus	-18	-10	65	39.1	1107
Precuneus	9	-67	41	25.34	144
Central Opercular Cortex	-54	-1	5	24.4	167
Parietal Operculum Cortex	57	-31	26	24.07	102
Precentral Gyrus	63	8	38	23.73	94
Supramarginal Gyrus	-48	-19	26	21.15	101
Inferior Frontal Gyrus	-30	35	17	25.84	61
Right Putamen	27	-1	14	25.67	72
Supramarginal Gyrus	66	-34	50	23.98	37
Lingual Gyrus	0	-70	5	22.84	43
Superior Temporal Gyrus	72	-28	17	20.43	46
Cerebellum VIIIa	-18	-67	-49	19.33	58
Lateral Occipital Cortex	39	-82	8	18.86	39
Precentral Gyrus	48	-4	59	18.14	33

Table A.2 Differences in money in self trials

Regions	*x*	*y*	*z*	*F*	*k*
Angular Gyrus	51	-52	20	33.88	474
Angular Gyrus	-45	-55	20	23.66	312
Superior Temporal Gyrus	57	-1	-16	22.93	150
Superior Temporal Gyrus	-51	-31	2	21.7	93
Right Amygdala	24	-1	-19	21.83	53
Left Amygdala	-18	-7	-19	20.12	42
Posterior Cingulate Gyrus	-9	-43	38	15.62	72

Table A.3 Differences in shocks in other trials

Regions	*x*	*y*	*z*	*F*	*k*
Lingual Gyrus	-12	-82	-13	48.27	450
Left Caudate	-15	-7	26	19.78	38
Superior Parietal Lobule	42	-43	68	19.24	117
Right Caudate	21	-4	26	19.15	35
Precentral Gyrus	54	8	23	18.9	38
Cerebellum VIIb	-30	-73	-58	17.69	35
Postcentral Gyrus	-45	-37	65	15.29	37
Precentral Gyrus	33	-7	65	14.34	35

Table A.4 Differences in money in other trials

Regions	*x*	*y*	*z*	*F*	*k*
Lateral Occipital Cortex	-54	-67	26	33.45	273
Superior Parietal Lobule	24	-55	50	23.26	39
Frontal Pole	-12	59	20	23.25	42
Superior Frontal Gyrus	-24	35	44	18.55	34
Angular Gyrus	42	-49	-29	18.22	62

Appendix B

Supplementary tables for whole brain analyses in Chapter 6.

Table B.1 – B.2. Uncorrected ($p < 0.01$) results of correlations between each of the 200 ROIs from Neurosynth (https://identifiers.org/neurovault.collection:2099) with interoception score and sensitivity to others' rewards in the prosocial effort task. Only (uncorrected) significant results are shown.

Table B.1. Correlations between ROIs and interoception

Regions	*x*	*y*	*z*	*r*	*p <*
Anterior Insula	38	11	5	-0.36	0.008
Posterior Cingulate Cortex	1	-26	28	-0.35	0.009
Precentral Gyrus	-51	10	38	-0.39	0.004
Supramarginal Gyrus	58	-34	46	-0.4	0.003

Table B.2. Correlations between ROIs and other reward betas in the prosocial effort task.

Regions	*x*	*y*	*z*	*r*	*p <*
Thalamus	-26	-24	8	-0.38	0.004