Sustainability of empathy as driver for prosocial behavior and social closeness: insights from computational modelling and functional magnetic resonance imaging

Nachhaltigkeit von Empathie als Motiv für
prosoziales Verhalten und soziale Nähe:
Erkenntnisse auf Grundlage von computational modelling und funktioneller
Magnetresonanztomographie

DISSERTATION

For a doctoral degree (Dr. rer. nat.)
at the Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section Neuroscience

submitted by

Anne Christin Saulin

from Halle (Saale)

Würzburg, 2022

Submitted on: …………………………………………………………..……..

Office stamp

Members of the Thesis Committee

Chairperson: Prof. Dr. Keram Pfeiffer

Primary Supervisor: Prof. Dr. Grit Hein

Supervisor (Second): Prof. Dr. Martin Herrmann

Supervisor (Third): Prof. Dr. Matthias Gamer

Date of Public Defence: ……………………………………………….………

Date of Receipt of Certificates: ……………………………………………….

Acknowledgements

Firstly, I am very thankful to my primary supervisor Prof. Dr. Grit Hein who greatly supported me throughout this whole highly non-linear journey, gave valuable advice, and taught me a thing or two (or three) about writing a good scientific paper. I am grateful for her patience, her openness towards my own ideas, and for giving me the space to realize them.

I further want to thank my second supervisor, Prof. Dr. Martin Herrmann, who introduced me to the practicalities of the behavioral laboratory and the fMRI scanner, helped me solve the administrative riddles of the psychiatry in general, and was always around when I had questions. Finally, I would like to thank Prof. Dr. Matthias Gamer whose expertise in neuroimaging was invaluable in preparation for and during data collection at the scanner in Würzburg, and whose methodological insights and knowledge have been a true inspiration.

I would also like to thank Prof. Dr. Martin Lotze and his group who welcomed me so warmly to their lab in Greifswald and went above and beyond to help whenever I needed it. Thank you, Ulrike, for being a better scanning buddy, cover story actress, babysitter, and friend than I could ever have hoped for. Thank you also to Annett, Theresa, Miriam, Carina, and Kathrin, Lea, and Larissa, who were reliable confederates and co-experimenters and just wonderful to work with – I literally couldn't have done it without you.

Another big thank you goes out to former and present colleagues at the Translational Social Neuroscience Unit for countless uplifting lunches, fun coffees, conference neologisms, more moral support than words can describe, and for being not only stellar colleagues but great friends. Marthe, Martin, Jasper, Vasko, Anna – you're the best. I also want to thank Yanyan, my first office mate in Würzburg for a cool start together, and Yuqing, our very competent zoom colleague, for her critical thoughts and valuable input.

I am also grateful to the Marcusstraßen colleagues, especially Lea, who taught me about open science and is my mixed-models-savior, and Mario, who was a most collaborative partner in crime at the scanner.

I also want to thank my collaborators, especially Prof. Dr. Jan Engelmann and Dr. Chih-Chung Ting, from whom I learned a lot about reinforcement learning modelling, for their patience, openness, support, and constructive feedback throughout the whole project.

Further, I want to thank the Studienstiftung des deutschen Volkes whose scholarship with the invaluable benefits of flexible hours, a supportive network, and inspiring workshops made coordinating and mastering PhD life with kids possible.

Above all, I thank my friends and family who were with me every step of the way and kept me sane. I thank my mother and Christian, and my grandparents, who supported all my decisions, helped us pack and unpack boxes more than once, and thanks to whom I practiced explaining what I do in non-jargon German (it's hard!). I also thank my children Frieda and Arthur for being patient and understanding when I had to work on the weekend or answer work-related calls on the playground. Most importantly, I am grateful to Eric, the person who over the course of this dissertation became my husband, a wonderful father, moving expert, and even a "Dr." himself. Without your unconditional support, openness, and love, this would not have been possible.

Financial support

This work was supported by a PhD scholarship provided by the German Academic Scholarship Foundation awarded to Anne Saulin and by two Heisenberg Grants (HE 4566/2–1; HE 4566/5–1) provided by the German Research Foundation awarded to Grit Hein.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 declaration of Helsinki.

Conflict of interest

The author declares no competing interests.

Table of Contents

List of figures

List of tables

Abbreviations

ACC	anterior cingulate gyrus
AI	anterior insula
DDM	drift-diffusion modelling
dlPFC	dorso-lateral prefrontal cortex
dmPFC	dorso-medial prefrontal cortex
fMRI	functional magnetic resonance imaging
IFG	inferior frontal gyrus
TPJ	temporo-parietal junction
M	mean
mPFC	medial prefrontal cortex
ms	milliseconds
RL	reinforcement learning
RT	reaction time
RW	Rescorla-Wagner
SD	standard deviation
SE	standard error
vmPFC	ventro-medial prefrontal cortex
VS	ventral striatum

Summary

Empathy, the act of sharing another person's affective state, is a ubiquitous driver for helping others and feeling close to them. These experiences are integral parts of human behavior and society. The studies presented in this dissertation aimed to investigate the sustainability and stability of social closeness and prosocial decision-making driven by empathy as well as other social motives. In this vein, four studies were conducted in which behavioral and neural indicators of empathy sustainability were identified using model-based functional magnetic resonance imaging (fMRI).

Applying reinforcement learning (RL), drift-diffusion modelling (DDM), and fMRI, the first two studies were designed to mathematically understand the temporal evolution of empathy-related social behavior. That is, we investigated the formation and sustainability of empathy-related social closeness (study 1) and examined how sustainably empathy led to prosocial behavior (study 2). Additionally, empathy-related social behavior was compared to social closeness and prosocial decision-making related to reciprocity, i.e., the social norm to return a favor. Using DDM and fMRI, the last two studies investigated how empathy combined with reciprocity on the one hand and empathy combined with the egoistic motive of outcome maximization on the other hand altered the behavioral and neural social decision process.

The results of studies 1 and 2 showed that empathy-related social closeness and prosocial decision tendencies persisted even if empathy was only weakly reinforced. The sustainability of empathy-associated effects was related to a recalibration of the empathy-related social closeness learning signal (study 1) and the maintenance of a prosocial decision bias (study 2). The findings of study 3 showed that empathy influenced the processing of reciprocity-based social decisions, but not vice versa. Study 4 revealed that empathy-related decisions were modulated by the motive of outcome maximization, depending on individual differences in state empathy.

Together, the results in this dissertation provide valuable insights into the mechanisms underlying empathy-related social closeness and decision-making. The studies strongly support the concept of empathy as a sustainable driver of social closeness and prosocial behavior, that is stronger than another important social motive, can enhance prosocial behavior based on other motives, and is resilient to potentially undermining motives.

Zusammenfassung

Empathie, d.h. das Teilen des Affekts einer anderen Person ist eine allgegenwärtige Motivation, anderen Menschen zu helfen und sich ihnen nahe zu fühlen. Diese Erfahrungen sind wesentliche Bestandteile menschlichen Verhaltens und zentral für das Fortbestehen unserer Gesellschaft. Die Arbeiten der hier vorgestellten Dissertation setzten sich zum Ziel, die Nachhaltigkeit und Stabilität von sozialer Nähe sowie prosozialem Entscheidungsverhalten basierend auf Empathie und anderen sozialen Motiven zu beleuchten. Hierfür wurden vier Studien durchgeführt, in denen Verhaltensmaße und neuronale Indikatoren für die Nachhaltigkeit von Empathie unter Einsatz von modellbasierter funktioneller Magnetresonanztomographie (fMRT) erhoben wurden.

Unter Verwendung von Verstärkungslernmodellen, Drift-Diffusionsmodellen (DDM) und fMRT wurden die ersten zwei Studien entwickelt, um die Entwicklung von empathiebasiertem sozialen Verhalten mathematisch zu verstehen. Wir untersuchten somit den Aufbau und die Nachhaltigkeit von empathiebasierter sozialer Nähe so wie empathiebasiertem prosozialen Verhalten. Des Weiteren wurde empathiebasiertes Verhalten mit sozialer Nähe und prosozialem Entscheidungsverhalten basierend auf Reziprozität verglichen, der sozialen Norm, Gefallen zurückzuzahlen. Mit Hilfe von DDM und fMRT wurde in den letzten beiden Studien untersucht, wie Empathie in Kombination mit Reziprozität einerseits und Empathie in Kombination mit dem egoistischen Motiv der Gewinnmaximierung andererseits den verhaltensbezogenen und neuronalen sozialen Entscheidungsprozess verändert.

Die Ergebnisse der Studien 1 und 2 zeigten, dass empathiebasierte soziale Nähe und empathiebasierte prosoziale Entscheidungstendenzen selbst dann fortbestanden wenn Empathie nur noch selten verstärkt wurde. Die Nachhaltigkeit dieser Effekte hing mit der Rekalibrierung des empathiebasierten Lernsignals für soziale Nähe (Studie 1) und dem Aufrechterhalten eines prosozialen Entscheidungsbias zusammen (Studie 2). Die Ergebnisse von Studie 3 zeigten, dass Empathie die Verarbeitung von reziprozitätsbasierten sozialen Entscheidungen beeinflusst, aber nicht umgekehrt. Studie 4 zeigte, dass empathiebasierte soziale Entscheidungen durch das Motiv der Gewinnmaximierung moduliert werden abhängig von individuellen Unterschieden im Empathiezustand.

Zusammengefasst liefern die Ergebnisse der vorliegenden Dissertation wertvolle Einblicke in die Mechanismen, die empathiebasierter sozialer Nähe und sozialen Entscheidungen zu Grunde liegen. Die Studien unterstützen nachdrücklich das Konzept von Empathie als nachhaltige Triebkraft für soziale Nähe sowie prosoziales Entscheidungsverhalten, die stärker ist als ein anderes wichtiges soziales Motiv, prosoziales Verhalten basierend auf anderen Motiven zusätzlich verstärken kann und widerstandfähig gegenüber potentiell unterminierenden Motiven ist.

1 Introduction

Social closeness and prosocial behavior are two key ingredients for cooperation on an individual as well as societal level. The studies in this dissertation aimed at shedding light on the computational and neural processes underlying social closeness and prosocial behavior as induced by empathy, one principal driver of social closeness and prosocial behavior (Batson, 2010; Batson et al., 1991; FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; Grynberg & Konrath, 2020; Hein, Morishima, Leiberg, Sul, & Fehr, 2016; Majdandžić, Amashaufer, Hummer, Windischberger, & Lamm, 2016; Morelli, Rameson, & Lieberman, 2014; Singer & Hein, 2012). Specifically, we investigated the formation and sustainability of empathy-related social closeness as well as the sustainability, benefits, and resilience of prosocial decision behavior related to empathy and other social motives.

The first section introduces the reader to the definition of empathy and the other social motives as well as to how motives were experimentally activated in the studies presented in this dissertation. In the following sections, the tasks and methods used to assess motive-driven behavior and the two computational modelling methods applied for data analysis (the Rescorla-Wagner (RW) learning model and the drift-diffusion model, DDM) are described. The final sections of the introduction focus on the method to measure neural activation in the studies presented (functional magnetic resonance imaging, fMRI) and outlines the results of previous studies relevant to the research questions addressed in this dissertation.

1.1 Social motives

Social behavior, such as helping a friend move houses, carrying an elderly lady’s groceries, or sharing study notes, is ubiquitous in our daily lives. It is also key to a peaceful human coexistence and societal stability. Whether a person decides to be prosocial is decisively determined by the person’s current motivation to act prosocially towards a specific other person. One key motive that drives prosocial behavior is the empathy motive (Batson, Ahmad, & Stocks, 2011; Cialdini et al., 1987; Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016; Preston, 2007), that is the motive of sharing another person’s affective state that elicits the goal to increase the well-being of that other person (Batson, 2010).

While motives can sometimes be understood in terms of traits, i.e., stable person characteristics (McClelland, 1985, 2014), they can also be defined as transient drivers of behavior that can be more or less active in a given situation (Kruglanski et al., 2018; Lewin, 1951). In this work, motives are understood in terms of the latter definition. This allowed us to experimentally activate empathy and other motives in order to investigate participants' motive-related social closeness and social decision-making behavior. The formation and sustainability of empathy-related social closeness and prosocial behavior is the focus of this dissertation. However, motives rarely act in isolation. Thus, another important social motive was investigated in this dissertation: the reciprocity motive, i.e., the social norm to return a previously given favor (Gouldner, 1960). Additionally, the combination of empathy with the egoistic motive of outcome maximization, another frequent driver of prosocial behavior (Batson & Shaw, 1991; Cutler & Campbell-Meiklejohn, 2019; Tabibnia & Lieberman, 2007), was investigated. In the following sections, these three motives are introduced in more detail.

Empathy

Empathy is a multi-dimensional construct, and researchers differ widely in how they define their working concepts of empathy. While some stress the difference between cognitive and emotional empathy (Harari, Shamay-Tsoory, Ravid, & Levkovitz, 2010; Perry & Shamay-Tsoory, 2013), others highlight its distinctiveness from compassion (Bloom, 2017; Singer & Klimecki, 2014) and theory of mind (Böckler, Kanske, Trautwein, & Singer, 2014; Kanske, Böckler, Trautwein, Lesemann, & Singer, 2016). However, what most of them agree on is the notion that certain facets of empathy can drive prosocial behavior. Batson most prominently coined this so called empathy-altruism hypothesis (Batson, 2010; Batson, Ahmad, & Tsang, 2004; Batson et al., 1991) making a strong case for empathy as a key driver for behaving prosocially. Since then a lot of works, particularly those that operationalised empathy for pain, have demonstrated that explicit activation of the empathy motive towards another person increased participants' self-reported empathy and subsequent prosocial decision behavior towards that person (Gu & Han, 2007; Hein, Engelmann, Vollberg, & Tobler, 2016; Hein, Morishima, et al., 2016; Klimecki, Mayer, Jusyte, Scheeff, & Schönenberg, 2016; Singer & Lamm, 2009). In the studies presented in this dissertation, we have built on these works and explicitly activated participants' empathy motive towards an interaction partner. This is commonly accomplished by participants observing that an interaction partner repeatedly receives painful stimulation (e.g., Hein, Morishima, et al., 2016). To behaviorally assess participants' empathic response, they typically report their emotional reaction on an analogue scale, i.e., a self-report measure of how they feel after observing the other's stimulation. The worse participants feel in response to the other's painful stimulation, the higher the empathic response (Böckler et al., 2014; Hein, Morishima, et al., 2016; Lamm, Batson, & Decety, 2007). To facilitate affect sharing, participants themselves experienced the same painful stimulation beforehand.

Reciprocity

The reciprocity motive is based on the social norm to reciprocate previously received helping behavior (Falk & Fischbacher, 2006; Fehr & Gächter, 2000; Gouldner, 1960) and is essential in building and maintaining cooperation across society in general (Axelrod & Hamilton, 1981). As such, the reciprocity motive is one principal driver for prosocial behavior and fosters feelings of social closeness (Adams & Miller, 2022; Fehr, Fischbacher, & Gächter, 2002; Hein, Morishima, et al., 2016). In general one can distinguish between positive and negative reciprocity, as well as direct and indirect reciprocity (Fehr & Gächter, 2000; Nowak, 2006; Perugini, Gallucci, Presaghi, & Ercolani, 2003). Positive reciprocity describes the norm of reciprocating acts of kindness with kind behavior, whereas negative reciprocity describes the norm of reciprocating unkind behavior with unkind behavior in return (Chernyak, Leimgruber, Dunham, Hu, & Blake, 2019; Kaltwasser, Hildebrandt, Wilhelm, & Sommer, 2016). Moreover, reciprocity can be direct, meaning that a favor is directly returned to the person who has paid the favor in the past. Indirect reciprocity, however, describes acts of reciprocity whereby a previously paid favor is indirectly repaid to a different person. Indirect reciprocity strongly builds on the assumption of a shared social norm of reciprocity and cooperation in society at large (Hilbe, Schmid, Tkadlec, Chatterjee, & Nowak, 2018; Simpson & Willer, 2008). In the works of this dissertation, we focussed on reciprocity as driver for prosocial behavior towards a specific previous interaction partner, i.e., positive direct reciprocity. Previous works have shown that activating the reciprocity motive by paying someone a favor, e.g., by cooperating in an economic game (McCabe, Rigdon, & Smith, 2003) or forgoing a monetary reward to spare someone from pain (Hein, Morishima, et al., 2016), increases this person's likelihood of making prosocial decisions in subsequent interaction tasks. Based on such previous works, we activated the reciprocity motive towards an interaction partner by having participants observe that this interaction partner repeatedly forwent a monetary reward in order to save the participant from a painful stimulation. Analogously to the empathy motive, the reciprocity response was assessed using a self-report scale asking participants how the feel in response to the other's decision to forgo a monetary reward to save the participant from pain. Previous research showed that the better participants feel in response to the other person's decision to help, the stronger the reciprocity motive (Hein, Morishima, et al., 2016).

Egoism

In contrast to empathy and reciprocity, egoism may not be primarily perceived as a motive that elicits prosocial behavior. However, egoistic motives as incited by monetary rewards can lead to prosocial behavior and can be considered a social motive (Batson et al., 2011; Besley & Ghatak, 2018; Cialdini et al., 1987). When driven by egoism, the goal of the prosocial behavior is not to improve the well-being of the other person (as is the case for empathy) but to improve one's own well-being. Egoism can hence incite prosocial behavior in situations in which the action that improves one's own well-being aligns with the behavior that improves the well-being of the other person. Inspired by previous work, we activated the egoism motive by offering participants a monetary bonus that was additionally paid out if they behaved prosocially (Balliet, Mulder, & Van Lange, 2011; Besley & Ghatak, 2018). Previous studies have demonstrated that reputation is an important confounding factor for effects of monetary incentives (Engelmann & Fischbacher, 2009; Exley, 2018; Izuma, Saito, & Sadato, 2008). Hence, to reduce potential influences of reputation and carve out the effect of the monetary incentive alone, we offered the monetary bonus to the participant in private.

1.2 Measuring the effects of motive activation

The studies outlined in this dissertation made use of two principal measures for motive-related behavior: ratings of social closeness and social decision-making behavior (see **Figure 1.2.1** for visualization). The first measure yielded continuous trial-by-trial indications of social closeness. The second measure aimed at assessing participants' likelihood for prosocial behavior driven by empathy, reciprocity and/or egoism and yielded binary trial-by-trial results (prosocial decision vs. egoistic decision). Both measures and the underlying assumptions are discussed in the following two sections.

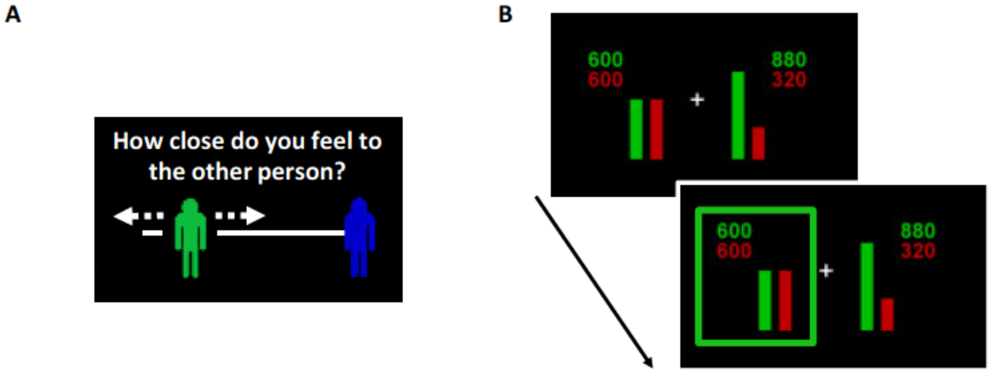


Figure 1.2.1 Assessing motive-related social behavior. **A** To measure social closeness, participants indicated how close they felt to the other by moving their own mannequin (green) closer to or further away from the other mannequin representing the respective partner (here blue) using a continuous slider scale. **B** To measure prosocial decision behavior, participants in each trial chose between a prosocial and an egoistic distribution of points between themselves (green point values) and the respective partner (here red point values). The option chosen by the participant was highlighted with a green rectangle. In this example, the participant chose the prosocial point distribution.

Social closeness

Previously, measuring trial-by-trial social closeness has been successfully applied to capture processes of social learning (Zhou et al., 2022). Moreover, social closeness is strongly associated with empathy-related and reciprocity-related social behavior and concurrent neural activation (Bohnet & Frey, 1999; Hoffman, McCabe, & Smith, 1996; Strombach et al., 2015). The closer someone feels to another person, the more likely they are to choose a more prosocial division of points in the dictator game (Bohnet & Frey, 1999; Hoffman et al., 1996), the more money they are willing to forego for another person's gain (B. Jones & Rachlin, 2006; Strombach et al., 2015), and the more empathic affect sharing they engage in (Beeney, Franklin, Levy, & Adams, 2011; Grynberg & Konrath, 2020). Previous work hence demonstrated that social closeness is positively associated with empathy. There is also evidence that feelings linked to reciprocal behavior can promote feeling closer to another person (Adams & Miller, 2022). Thus, explicitly activating empathy or reciprocity should increase feelings of social closeness. We were thus able to use the same measures of social closeness to investigate social closeness sustainability based on the empathy and based on the reciprocity motive. Additionally, this measure provided an indicator of how close participants felt to their interaction partner at any given timepoint on a continuous scale. In particular, we hypothesized that the more sustainably empathy induces social closeness, the less social closeness should decrease when empathy is only rarely reinforced.

Social decision-making and the dictator game

In everyday life, humans face multiple situations in which they have to make social decisions, e.g., whether to hold the door for someone or not, whether to share the last piece of chocolate with a sibling or eat it all alone, or whether to share one's apartment with a refugee family or not. All these scenarios are examples for social decision-making (for reviews, see e.g., Rilling & Sanfey, 2011; Van Dijk & De Dreu, 2021). In behavioral economics, various tasks, termed economic games, have been developed to assess and quantify people's (social) decision preferences in different scenarios. Based on the idea of the *homo oeconomicus* (Mill, 1836), behavioral economists assumed that in each game, participants should only make rational decisions such that their own utility is maximized, i.e., make decisions that result in the most money for themselves (von Neumann & Morgenstern, 1944). However, what economists actually observed was that participants did not strictly adhere to this assumption but oftentimes decided prosocially instead (Camerer, 2003; Eckel & Grossman, 1996; Forsythe, Horowitz, Savin, & Sefton, 1994). Decisively, the more pronounced an individual's motivation for prosocial behavior, the more frequent and pronounced were prosocial decisions (Ben-Ner & Kramer, 2011; Edele, Dziobek, & Keller, 2013; Hein, Morishima, et al., 2016; Klimecki et al., 2016; Schier, Ockenfels, & Hofmann, 2016). Economic games hence provide a quantitative measure for prosocial behavior that is sensitive to a participant's motivational state.

For the studies in this dissertation, we used an adopted version of one specific economic game, the dictator game, to assess participants' changes in motive-driven prosocial behavior. In the original version of the dictator game (Forsythe et al., 1994), an allocator the dictator divides a given amount of money between herself and another person. In the binary version of the dictator game used in this dissertation, participants can repeatedly choose between two predefined distribution options, one of which yield a relatively more prosocial whereas the other yields a relatively more egoistic distribution (e.g., Chen & Krajbich, 2018; Hein, Morishima, et al., 2016; Hutcherson, Bushong, & Rangel, 2015). It has been shown that the specific distribution options influence participants' decision preferences (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Schmidt, 1999). Specifically, the higher the potential gain for the self associated with a particular distribution option, the more likely it is that a participant will choose this option. Likewise, the larger a potential gain for the other associated with a particular distribution option, the more likely is a participant to choose this option. Specifically, participants are more likely to choose the relatively more egoistic option in situations of so-called disadvantageous inequality (smaller initial pay-off for the allocator in both distribution options, e.g., Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012). Advantageous inequality, however, describes the situation in which for both distribution options in each trial, the allocator would receive more than the receiver, i.e., the initial pay-off is always higher for the allocator. Nonetheless, one option is more prosocial than the other option in that it maximizes the gain for the receiver relative to the less prosocial option. In the works of this dissertation, we aimed at investigating the decision process underlying prosocial decisions. Hence, to maximize participants' likelihood to make a prosocial decision, participants in the studies of this dissertation performed the dictator game with distribution options yielding advantageous inequality. Using this optimized variant of the binary dictator game, we investigated the sustainability of the prosocial decision process driven by the empathy motive in comparison to and combination with the reciprocity motive and the egoistic motive of outcome maximization.

1.3 Computational models

The directly observable information that can be obtained based on the two measures outlined above provide helpful indicators of how close someone feels to a given point in time or how likely and how fast decides prosocially after different motive activation procedures. However, they do not offer mechanisms that may subserve the respective behavior. Thus, to gain a better understanding of the potential mechanisms underlying empathy-related sustainability in the studies of this dissertation, we used the method of computational modelling.

Computational modelling is a technique by which a certain phenomenon or behavior can be mechanistically described by means of mathematical formulations. Computational models in cognitive neuroscience synthesize the information provided by different outcome measures (e.g., ratings or choices made and reaction times) and this way provide valuable insights into the mechanisms shaping the respective behavior under investigation. In recent years, computational modelling approaches have increasingly been applied in the field of social neuroscience (Charpentier & O'Doherty, 2018; Decety, Jackson, Sommerville, Caminade, & Meltzoff, 2004; Forstmann, Ratcliff, & Wagenmakers, 2016; Hein, Engelmann, et al., 2016; Hutcherson et al., 2015; Lockwood, Apps, Valton, Viding, & Roiser, 2016) and have furthered the understanding of the neural social decision-making process (e.g., Hutcherson et al., 2015) and instances of social learning (e.g., Hein, Engelmann, et al., 2016). In the studies outlined in this dissertation, two different modelling techniques were applied. Each of these techniques allowed for closer investigation of the potential mechanisms underlying participants' empathy-driven social closeness (Rescorla-Wagner learning model, study 1) and empathy-driven social decision process (drift-diffusion model, studies 24). In the following two sections, each modelling technique will be introduced in detail.

The Rescorla-Wagner learning model

Associative learning, that is, forming an association between two concepts by continuously updating their coinciding, is common in daily life. One frequently used model to mathematically describe such a process of learning specific stimulus-outcome associations (e.g., learning to associate a certain stimulus with a rewarding outcome such as a monetary gain or with a punishing outcome such as a monetary loss) is the Rescorla-Wagner learning model (Rescorla & Wagner, 1972).

In the basic Rescorla-Wagner model, the estimated association strength *V* at trial *t* is updated with prediction error δ and free parameter α only. Specifically, the prediction error is calculated as difference between the actual outcome and the expectation, i.e., the association strength from the previous trial:

$$\delta_t = R_t - V_{t-1} \tag{1}$$

In equation 1, R_t refers to the actual outcome: for example, R_t can be set to 1 for reinforced trials (i.e., reward or gain) and to 0 for non-reinforced trials (i.e., punishment or loss) at trial *t*. The prediction error δ from the current trial *t* is then used to update the value *V* for this trial weighted by the learning rate α:

$$V_t = V_{t-1} + a \times \delta_t \tag{2}$$

The learning rate α is a free parameter bounded between 0 and 1 and reflects to what extent more recent feedback influences the learning process. The larger this parameter, the more the most recent feedback dominates over previously received feedback.

Based on this basic model, extensions can be developed serving the research question at hand. Garrett & Daw (2020) for example observed that in a foraging learning task, a differential model that assumes separate updating of positive and negative prediction errors can more accurately describe participants' behavior than assuming only one learning rate. In their model, the prediction error was calculated as in equation (1), but learning rates depended on whether the present trial resulted in a negative or a positive prediction error. That is, a positive δ is multiplied by learning rate α^+, and a negative δ is multiplied by learning rate α^- to update V:

$$V_t = \begin{cases} V_{t-1} + \alpha^+ \times \delta_t \, if \; \delta > 0 \\ V_{t-1} + \alpha^- \times \delta_t \, if \; \delta \; < 0 \end{cases} \tag{3}$$

This distinction revealed that participants weighted trials of rewards (i.e., experiences resulting in a positive prediction error) more strongly than trials of losses (i.e., experiences resulting in a negative prediction error).

Another extension of the basic model was applied by Palminteri and colleagues (2015). They assumed that the outcome values of the respective feedback (i.e., *R* = 1 for reinforcer feedback and *R* = 0 for non-reinforcer feedback) may be recalibrated depending on the context in which they are learnt. In this model, the proposed outcome value is recalibrated by subtracting an additional free parameter ω ϵ [0,1] (see equation 4).

$$\mathrm{\delta_t} = |\mathrm{R_t} - \omega_t| - \mathrm{V_{t-1}} \tag{4}$$

Hence, according to this model, an individual's actual outcome value for reinforced trials corresponds to 1 minus the individual recalibration value ω_t, and the actual outcome value of a non-reinforced trial corresponds to ω_t. Thus, the larger the value of ω_t, the smaller the prediction error associated with reinforced trials and the larger the prediction error associated with non-reinforced trials. The study showed that this relative model (equation 4) which accounts for the context (here the reward associated with the option the participant did not choose) better described participants' behavior than the simple learning model without recalibration (equation 2).

More recently, the Rescorla-Wagner learning rule was successfully used to describes processes of social learning, such as imitation learning (Najar, Bonnet, Bahrami, & Palminteri, 2020), learning about whether another person searches social contact (R. M. Jones et al., 2011), learning about other people's personalities (Frolichs, Rosenblau, & Korn, 2021), or learning to empathize with outgroup members (Hein, Engelmann, et al., 2016).

In this dissertation, we built on this idea and tested which out of the three variants (simple learning model, differential model, and recalibration model) can best explain the temporal evolution of empathy-related social closeness.

The drift-diffusion model

The drift-diffusion model (DDM) is one of the most prominent representative of so-called sequential sampling models (Forstmann et al., 2016). DDMs formalize the choice process as noisy accumulation of evidence towards two different choice option boundaries (Ratcliff & McKoon, 2008). That is, when faced with two choice options, participants start to accumulate evidence towards these options. Once they have accumulated sufficient evidence for one of these options, this option is chosen. This process can be parametrized in terms of three principal components: the speed at which evidence is accumulated (*v*-parameter), the bias towards one of the options before entering the choice process (*z*-parameter), and the amount of relative evidence required in order to reach a decision (decision threshold, *a*-parameter). In addition to these principal parameters, the choice process can be characterized in terms of the so-called non-decision time t_0 as well as the trial-by-trial variability of the aforementioned components (*sv*, *sz*, *sa*, *st*). In order to estimate these parameters, the various estimation approaches (Vandekerckhove & Tuerlinckx, 2008; Voss & Voss, 2007; Wiecki, Sofer, & Frank, 2013) synthesize the information obtained from participants' trial-by-trial choices (in this dissertation: participants' prosocial vs. egoistic decisions) and reaction times.

The *v*-parameter describes the speed at which information is accumulated towards the different choice options. Hence, this parameter provides an indicator for the efficiency of the choice process itself. Previous studies demonstrated that the *v*-parameter is sensitive to basic task affordances. The more difficult the task is , e.g., a perceptual discrimination task, the smaller is the *v*-parameter (Voss, Rothermund, & Voss, 2004). In the realm of social decision-making, the *v*-parameter was modulated by various factors such as information about other's choice behavior (Yu, Siegel, Clithero, & Crockett, 2021), the degree of self-relevance of the choice options at hand (Bottemanne & Dreher, 2019; Falbén et al., 2020), and social motivation (Leong, Hughes, Wang, & Zaki, 2019). Assuming a similar modulation of the v-parameter after activation of the empathy motive , an empathy-related increase of the v-parameter in the studies of this dissertation would imply that empathy increases the efficiency of the prosocial decision process itself.

The z-parameter reflects the initial choice bias, i.e., the degree to which an individual prefers one of the choice options prior to making the choice. Thus, in contrast to the v-parameter, which models the choice process itself, the z-parameter models the individual bias with which a person enters the choice process. Generally, changes in this parameter have been associated with manipulation of the reward structure (Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; Voss et al., 2004). That is, choices are biased toward the option associated with the higher reward. In studies investigating social decision-making, peer behavior (Germar, Albrecht, Voss, & Mojzisch, 2016; Toelch, Panizza, & Heekeren, 2018), personal preferences (Chen & Krajbich, 2018), and situational factors such as the relationship between the decider and the other people involved (Son, Bhandari, & FeldmanHall, 2019) modified the *z*-parameter. Toelch et al. (2018) for example observed that participants' choices were biased towards the option chosen by the majority. Other work showed that when faced with the binary choice to allocate points in a prosocial (relatively more points for the other person, prosocial choice option) or an egoistic way (relatively more points for the participant herself, egoistic choice option), prosocial individuals exhibited a bias towards the prosocial choice options whereas more egoistic individuals exhibited a bias towards the egoistic choice options (Chen & Krajbich, 2018). In the context of the studies presented here, for a person with a strong initial bias towards making prosocial choices (reflected by a large value of the *z*-parameter), the starting point of the choice computation is located closer to the prosocial choice boundary, and as a result this person is more likely to choose the prosocial option. Hence, if activation of empathy increased the *z*-parameter, this would imply that empathy increases an individual's initial bias towards making a choice prior to the choice process itself.

The third component, the *a*-parameter or decision threshold, quantifies the amount of relative evidence that is required to choose one of the options, and hence provides a measure of response caution. That said, it reflects a participant's speed-accuracy trade-off, with larger *a*-parameters indicating a stronger emphasis on accuracy over speed (Voss et al., 2004). This parameter is generally influenced by instruction manipulations such as telling participants that it is important that they respond correctly (e.g., Katsimpokis, Hawkins, & van Maanen, 2020; Zhang & Rowe, 2014). In social decision-making, the *a*-parameter was thus far not prominently linked to specific factors (see Son et al., 2019, for an exception). However, the *z*-parameter and the *a*-parameter are closely associated and both influence similar properties of the reaction time distribution (Ratcliff, Smith, Brown, & McKoon, 2016). If activation of empathy increased the *a*-parameter, this would imply that empathy increases an individual's response caution while making the decision to act prosocially or egoistically.

The two modelling approaches applied in the studies of this dissertation can be used in order to better understand the neuro-computational mechanisms associated with social learning and social decision-making by relating functional brain activation to specific modelling parameters (Forstmann et al., 2016; Lockwood & Klein-Flügge, 2021). A short introduction to functional magnetic resonance imaging (fMRI) and its application in the studies conducted as part of this dissertation is provided in following section.

1.4 Neural correlates of motive-driven behavior

Functional magnetic resonance imaging (fMRI)

In the studies presented in this dissertation, neural activation was assessed using functional magnetic resonance imaging (fMRI). This neuroimaging technique relies on the blood oxygenated level-dependent response (BOLD response) for localizing neural regions of increased activation (Haacke et al., 1997; Logothetis, 2002; Ogawa, Lee, Kay, & Tank, 1990). The BOLD response is based on the cerebral blood flow, the co-dependent local cerebral blood volume, and the cerebral metabolic rate of oxygen consumption in regions of increased neuronal activity (but see e.g., Blockley, Griffeth, Simon, & Buxton, 2013, for additional determinants of the BOLD response). This change in oxygen level and blood flow in turn alters the nuclear spin of the hydrogen molecules in the blood. Cerebral blood in regions of recent neural activation thus contains hydrogen molecules with different nuclear spin properties than in regions of comparably less neural activation. Magnetic resonance signals depend on spin echoes and thus differ between those regions. This allows for the imaging of the contrast between regions of high neural activation and regions of low neural activation, for example, in response to a certain stimulus. Such event-related fMRI measurements (Buckner, 1998; Friston et al., 1998) have enabled a wealth of neuroscientific research linking cognitive processes to neural activation on a trial-by-trial basis. In the works of this dissertation, we took advantage of this technique by studying the neural activation linked to the emotional reaction during the development of empathy-related social closeness as well as neural activation when making motive-based social decisions.

Neural correlates of empathy for pain and empathy-related behavior

Generally, the extent of an individual's empathy for another's pain has frequently been associated with neural activation in specific brain regions. Most prominently, increased activation in the AI, the inferior frontal gyrus (IFG), and the ACC have been linked to a person's response to another individual's painful experience (Hein, Engelmann, et al., 2016; Hein, Morishima, et al., 2016; Lamm et al., 2007; Lamm, Decety, & Singer, 2011; Y. Li et al., 2021; Naor et al., 2020; Singer & Lamm, 2009; Völlm et al., 2006). The larger an individual's emotional reaction to another's pain was, the larger was the neural activation in the AI, IFG, and the ACC. Neural activation in these regions while witnessing another person in pain was also predictive of future prosocial behavior (Masten, Morelli, & Eisenberger, 2011; Morelli et al., 2014). Additionally, together with the supplementary motor area, the AI and the ACC were identified as the core regions of empathy (Fan, Duncan, de Greck, & Northoff, 2011). Moreover, neural regions comprising the mentalizing network, i.e., dmPFC, TPJ, STS, posterior cingulate cortex, precuneus, and temporal poles, have been associated with empathic reactions to another's painful experience (Bruneau, Pluta, & Saxe, 2012; Dvash & Shamay-Tsoory, 2014; Lamm et al., 2007; Lieberman, 2007; Shamay-Tsoory, 2011; Singer et al., 2004).

Research more specifically related to empathy-based social learning has observed that, AI activation was associated with an empathy-related learning prediction error when learning to empathize with an outgroup member (Hein, Engelmann, et al., 2016) and stronger IFG activation was linked to stronger reappraisal of an empathic reaction (Naor et al., 2020). In addition to signalling the increase in empathy or the adaptation of the empathics response, neural activation in these two regions may hence also be linked to the sustainability of empathy-related social closeness as assessed in study 1 of this dissertation.

In the work that served as a basis for the studies of this dissertation and investigated motive-related social decision-making, neural activation in the AI, the ACC, and the ventral striatum (VS) was increased when participants made prosocial choices in favor of a person who had previously received painful stimulation as compared to making prosocial choices in favor of a person towards whom empathy was not explicitly activated (Hein, Morishima, et al., 2016). These results indicate that these regions are more strongly involved when prosocial behavior is driven by the empathy motive as compared to when prosocial behavior is solely based on 'home-grown' motivation to act prosocially. Neural activation in these regions may hence also be indicative of the sustainability of the empathy-related social decision-making process, as assessed in studies 2-4 in this dissertation.

Neural correlates of decision behavior related to reciprocity and the egoistic motive of outcome maximization

Compared to empathy-based behavior, the neural bases of reciprocal behavior is less clearly defined since different neural and behavioral models exist to explain acts of reciprocity (Rilling & Sanfey, 2011). One approach implies the human reward system with the vmPFC as one key neural region (Hare, Schultz, Camerer, O'Doherty, & Rangel, 2011; Kable & Glimcher, 2007; Strait, Sleezer, & Hayden, 2015). In this model, reciprocity increases the value an individual associates with long-term mutual cooperation, which is reflected by increased neural activation in vmPFC (Wood, Rilling, Sanfey, Bhagwagar, & Rogers, 2006). According to this view, the act of reciprocating is inherently rewarding for people who strongly act according to this motive. Moreover, strategic prosocial decisions which may be driven by reciprocity have been associated with increased neural activation in the striatum and especially the anterior vmPFC (Cutler & Campbell-Meiklejohn, 2019). Interestingly, neural activation linked to maximizing one's outcome, i.e., neural responses to stimuli yielding larger outcome value for the participant have implicated a similar network of neural regions. That is, stimuli that are associated with a large outcome value yield increased neural responses in the striatum and vmPFC as compared to stimuli with small outcome values (U. Basten, Biele, Heekeren, & Fiebach, 2010; Hare, Camerer, Knoepfle, & Rangel, 2010; Strait et al., 2015). Additionally, neural activation in the dlPFC has been associated with the accounting for non-immediate reward during decision-making (Hare, Hakimi, & Rangel, 2014). Work directly related to the question addressed in study 4, i.e., the question of whether empathy-based prosocial behavior can be undermined by the egoistic motive of outcome maximization, have linked the striatum and dlPFC to changes in social behavior due to financial incentives. Specifically, the undermining effect of financial incentives on prosocial behavior was associated with decreased neural activation in the striatum and the dlPFC (Murayama, Matsumoto, Izuma, & Matsumoto, 2010). These two regions are generally implicated in valuation of stimulus outcomes (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005) and self-control (Hare, Camerer, & Rangel, 2009; Schmidt et al., 2018).

Another explanatory mechanism for reciprocity interestingly also suggests the involvement of self-control (Rilling & Sanfey, 2011). According to this view, acts of reciprocity are driven by obviating the guilt associated with acting against the social norm of reciprocating kind behavior. Hence, neural structures that reflect feelings of guilt, as they may be implicated when breaking a promise, may be associated with reciprocal behavior. Studies investigating such behavior have observed that increased neural activation in the ACC and the dlPFC was related to breaking a previously given promise to cooperate (e.g., Baumgartner, Fischbacher, Feierabend, Lutz, & Fehr, 2009).The ACC and the dlPFC, as discussed above, have been implicated in processes of conflict monitoring and self-control, respectively, with increased activation reflecting the increased need for conflict-monitoring or self-control (Botvinick, 2007; Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006).

Yet other studies have explicitly activated the reciprocity motive before participants performed a social decision-making paradigm (Hein, Morishima, et al., 2016) or focussed on neural activation associated with the act of reciprocating in the trust game (van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2009). In the study by Hein and colleagues (2016), participants frequently observed that another person decided to forgo a monetary reward in order to save the participant from a painful stimulation. Hence, this other person paid the participant a favor that she may want to repay, which is the essence of the reciprocity motive as driver for cooperative behavior (Nowak, 2006). In the task following the motive activation, participants repeatedly chose between an egoistic and a prosocial option to divide points between themselves and that other person (i.e., the reciprocity partner). Hein and colleagues (2016) observed that neural activation during the decision to choose the prosocial option towards the reciprocity partner was increased in the ACC, the AI, and the VS compared to prosocial choices for a baseline partner towards whom the reciprocity motive was not activated. Hence, this work suggests that reciprocity and empathy increases neural activation during the prosocial decision-making process in overlapping neural systems.

Taken together decision behavior based on reciprocity as well as based on the egoistic motive of outcome maximization have frequently been associated with neural activation in neural regions linked to value computations (striatum and vmPFC) as well as regions linked to conflict monitoring and self-control (ACC and dlPFC).

1.5 Objectives

In the studies conducted for this dissertation we aimed at investigating the sustainability of empathy with respect to empathy-based social closeness and the stability and resilience of empathy-based prosocial behavior alone and in combination with other drivers of social behavior. More specifically, we investigated how sustainably empathy leads to social closeness and prosocial behavior (in contrast to social closeness and prosocial behavior based on reciprocity) and whether empathy-based prosocial behavior is enhanced or undermined when combined with reciprocity or the egoistic motive of outcome maximization.

In their study, Hein and colleagues (2016) used a between-subject paradigm, in which participants performed a social decision task either following an induction of the empathy motive or following an induction of the reciprocity motive. In the first part of this dissertation encompassing studies 1 and 2, we combined the design developed by (Hein, Morishima, et al., 2016) with RL paradigms (Dunsmoor et al., 2018; Shiban, Wittmann, Weißinger, & Mühlberger, 2015) and optimized the induction phase in correspondence with RL paradigms. That is, we controlled how strongly participants' empathy motive or reciprocity motive was induced in each block, varying the frequency of motive reinforcing events (i.e., empathy: the frequency of observing the interaction partner receive painful stimulation; reciprocity: the frequency with which the interaction partner forgoes a monetary reward to save the participant from a painful stimulation). This enabled us to model the temporal evolution of empathy-based and reciprocity-based social closeness in situations of frequent and rare motive reinforcement along with their neural correlates using RL models (study 1). Additionally, we aimed to assess which components of the social decision process change after frequent compared to after subsequent rare reinforcement of the underlying social motive (study 2). This approach allows us to gain an understanding of how empathy-related social closeness is formed, how sustainable it is (also in comparison to reciprocity-related social closeness), and to what extent motive activation strength affects prosocial decision behavior. The results may provide valuable information about how sustainable empathy-related social closeness and prosocial behavior are, and to what extent empathy-related prosocial behavior is sensitive to the respective underlying motive strength.

In part 2 of this dissertation, we addressed effects of motive combination on the prosocial decision process and developed a within-subject design paradigm (study 3) in which participants performed the social decision task following an induction of the empathy motive (empathy partner), the reciprocity motive (reciprocity partner), both motives (multi-motive partner), and no motive induction (baseline partner). In our analyses, we modelled participants' decision behavior towards the different interaction partners using DDM and identified which choice components were affected by motive combination, including their neural correlates. Based on these results, we could provide first insights into the neuro-computational mechanisms underlying the combination and possible interaction of different social motives as well as the influence of empathy on the prosocial decision process relative to reciprocity.

Testing how the combination of empathy and the motive of outcome maximization influences the social decision process, we performed a final experiment (study 4), in which participants performed the social decision task following an induction of the empathy motive (empathy condition) and following the induction of both, the empathy motive and the motive of outcome maximization, i.e., offering a monetary bonus for making prosocial decisions in the clear majority of the trials (empathy-bonus condition). In the analyses, we again modelled participants' decision behavior in the two conditions using DDM and identified which choice components of the empathy-driven social decision process and concurrent neural activation were affected by offering a financial incentive. Based on these results, we were able to test whether empathy-driven prosocial decision behavior is resilient to the additional activation of the motive of outcome maximization.

2 Manuscripts and publications

In accordance with the goals formulated above, four studies were conducted investigating different aspects of empathy sustainability with respect to social closeness as well as prosocial decision-making. The content of the two studies that are published (studies 3 and 4) corresponds to the published version in the respective journal but has been edited to fit the formatting of this dissertation.

After each manuscript, the implications of the results obtained as well as the open question relevant for the subsequent manuscript is shortly discussed.

Study 1: Saulin, A., Ting, C.-C., Engelmann, J.B., & Hein, G. Empathy induces sustained social closeness.

Materials: https://github.com/AnneSaulin/empathy_sustainability

Study 2: Saulin, A. & Hein, G. Empathy incites a sustainable prosocial decision bias.

Materials: https://github.com/AnneSaulin/empathy_sustainability

Study 3: Saulin, A., Horn, U., Lotze, m., Kaiser, J., & Hein, G. (2022). The neural computation of prosocial decisions in complex motivational states. *NeuroImage,* 247, 118827. https://doi.org/10.1016/j.neuroimage.2021.118827

Materials: https://github.com/AnneSaulin/complex_motivations

Study 4: Iotzov, V., Saulin, A., Kaiser, J., Han, S., & Hein, G. (2022). Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females. *Social Neuroscience*. https://doi.org/10.1080/17470919.2022.2115550

Materials: https://github.com/Vassil-Iotzov/empathy_incentives

2.1 Empathy induces sustained social closeness

Anne Saulin[1*], Chih-Chung Ting[2], Jan B. Engelmann[3], & Grit Hein[1*]

[1] Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, 97080 Würzburg, Germany.

[2] Department of Psychology, University of Hamburg, 20146 Hamburg, Germany

[3] CREED, Amsterdam School of Economics (ASE), Universiteit van Amsterdam, 1018 WB Amsterdam, the Netherlands

*corresponding authors:
Anne Saulin, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Saulin_A@ukw.de

Prof. Dr. Grit Hein, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Hein_G@ukw.de

Abstract

Empathy generates the feeling of social closeness which is key for connecting humans on the individual and the societal level. However, despite its importance for everyday life, it is unclear how empathy-related social closeness is formed and how sustainable it is. Here we applied an acquisition-extinction paradigm, combined with computational modelling and fMRI to investigate the formation and sustainability of empathy-related social closeness. Participants observed painful stimulation of another person with high probability (acquisition phase), low probability (extinction phase) and at chance level (control blocks) and rated their closeness to the other person. The results of two independent studies showed an increase in social closeness in the acquisition phase that persisted in the extinction phase. Providing insights into the underlying mechanisms, reinforcement learning modelling revealed a recalibration of the observed feedback value allowing for an increase in social closeness based on observing another's pain as well as non-pain. The results of a control study in which we induced a different social motive showed that the observed effects and learning mechanisms were specific for empathy-related social closeness. On the neural level, the recalibration of the feedback signal was associated with neural responses in anterior insula and adjacent inferior frontal gyrus and the bilateral superior temporal sulcus/temporo-parietal junction (TPJ), modulated by individual differences in trait empathic concern and mentalizing, respectively. Taken together, our studies demonstrate that empathy-related social closeness persists even if the other person is no longer suffering and provides insights into the computational and neural mechanisms that drive the longevity of empathy-related effects. These finding are important, because they show that once empathy is activated, empathy-related responses are a robust driver of social closeness.

keywords:
empathy, social closeness, Rescorla-Wagner model, fMRI, STS/TPJ, IFG

Introduction

Empathy enables us to share another's emotions and thereby provides an important way to connect with other people. For example, there is abundant evidence that observing another's pain results in an empathic reaction (Lamm et al., 2011), increases the perceived closeness to others (Beeney et al., 2011), and predicts prosocial behaviour towards the suffering person (Batson, 2010; Hein, Morishima, et al., 2016; Saulin, Horn, Lotze, Kaiser, & Hein, 2022). However, it remains unknown how long these relationship-enhancing effects of empathy last. In other words, does empathy-induced closeness prevail after the other person's suffering is relieved or does it decay?

Empathy itself is a multidimensional construct (Timmers et al., 2018). Commonly, researchers distinguish between so called cognitive empathy or theory of mind (ToM) and emotional empathy – a distinction that is even mirrored on a neural level (Cox et al., 2012; Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Kanske, Böckler, Trautwein, & Singer, 2015; Preckel, Kanske, & Singer, 2018; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). Cognitive empathy has often been associated with neural activation of the medial prefrontal cortex (mPFC), the superior temporal sulcus (STS), the temporal poles (TP), and the temporo-parietal junction (TPJ; Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019), while affective empathy is often associated with the anterior insula (AI), the anterior cingulate cortex (ACC), and inferior frontal gyrus (IFG) (Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Fan et al., 2011; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019; Walter, 2012).

Most previous studies investigated empathic responses in a given moment, for example when observing pain in another person (Hein, Morishima, et al., 2016; Morelli, Lieberman, & Zaki, 2015; Singer & Lamm, 2009). Recent studies have shown that the dynamic formation of empathic responses in the realm of affective (Churamani, Barros, Strahl, & Wermter, 2018; Hein, Engelmann, et al., 2016; Olsson & Spring, 2018; Singer, Critchley, & Preuschoff, 2009) as well as cognitive empathy (Bagheri, Roesler, Cao, & Vanderborght, 2021) can be captured by reinforcement learning models. Reinforcement learning models mathematically describe the process of learning specific stimulus-outcome (i.e., reward vs. punishment) associations (e.g., Rescorla & Wagner, 1972), which can be extended to associations between persons and outcomes. The Rescorla-Wagner model assumes that the associative strength, e.g., between a person and an action, in a given trial can be described by the associative strength in the previous trial and a prediction error that quantifies the difference between the feedback that is actually observed in the present trial and the experience based on the previous trial. This prediction error is weighted by the learning rate which indicates how strongly the most recent experiences influence the change in associative strength. Originally, reinforcement learning models have been used to investigate various instances of reward and punishment learning (e.g., Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Klein, Ullsperger, & Jocham, 2017; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; Palminteri et al., 2015).

More recent works demonstrated that mechanisms underlying social preferences in general (FeldmanHall, Montez, Phelps, Davachi, & Murty, 2021; Lockwood & Klein-Flügge, 2021; Olsson, Knapska, & Lindström, 2020) and empathy-related behavior in particular (Hein, Engelmann, et al., 2016; Lockwood et al., 2016; Shamay-Tsoory & Hertz, 2022) may also be understood within the framework of reinforcement learning. Specifically, processes such as learning to react in an empathic fashion (Shamay-Tsoory & Hertz, 2022), obtaining rewards for another person (Lockwood et al., 2016), or empathizing with outgroup members (Hein, Engelmann, et al., 2016) can be captured by reinforcement learning models.

These studies have started to shed light on how empathy is formed. However, it remains unclear whether empathy-related closeness persists in the absence of empathy-inducing events. Answering this question is important to understand the longevity of empathy-induced effects such as social closeness.

Here, we conducted two studies to investigate the longevity of empathy-related social closeness and the underlying neural circuitries using an adapted reinforcement learning acquisition-extinction paradigm (Dunsmoor et al., 2018; Palminteri et al., 2015; Shiban et al., 2015), reinforcement learning modelling and functional magnetic resonance imaging (fMRI). In a third study, we tested whether the mechanisms underlying empathy-related closeness generalized to the formation and sustainability of another source, namely, the social norm of reciprocity.

To test the longevity of empathy-related social closeness, participants observed painful stimulation of another person, known to elicit empathy for pain (Beeney et al., 2011;

Grynberg & Konrath, 2020; Hein, Engelmann, et al., 2016; Lamm et al., 2007; Marsh, 2018) in a treatment condition and a control condition. In the first block of the treatment condition, participants observed painful stimulation of the other person (treatment partner) with high probability (80%), corresponding to the acquisition phase. In a second block, they observed the empathy partner receiving painful stimulation only with low probability (20%), corresponding to the extinction phase. In the control condition, participants observed painful stimulation in another person (control partner) at chance level in both blocks (50%; **Figure 2.1.1A**). In each trial, after observing the stimulation of the other person, participants rated their emotional reaction to the stimulation, and subsequently indicated how close they felt to the respective partner. To do so, they moved a mannequin (representing themselves) towards or away from a mannequin representing the other person (**Figure 2.1.1B**).

This set up allowed us to investigate the formation of empathy-related closeness in the acquisition phase, and the sustainability of empathy-related closeness in the extinction phase. Inspired by previous work demonstrating that watching other's receive painful stimulation elicits empathy (Beeney et al., 2011; Grynberg & Konrath, 2020; Hein, Engelmann, et al., 2016; Lamm et al., 2007; Marsh, 2018), we hypothesized that watching another person receiving painful stimulation constitutes feedback that is relevant to the process of learning empathy-related social closeness. Thus, the prediction error in our studies quantifies the difference between a hypothetical social closeness linked to observing the other person receive painful stimulation (i.e., the feedback), and the social closeness from before watching the other person receive pain (i.e., social closeness in the previous trial). In more detail, we hypothesized an increase in empathy-related social closeness in the acquisition phase. In the extinction phase, when the other person only rarely received empathy-inducing painful stimulation, we hypothesized a decay of empathy-related social closeness. However, if empathy-related closeness is sustainable, empathy-related social closeness should not decay in the extinction phase. On a neural level, learning-related changes and the extent to which empathy-related social closeness resists extinction should be associated with changes in activation in brain regions related to cognitive empathy such as the TPJ, the STS, the mPFC and the temporal poles, and to regions related to affective empathy such as the AI and the adjacent IFG, and the anterior and mid ACC.

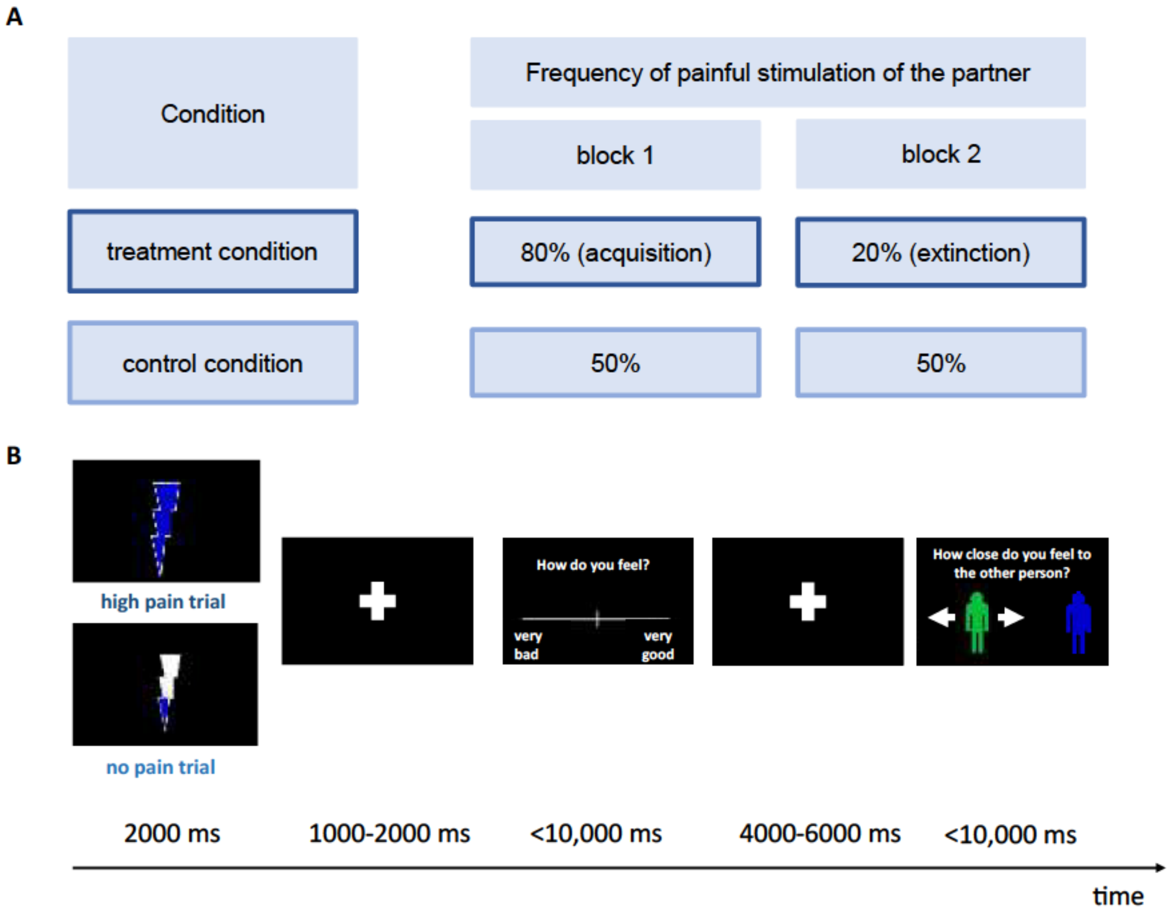


Figure 2.1.1 Visualization of the design and trial structure. **A** Participants sequentially underwent two counterbalanced conditions. In the treatment condition, they interacted with a first partner (confederate of the experimenter) and performed two blocks of the motive task. In block 1, empathy was reinforced in 80% of the trials and in block 2 in 20% of the trials. They performed the same tasks again with a new interaction partner (also confederate of the experimenter) in the control condition (order of treatment and control condition was counterbalanced across participants). Here, empathy was reinforced in 50% of the trials in both blocks. **B** At the beginning of each trial participants observed that the other person received a painful stimulation (high pain trial = reinforced trial) or a non-painful stimulation (no-pain trial = non-reinforced trial). Then participants rated how they felt after observing this feedback. After 4000-6000 ms, participants (green mannequin) indicated how close they felt to the other person.

Material and methods

Participants

We recruited 107 right-handed female participants via online platforms and flyers posted around the university campus in Würzburg. Participants were assigned to three different studies: two studies investigating the formation and sustainability of empathy-related social closeness (one fMRI and one behavioural replication study), and one behavioural control study investigating the formation and sustainability of social closeness driven by a social norm that is distinct from empathy, namely reciprocity. We trained two female students that served as confederates in all three studies.

We chose female participants as well as female confederates to control for gender and avoid cross-gender effects. The confederates were students who had been trained to act as naïve participants. We ensured that participants did not know either of the confederates prior to the experiment by asking confederates beforehand. Before the experiment began, written informed consent was obtained from all the participants. The study was approved by the local ethics committee (268/18). Participants received monetary compensation (26.80 ± 3.30 Euros (mean ± sd)).

We had to exclude seven data sets (five from the fMRI study and one each from the behavioural replication and the control study), because the estimation of learning models was not possible due to a lack of variance in ratings (four participants), sleepiness (two participants), or technical problems (one participant). We thus analysed 46 data sets for the fMRI study, 27 data sets for the behavioural replication study and 27 data sets for the control studies. The mean age was comparable between studies ($F(2, 106) = .987$, $P = .376$, see **Table S2.1.1** for overview of sample characteristics). A post-hoc sensitivity analysis using G*Power 3.1 indicated that given $\alpha = 5\%$ and considering 3 predictors in the regression model, the sample size in the fMRI study had 80% power to detect a true effect with an effect size of $f \geq .18$ ($F = 2.68$), and an effect size of $f \geq .23$ ($F = 2.73$) in the behavioral replication and the control study.

fMRI study and behavioural replication study

Procedure

Prior to the tasks, the individual thresholds for pain stimulation (see section pain stimulation for details) were determined for the participants and the confederates. Thus, participants had a first-hand experience of the pain stimulation they would observe in others.

Next, the participants and confederates were assigned their different roles in a manipulated lottery of drawing matches. Participants always drew the last match in order to ensure she was assigned her designated role (observer). The confederates were assigned the role of pain recipients and served as treatment or control partner counterbalanced across participants. In the fMRI study, the respective confederate (treatment partner in the treatment condition and control partner in the control condition) was seated on a chair to the left of the participant with her hand visible by the participant. In the behavioural replication study, the respective confederate was seated next to the participant in a soundproof cabin facing the opposite direction such that no one could see the other's screen.

The fMRI experiment consisted of the treatment condition, in which the participants observed painful stimulation of one of the confederates (treatment partner) with high probability (acquisition phase) or low probability (extinction phase), and the control condition in which participants observed painful stimulation of the other confederate (control partner) with chance probability in both blocks (**Figure 2.2.1**). Each block consisted of 25 trials. In between blocks, participants performed an additional task which was part of another experiment. In the acquisition phase, participants observed that the partner received ostensibly painful stimulation in 80% of the trials. In the extinction phase, they observed painful stimulation of the same confederate in 20% of the trials. In the control condition, participants observed painful stimulation of the second confederate in 50% of the trials of both blocks. Participants observed painful stimulation of different individuals in the treatment and the control condition to avoid spill-over effects and to keep the ostensible pain stimulation of the other person in a reasonable range. The order of treatment and control condition were counter-balanced across participants.

Participants spent approximately 60 minutes in the scanner and the entire procedure took about 2.5 hours. The behavioural replication study lasted approximately 2 hours. To avoid possible reputation effects (e.g., Engelmann & Fischbacher, 2009; Gächter & Falk, 2002), which could influence participants' behavior, participants were informed at the beginning that they would not meet the others after the experiment. In more detail, at the end of the fMRI study, the second confederate left and the participant remained in the scanner for anatomical image acquisition. At the end of the behavioural replication study, the confederate left and participants remained in the cabin to complete the same questionnaires as in the fMRI study, (outlined in detail below).

Task

Each trial started with a fixation cross displayed for 4000-6000 ms, followed by a continuous slider scale (internally ranging from 0-100) that asked the participant to indicate how close they felt to the other person at this moment ("How close do you feel to the other person?" in German). Participants were asked to respond within 10 seconds (6 seconds in the laboratory study). After a second fixation cross (1000-2000 ms), participants were either shown a fully filled flash in the partner's color (symbolizing a painful stimulation of the partner, i.e., a reinforced trial) or a partly filled flash in the partner's color (symbolizing a non-painful stimulation of the partner, i.e., a non-reinforced trial) for 2000 ms. The respective flash was followed by a fixation cross (1000-2000 ms). At the end of each trial, participants indicated how they felt ("How do you feel?" in German) after having observed the partner's stimulation on a visually displayed continuous slider scale (internally ranging from 0-100), and again had to respond within 10 seconds (6 seconds in the laboratory study).

Behavioral control study

Procedure

The procedure was identical to the behavioural replication study, except that now the participants were assigned as pain recipients and the confederates could decide to give up money to save them from pain, a procedure that has been shown to induce positive reciprocity (Hein, Morishima, et al., 2016; Saulin et al., 2022).

Each block of the reciprocity learning task consisted of 25 trials. In the treatment condition (corresponding to two interaction blocks with one confederate), participants observed that the partner ostensibly decided to help them in 80% of the trials in block 1 and in 20% of the trials in block 2. In the control condition (corresponding to the two interaction blocks with the other confederate), participants observed that the partner ostensibly helped them in 50% of the trials in block 1 as well as block 2. Again, the order of treatment and control condition were counter-balanced across participants. To avoid possible reputation effects (e.g., Engelmann & Fischbacher, 2009; Gächter & Falk, 2002), which could influence participants' behavior, participants were informed at the beginning that they would not meet the ostensible other participants after the experiment. At the end of study, the confederate left and participants remained in the cabin to complete the same questionnaires as in the other two studies.

Task

The trial structure was comparable to the fMRI study and the behavioural replication study described above. Each trial started with the display of a jittered fixation cross (4000-6000 ms). Then participants were asked to indicate how close they felt to the other person at that moment ("How close do you feel to the other person?" in German) on a continuous slider scale (internally ranging from 0-100) and were asked to respond within 6 seconds. After a fixation cross (1000-2000 ms), participants saw the deliberation screen of the interaction partner, in which the two possible options were visualized side-by-side using a fully filled flash in the color of the participant (symbolizing the option to take the monetary reward and not help the participant) and a crossed out fully filled flash in the color of the participant (symbolizing the option to forego the monetary reward and help). This screen was shown for a jittered length of 2000-4000 ms, followed by the display of the ostensible decision of the interaction partner. If the decision was to help (reinforced trial), the crossed-out flash was highlighted by a box in the color of the interaction partner. If the decision was not to help (non-reinforced trial), the fully filled flash was shown highlighted by a box in the color of the interaction partner. After another fixation cross (1000-2000 ms), the emotion rating scale was shown asking the participant how they felt after observing the partner's decision ("How do you feel?" in German). Again, participants were asked to respond within 6 seconds. Then, the next trial started.

Questionnaires

At the end of the respective main experiments, participants filled out questionnaires capturing trait empathic concern and perspective taking/cognitive empathy (empathic concern and perspective taking subscales of the Interpersonal Reactivity Index (IRI, Davis (1980)). Conceptually, scores on the empathic concern subscale have been related to emotional empathy, and scores on the perspective taking subscale to cognitive empathy (Davis, 1980, 1983). Moreover, they completed questionnaires measuring individual differences in trait reciprocity (Personal Norm of Reciprocity, PNR; Perugini, Gallucci, Presaghi, & Ercolani, 2003) as well as participants' impressions of the other individuals (confederates) (Hein, Engelmann, et al., 2016; Hein, Silani, Preuschoff, Batson, & Singer, 2010) modified from Batson et al. (1988).

Pain stimulation

In the fMRI study, painful stimulation was applied using a Digitimer DS7A constant current stimulator (Hertfordshire, United Kingdom) and an MRI compatible surface electrode attached to the left lower inner arm. Shock segments consisted of a single 1 ms square-wave pulses. For pain stimulation in the laboratory, we used a mechano-tactile stimulus generated by a small plastic cylinder (612 g). The projectile was shot against the cuticle of the left index finger using air pressure (Impact Stimulator, Labortechnik Franken, Release 1.0.0.34).

In all studies, the criterion for painful stimulation was a subjective value of 8 on a pain scale ranging from 1 (no pain at all, but a participant could feel a slight tingling) to 10 (extreme, hardly bearable pain). The participants were told that a value of 8 corresponded to a painful, but bearable stimulus, and a non-painful stimulus corresponded to a value of 1 on the same subjective pain scale. These subjective pain thresholds were determined using a stepwise increase in shock strength (air pressure in the laboratory) starting with the lowest value of 0.00 mA (0.25 mg/s in laboratory) in steps of 0.05 mA (0.25 mg/s in laboratory) until it reached the individual's value of 8 (range fMRI= 0.25-1.50 mA; range behavioral replication study and control study= 2.00–6.00 mg/s). Hence in all studies participants experienced the same threshold procedure and a painful stimulation corresponding to strong but bearable pain.

Regression analyses

In all linear mixed effects regression models, we conducted, we included participant as random intercept in order to account for shared error variance across multiple data points, i.e., the within-subjects variables. Random slopes were included for continuous variables if these variables were also included as a fixed effect. As our categorical variables only yielded two levels, we did not include random slopes for categorical variables.

As manipulation check, we first checked whether emotion ratings significantly differed for observed pain vs. no-pain. To test this, we ran linear mixed models analyses with the fixed effects of trial type (reinforced vs. non-reinforced), block (block 1 vs. block 2) and condition (treatment vs. control), participant as random intercept and the dependent variable emotion rating.

In order to test whether we successfully reinforced empathy, we conducted a linear mixed models analysis with trait empathy (fMRI study and behavioral replication study: empathic concern subscale of the IRI (Davis, 2006)), trial type, study, and their interaction as fixed effects, participant and trial number as random intercept, and emotion ratings as dependent variable. In the behavioral control study, the analogous analysis was conducted but using positive reciprocity as trait measure of reciprocity (positive reciprocity) subscale of the PNR (Perugini et al., 2003).

In order to test the influence of condition, block, and trial number on social closeness, we conducted linear mixed models with condition, block (block 1 vs. block 2), and trial number (1-25) as fixed effects, participant as random intercept, trial number as random intercept for participant and trial-by-trial closeness ratings as dependent variable. In order to test whether the resulting effects were comparable across the fMRI and the behavioral replication study, we reran this analysis using the pooled data of these two studies and adding a predictor variable for Study.

Linear mixed model analyses were conducted in R (R version 4.0.4, R Core-Team, 2018) using the packages *lme4* (Bates et al., 2014) and *car* (Fox et al., 2018). For mixed models, we report the chi-square values derived from Wald chisquare tests using type 3 sum of squares from the Anova() function (*car* package). For predefined contrasts we report the t-values derived from the summary() function. Simple slopes extracted from the linear mixed models are reported with 95% confidence intervals using the *emtrends* function (*emmeans* package; Lenth, Singman, Love, Buerkner, & Herve, 2019).

Computational modelling

To identify the computational mechanisms of the formation and maintenance of empathy-related social closeness, we tested three different learning models against each other (**Figure 2.1.2**). Specifically, our baseline model, which implemented only the standard Rescorla-Wagner learning rule (model 1, **Figure 2.1.2A**), was compared to two recent adaptations (models 2 and 3) that allowed us to test the role of specific processes, namely differential learning rates for positive and negative feedback (e.g., Garrett & Daw, 2020) and context-dependent recalibration of the prediction error (e.g., Bavard, Lebreton, Khamassi, Coricelli, & Palminteri, 2018). The first adaptation (model 2, **Figure 2.1.2B**) assumes different learning rates for positive prediction errors and negative prediction errors, i.e., for the learning and the unlearning of an association. If, for example recent experiences more strongly influence surprisingly positive than surprisingly negative feedback, the learning rate for positive prediction errors will be larger than the learning rate for negative prediction errors. In the context of empathy-related and reciprocity-based social closeness such a finding would entail that social closeness more rapidly increases in the acquisition phase than it decreases in the extinction phase.

In the second adaptation (model 3, **Figure 2.1.2C**), we hypothesized that the assumed outcome values of the respective feedback (i.e., R = 1 for reinforcer feedback and R = 0 for non-reinforcer feedback) may vary depending on the respective context (e.g., empathy motive vs. reciprocity motive). Thus, the prediction error is directly recomputed which means that the learning signal itself is recalibrated. The larger this recalibration, the smaller the learning signal associated with a reinforced trial and the larger the learning signal associated with a non-reinforced trial, and vice versa. Context-dependent recalibration therefore allows social closeness to continue to increase in the extinction block despite a high probability for non-reinforced trials.

Based on these models, we aimed to test whether empathy sustainability can be understood (i) in terms of asymmetrical updating of the learning signal (i.e., different learning rates for reinforced and non-reinforced trials) or (ii) in terms of recalibration of the value associated with the feedback in each trial (i.e., a value different from 1 in reinforced trials and different from 0 in non-reinforced trials). We hence tested which out of three models in our model space best describes participants' behavior.

In the simplest model (*basic model*), the estimated motive-driven closeness V at trial t is updated with prediction error δ and free parameter α only. Specifically, the prediction error is calculated as difference between the actual outcome and the prediction:

$$\delta_t = R_t - V_{t-1} \tag{1}$$

In equation (1), R_t refers to the actual outcome: 1 for reinforced feedback (painful stimulation of the partner in the fMRI and behavioral replication study, decision of the partner to help in the behavioral control study) and 0 for non-reinforced feedback (non-painful stimulation of the partner in the empathy group, decision of the partner not to help in the reciprocity group) at trial t.

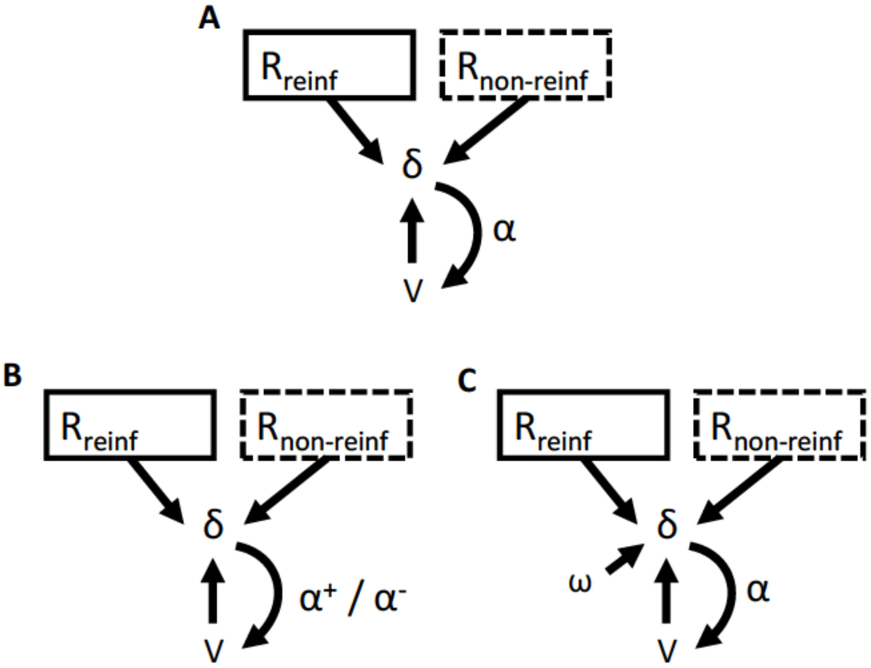


Figure 2.1.2 Model space. **A** Basic model. In the basic model, social closeness in the next trial V_t depends on the closeness rating in the present trial V_t and the learning rate α multiplied by the prediction error δ. This prediction error is computed as the difference between the reinforcer value in the current trial R (1 vs. 0) and the closeness rating of the previous trial V_{t-1}. **B** Differential model. Same as the basic model with the addition that alpha is different for reinforced (α^+) and non-reinforced (α^-) trials. **C** Individual calibration model. Same as the basic model except that a recalibration parameter ω is added to the computation of the prediction error δ. That is $\delta = (R-\omega)-V_{t-1}$ if R = 1 (reinforced trials) and $\delta = \omega-V_{t-1}$ if R = 0 (non-reinforced trials).

The prediction error is weighted by the learning rate and used to update V:

$$V_t = V_{t-1} + a \times \delta_t \quad (2)$$

In the second model (*differential model*), we tested if positive and negative prediction errors are updated separately, inspired by previous studies on reward learning (cf. e.g., Garrett & Daw, 2020). In this model, the prediction error was calculated as in equation (1), but learning rates depended on whether the present trial was reinforced or not (see equation 3). That is, a positive δ will be multiplied by learning rate α^+, and a negative δ will be multiplied by learning rate α^- to update V.

$$V_{t+1} = \begin{cases} V_t + \alpha^{+} \times \delta_t \, if \; \delta > 0 \\ V_t + \alpha^{-} \times \delta_t \, if \; \delta \; < 0 \end{cases} \tag{3}$$

Hence, the learning of empathy-related closeness may be characterized by a stronger weight of the prediction error for reinforced compared to non-reinforced trials, thus leading to less decline in empathy-related closeness when reinforcer rates are low (as in the second block of the treatment condition in the fMRI and the behavioral replication studies).

Third, based on previous work (Palminteri et al., 2015), the assumed outcome values of the respective feedback (i.e., R = 1 for reinforcer feedback and R = 0 for non-reinforcer feedback) may actually be recalibrated depending on the respective context (empathy motive vs. reciprocity motive). To test whether the learning of motive-driven closeness can be understood in these terms, we added a third model (*individual calibration model*), in which the proposed outcome value is recalibrated by subtracting an additional free parameter ω (see equation 4).

$$\mathrm{\delta_t = |R_t - \omega| - V_{t-1}} \tag{4}$$

Hence, according to this model, an individual's actual outcome value for reinforced trials corresponds to 1 minus the individual recalibration value ω, and the actual outcome value of a non-reinforced trial corresponds to ω. Thus, the larger the value of ω, the more likely a positive prediction error and subsequent increase of social closeness after non-reinforced trials. Hence, the larger the value of ω, the less decline of empathy-driven closeness can be expected for the extinction phase (i.e., when non-reinforced trials are most frequent).

Model optimization and comparison

The parameters θ_M in each model *M* were optimized using the procedure of minimizing the negative logarithm of the posterior probability (nLPP): the combination of the likelihood for choosing a particular closeness value and the prior distribution of the parameters.

$$nLPP = \; -log\;(P(\theta\_M \,|\, D, M)) \propto \; -log\;(P(D \,|\, M, \theta\_M\;)) - log\;(P(\theta\_M \,|\, M)) \tag{5}$$

$P(D \,|\, M, \theta\_M\;)$ refers to the likelihood of choice value D (i.e., the actual rating) given the current model M and its parameters θ_M. Here, we assumed that the rating was selected from the normal distribution with the estimated rating as mean (given M and θ_M) and standard deviation of 0.4. Therefore, if the rating is correctly estimated and close to the actual rating D, the likelihood will be high. It is worth to note that this method deviates from the typical approach to estimate Q-learning models, in which the probability of a binomial decision is estimated with temperature parameter β. The temperature parameter β explains whether a decision is made based on the differences between two options, however, this is not appropriate in the context of our task that includes only one choice option on a continuous scale.

$P(\theta\_M \middle| M)$ is the likelihood of getting an estimate for θ_M within the prior probability distribution of the parameters. All parameters were selected from a beta distribution (α = β = 1.1) (Daw et al., 2011), so that the estimated value will always be located between 0 and 1. We then applied the model to fit the data.

A lower LPP value indicates that a model can explain the data better, however, the n*LPP* does not take a model's complexity into consideration. To address this issue, we then applied the Laplace approximation to the model evidence (*LAME*) to penalize goodness-of-fit (i.e., the measure of n*LPP* for each subject) with model complexity (i.e., number of parameters). The LAME for each model was computed according to equation 6.

$$LAME \equiv -LPP + df/2 \ \ log\ (2\pi) - 1/2\ log|H| \tag{6}$$

In this calculation, df is determined as the number of free parameters and $|H|$ is the determination of the Hessian. Again, these values were computed at individual level.

To test which model out of the model space is most likely to have generated a certain data set, we fed the LAME (from each subject in each model) to group-level random-effects analysis in the mbb-vb-toolbox (http://mbb-team.github.io/VBA-toolbox/; Daunizeau, Adam, & Rigoux, 2014). This toolbox performs Bayesian model selection and estimates two indicators of model performance: the exceedance probability (EP) and the expected model frequencies (*EF*) for each model. Specifically, the exceedance probability of a model quantifies the probability for a given model to have generated the data relative to the other models in the model space. Commonly, an EP of higher than 95% is an indicator of convincing evidence for a model to be most likely to have generated the data compared to other models. The expected frequency *EF* of a model quantifies the probability that the model generated the data for any randomly selected subject. Note that the EF should be higher than chance level given the number of models in the model space (in our case higher than 1/3).

The modelling was conducted using Matlab 2018b. The estimated rating (V) was initialized as the actual rating in the first trial in each block. All the parameters were optimized using Matlab's *fmincon* function with random starting points, ranging from 0 to 1.

fMRI data acquisition

Imaging data was collected at a 3T MRI-scanner (Skyra syngo, Siemens, Erlangen, Germany) with a 32-channel head coil. Functional imaging was performed with a multiband EPI sequence of 42 transversal slices oriented along the subjects' AC-PC plane and distance factor of 50% (multi-band acceleration factor of 2). The in plane resolution was 2 x 2 mm^2 and the slice thickness was 2 mm. The field of view was 216 x 216 mm^2, corresponding to an acquisition matrix of 108 x 108. The repetition time was 1340 ms, the echo time was 25 ms, and the flip angle was 60°. Structural imaging was conducted using a sagittal T1-weighted 3D MPRAGE with 240 slices, and a spatial resolution of 1 x 1 x 1 mm^3. The field of view was 256 x 256 mm^2, corresponding to an acquisition matrix of 256 x 256. The repetition time was 2,300 ms, the echo time was 2.96 ms, the total acquisition time was 3:50 min, and the flip angle was 9°. We obtained, on average, 1,215 (SE = 5.07 volumes) EPI-volumes in the control condition and 1,208 (SE = 4.26 volumes) EPI columes in the treatment condition for each participant. We used a rubber foam head restraint to avoid head movements.

fMRI Preprocessing

Preprocessing and statistical parametric mapping were performed with SPM12 (Wellcome Department of Neuroscience, London, UK) and Matlab version 9.2 (MathWorks Inc; Natick, MA). Spatial preprocessing included realignment to the first scan, and unwarping and coregistration to the T1 anatomical volume images. Unwarping of geometrically distorted EPIs was performed using the FieldMap Toolbox. T1-weighted images were segmented to localize grey and white matter, and cerebro-spinal fluid. This segmentation was the basis for the creation of a DARTEL Template and spatial normalization to Montreal Neurological Institute (MNI) space, including smoothing with a 6 mm (full width at half maximum) Gaussian Kernel filter to improve the signal-to-noise-ratio. To correct for low-frequency components, a high-pass filter with a cut-off of 128 s was used.

fMRI statistical analysis

First-level analyses

First-level analyses were performed with two general linear models (GLMs), using a canonical hemodynamic response function (HRF). Regressor lengths were defined from stimulus onset until the individual response was made by pressing a button (resulting in a time window of 1,000 ms + individual response time) for stimuli that required a response (emotion rating phase, closeness rating phase) and from stimulus onset to stimulus offset for stimuli that were just observed by participants (feedback phase, i.e., observing the partner's pain vs. no pain). The model included three regressors of interest the closeness phase (scale onset until button press), the feedback phase (stimulus onset until stimulus offset), and the emotion rating phase (scale onset until button press). Parametric modulators coded the trial type (PM trial type), i.e., whether the current trial was reinforced (value = 1) or non-reinforced (value = 0), separately for the closeness phase, the feedback phase, and the emotion rating phase. An additional task of no interest was modelled as additional regressor. The residual effects of head motions were corrected by including the six estimated motion parameters for each participant and each session as regressors of no interest. To allow for modelling all the conditions in one GLM, an additional regressor of no interest was included, which modelled the potential effects of session.

Second-level analyses

Based on the first-level model, we performed one-sample t-tests on the respective parametric modulator separately for each phase of interest (feedback, emotion-rating, closeness) across all blocks and conditions. In a next step, we computed second-level regressions with the same simple contrasts and individual ω values as covariate across all blocks and conditions separately for each phase. Next, we re-ran these second-level regressions using the difference in neural activation between conditions, i.e., PM trial type (treatment) > PM trial type (control) and individual ω values as covariate. The main manuscript focusses on the result from the emotion rating phase as this is the phase clearly linked to empathic reaction. The results for the other task phases are reported in the supplement (see supplementary results). As recommended, a cluster-forming threshold of P <.001 uncorrected (Eklund, Nichols, & Knutsson, 2016; Woo, Krishnan, & Wager, 2014;

Yeung, 2018) was used and where not stated otherwise whole-brain level FWE cluster-corrected statistics are reported at an α level < .05.

To test the relationship of neural activation related to individual recalibration with closeness ratings, emotion ratings and trait empathy, beta values during acquisition and extinction in the emotion rating phase were extracted from the resulting bilateral clusters in temporo-parietal junction/superior temporal sulcus and left inferior frontal gyrus/anterior insula using MarsBar (Matthew Brett, Anton, Valabregue, & Poline, 2002). Extracted beta values were added as predictors in two separate linear mixed models together with block (acquisition vs. extinction) and empathy subscale (empathic concern vs. perspective-taking subscale of the IRI (Davis, 2006)), trait score, and their interaction as fixed effects, participant and trial number as random intercepts, and social closeness as dependent variable.

Results

Results of the fMRI and the behavioural replication study

Manipulation Check

The analysis of emotion ratings in the fMRI study showed a main effect of trial type (pain vs. no-pain) (χ^2= 59.44, P < .001, β = .85, SE = .110). This effect was not modulated by condition (trial type X condition interaction: χ^2 = 1.95, P = .16, β = .22, SE = .16) or block (trial type X block interaction: χ^2 = .04, P = .84, β = .03, SE = .16), indicating that participants emotionally distinguished between those trials in which the partner received painful stimulation vs. non-painful stimulation and did so equally strongly in the treatment and control conditions and across the two blocks.

This effect was replicated for the laboratory replication study (χ^2= 52.27, P < .001, β = 1.22, SE = .168; trial type X condition interaction: χ^2 = .78, P = 377, β = .21, SE = .24; trial type X block interaction: χ^2 = .34, P = .56, β = .14, SE = .24) and was comparable across studies (fMRI vs behavioural replication study: trial type X condition X block X study interaction: χ^2= 1.82, P = .18, β = .19, SE = .14).

To test whether the emotion ratings were associated with trait empathy, we conducted a linear regression with the empathic concern subscale of the IRI (Davis, 2006) as predictor and study (fMRI vs. replication study) and trial type (observed pain vs. observed no-pain) as control variables. Results showed that trait empathic concern (EC) was generally associated with larger emotional reactions to the others stimulation on a marginal level (main effect of trait empathic concern: χ^2= 3.31, P = .07, β = -.31, SE = .17). This relationship, however, was stronger for observed pain in contrast to observed no-pain (trial type X trait EC interaction: χ^2= 8.07, P = .005, β = .53, SE = .19). Hence, for trials of observed pain, trait EC was more predictive of the emotional reaction than in trials of observed no-pain. This effect was more pronounced in the replication study than in the fMRI study (trial type X trait EC X study interaction: χ^2= 4.52, P = .03, β = -.49, SE = .23). Conducting the analogous models with trait perspective-taking (PT) showed no main effect of PT on emotional reactions but revealed a significant interaction of PT and trial type (trial type X trait PT interaction: χ^2= 12.10, P < .001, β = -.31, SE = .18). That is, emotional reactions to observed non-pain were relatively more positively linked to trait PT than emotional reactions to observed pain. Again, this effect was more pronounced in the replication study than in the fMRI study (trial type X trait PT X study interaction: χ^2= 4.58, P = .03, β = -.39, SE = .17)

Behavioral Results: Empathy motive activation leads to sustained social closeness

The main goal of the current studies was to understand how social closeness developed over time in the two blocks and conditions. To this end, a linear mixed model was conducted with trial number (1 to 25), block (block 1 vs block 2), and condition (control vs. treatment) as fixed effects and participant as random intercepts and trial number as random slope for participant. This analysis revealed that empathy-related closeness increased with trial number in all blocks and conditions (main effect of trial number: χ2 = 15.62, P <.001, β = .10, SE = .02, see **Table 2.1.1** for full results and **Figure 2.1.3A** for visualization). Average closeness was larger in block 2 than block 1 (main effect of block: χ2 = 47.41, P < .001, β = .14, SE = .02) and larger in the treatment than in the control condition (main effect of condition: χ2 = 18.50, P < .001, β = .09, SE = .02). Further, results showed a significant interaction between condition and block (χ2 = 26.87, P < .001, β = -.15, SE = .03), reflecting that in block 1, ratings in the treatment condition tended to be comparable to ratings in the control condition, whereas they were higher in the control condition in block 2. In contrast to a hypothesized decay in social closeness in block 2, post-hoc t-tests comparing the means of the last five trials in block 1 and the mean of the last five trials in block 2, revealed no significant difference in closeness (T(45) = -.96, P = .34), indicating sustained empathy towards another who is only rarely receiving painful stimulation. The corresponding analysis in the behavioral replication study replicated these results with the addition of a stronger effect of trial number in block 1 than block 2 (trial number × block interaction χ2 = 4.28, P = .039, β = -.06, SE = .03, see **Table 2.1.1** for full results and **Figure 2.1.3C** for visualization). Again, post-hoc t-tests comparing the means of the last five trials in block 1 and the mean of the last five trials in block 2, revealed no significant difference in social closeness (T(26) = 1.29, P = .208).

Combined analysis of both studies showed a larger main effect of block in the behavioral replication study and more pronounced interaction between condition and block number (block × study: χ2 = 4.67, P = .031, β = -.07, SE = .03; block × condition × study: χ2 = 7.66, P = .006, β = .13, SE = .05; **Figures 3A** and **3C**).

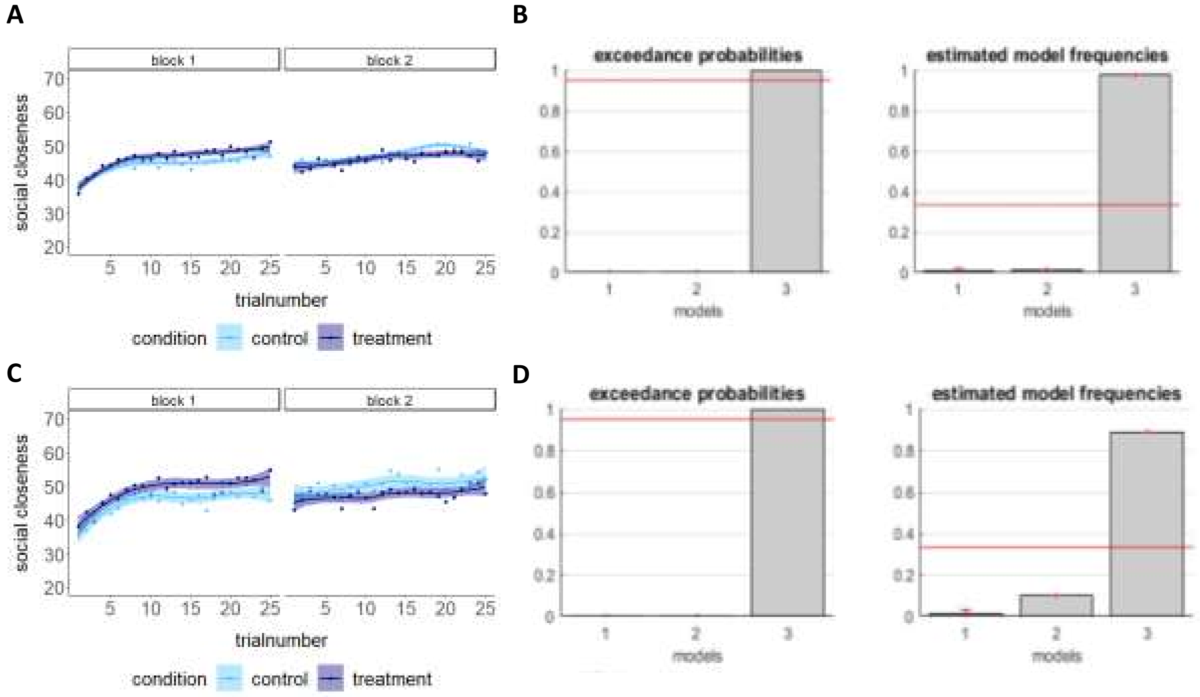


Figure 2.1.3 Mean empathy-related social closeness and results of Bayesian model comparison in the fMRI study (top) and the behavioral replication study (bottom). **A** Mean social closeness in the fMRI study with model free trend line and pointwise 95% confidence interval (loess function) by block, condition, and trial number. Social closeness increased in block 1 and plateaus/slightly increased in block 2 in both conditions, demonstrating sustainability of empathy-related social closeness. **B** Bayesian model comparison of three models (see **Figure 2.1.2** for model space) revealed that individual recalibration of the learning signal associated with observing another's pain vs. no-pain was most likely to explain participants' social closeness rating behavior. **C** Replication of the behavioral pattern and **D** of the modelling comparison results in the laboratory replication study.

Table 2.1.1 Results of the linear mixed models analysis with condition (treatment vs. control), trial number (1-25), block (block 1 vs. block 2) as fixed effects, participant as random intercept and trialnumber as random slope for participant. The dependent variable are participants' closeness ratings in the fMRI study (N = 46, 4600 observations, maximal VIF = 3.12) and the behavioral replication study (N = 27, 2700 observations, maximal VIF = 3.10). χ2 and P(χ2) are the type 3 Wald χ2 test statistics. VIF = variance inflation factor.

Factor	**beta**	**SE**	**t-value**	**χ2**	**P(χ2)**
fMRI study					
(Intercept)	-.076	.129	-.60	.35	.55
Condition	**.087**	**.021**	**4.15**	**18.50**	**<.001**
trial number	**.096**	**.015**	**6.47**	**15.50**	**<.001**
Block	**.140**	**.021**	**6.64**	**47.41**	**<.001**
condition*trial number	.028	.021	1.31	1.85	.174
condition*block	**-.149**	**.030**	**-5.00**	**26.87**	**<.001**
trial number*block	-.013	.021	-.62	.42	.519
condition*trial number*block	-.047	.030	-1.57	2.67	.101
Behavioral replication study					
(Intercept)	-.112	.165	-.68	.46	.496
Condition	**.152**	**.029**	**5.19**	**26.27**	**<.001**
trial number	**.117**	**.021**	**5.65**	**10.78**	**.001**
Block	**.214**	**.029**	**7.30**	**57.98**	**<.001**
condition*trial number	.022	.029	.76	.62	.430
condition*block	**-.283**	**.041**	**-6.83**	**50.68**	**<.001**
trial number*block	**-.058**	**.029**	**-1.99**	**4.29**	**.039**
condition*trial number*block	-.035	.041	-.843	.77	.399

Computational modelling of empathy-related social closeness

In a next step, we tested which of the three variants of the Rescorla-Wagner model best described the development of empathy-related social closeness (see **Figure 2.1.2** for visualization of the model space). As outlined in detail above, the first model (*basic model*) consisted of the basic Rescorla-Wagner model with one learning rate; the second model (*differential model*) allowed for a different learning rate in reinforced trials and non-reinforced trials; the third model included a recalibration parameter ω that directed the computation of the prediction error (*individual calibration model*).

Bayesian model comparison (see methods for details) revealed that in the fMRI study (**Figure 2.1.3B**), the *individual calibration model* is the winning model with an exceedance probability of over 99 % (probability that this model is more likely than all other models in the model space) and an estimated model frequency of 97 % (probability that this model generated the data of any randomly selected participant). This result was replicated in the behavioral replication study (**Figure 2.1.3D**).

The recalibration parameter ω

For empathy-related social closeness, the respective winning model included the recalibration parameter ω. The larger this parameter, the more likely are non-reinforced trials to elicit a positive prediction error and hence a positive updating of closeness. A large ω should thus entail less decay of social closeness in the extinction phase than a small ω.

The recalibration parameter ω was initially estimated across all blocks and conditions as one variable characterizing each individual. To test, whether strong recalibration was specific to the extinction block, we tested additional RL models in which ω was free to vary by block as well as condition, resulting in block-specific estimates of individual recalibration for both conditions (see supplementary online results for details). These analyses showed that on average, participants more strongly recalibrated in the extinction block than in the acquisition block (fMRI study: T(45) = 2.753, P = .009, CI = [.345, .054]); replication study (T(26) = 2.0, P = .056, CI = [-.005, .384]), but recalibration values did not differ between block 1 and block 2 for the control condition(fMRI study: T(45) = -.579, P = .568, CI = [-.139, .077]; replication study: T(26) = -1.027, P = .314, CI = [-.176, .059]; for visualization of the median and spread of the extracted parameters, see supplementary **Figure S2.1.6**).

Behavioral control study: Reciprocity does not induce sustained social closeness

So far, our results revealed the sustained nature of empathy-related closeness, because of the recalibration of the outcome value, associated negative emotion ratings (empathy for pain) in the acquisition phase and positive emotion ratings (empathic joy) in the extinction phase. To test if the observed recalibration of social closeness is a general phenomenon or specifically related to empathy, we conducted a behavioural control study using the identical experimental design to test the formation and sustainability of reciprocity-based closeness. Reciprocity, commonly defined as returning a previously given or an anticipated favor (Gouldner, 1960; Hein, Morishima, et al., 2016; McCabe et al., 2003), is one of the most important social norms worldwide (Axelrod & Hamilton, 1981; Falk & Fischbacher, 2006; Nowak, 2006; Perugini et al., 2003). Similarly to empathy, reciprocity can increase closeness (Adams & Miller, 2022; Neyer, Wrzus, Wagner, & Lang, 2011), and is a strong motivator of prosocial behaviour (Fehr et al., 2002). However, whereas empathy-related closeness and prosocialty is elicited by sharing the emotions of the other, reciprocity-based processes are conditional on the other's behaviour, i.e., reflect a "tit-for-tat" principle rather than shared emotions (Dufwenberg & Kirchsteiger, 2004; Eccles, Hughes, Kramár, Wheelwright, & Leibo, 2020; Rand, Ohtsuki, & Nowak, 2009; Zaki, 2014). Hence, to reinforce reciprocity in the present paradigm, the participant received help from the other person, i.e., the other person gave up a monetary reward to save the participant from pain, a procedure that has been-established for enforcing direct positive reciprocity towards the helper (Hein et al., 2010; Saulin et al., 2022). The trial structure was identical to the trial structure in the two empathy studies outlined above (for visualization of an exemplary trial, see supplementary **Figure S2.2.2B**).

Manipulation Check

Analogously to the empathy studies above, we first analysed participants' emotion ratings. Results of a linear mixed model revealed a main effect of trial type (χ^2= 62.89, P < .001, β = -1.06, SE = .134) independent of condition (main effect of condition: χ^2 = .15, P = .701, β = -.05, SE = .134, trial type X condition interaction: χ^2 = .004, P = .953, β = -.01, SE = .190) for the reciprocity motive. Thus, participants emotionally distinguished between those trials in which the partner had decided to help them vs. decided not to help them. They did so equally strongly in both conditions (treatment vs. control). We further tested whether trait positive reciprocity was associated with participants' emotion ratings and conducted a linear mixed models analysis with positive trait reciprocity scores (positive reciprocity subscale of the PNR questionnaire (Perugini et al., 2003) and trial type (reinforced vs. non-reinforced) and their interaction as fixed effects, participant as random intercept, and emotion ratings as dependent variable. This analysis showed that positive trait reciprocity was marginally linked to emotional reactions (main effect of trait positive reciprocity: χ^2 = 3.30, P = .07, β = .19, SE = .11). However, when controlling for block number (block 1 vs. block 2) and condition (control vs. treatment) this effect was significant (main effect of trait positive reciprocity: χ^2 = 4.34, P = .037, β = .26, SE = .12), demonstrating that our paradigm successfully reinforced positive reciprocity.

Behavioral results

Next, we conducted a linear mixed model with trial number, block, and condition as fixed effects, participant as random intercept and trial number as random slope for participant to analyse the development of reciprocity-related social closeness over time. This analysis revealed a significant three-way interaction of condition, trial number, and block (χ2 = 120.69, P < .001, β = -.53, SE = 05), which shows that the development of social closeness over time differentially depended on the block as well as the condition (see **Figure 2.1.4A** for visualization and **Table 2.1.2** for full results). Thus, in contrast to empathy-related social closeness, reciprocity-related social closeness was affected significantly by reinforcement frequency: in the treatment condition (**Figure 2.1.4A**, dark lines) social closeness increased when strongly reinforced during the acquisition block and decaying when weakly reinforced during the extinction block, while in the control condition (**Figure 2.1.4A**, light lines) where reinforcement remained at chance level in blocks 1 and block 2 little change in social closeness ratings was observed.

Computational modelling of reciprocity-related social closeness

Bayesian model comparison conducted analogously to the fMRI and the behavioral replication study revealed that in the control study, the *basic model* is quite likely to have generated the data as well as the *individual calibration model*. (**Figure 2.1.4B,** see supplementary **Table S2.1.2** for overview of model comparison metrics and **Figure S2.1.3C** for visualization of absolute model fit). Hence, in contrast to empathy-related social closeness formation and sustainability, reciprocity-related social closeness can be well captured by a simple learning rule, which is in line with the decrease in social closeness when the frequency of helping declined during block 2 of the treatment condition.

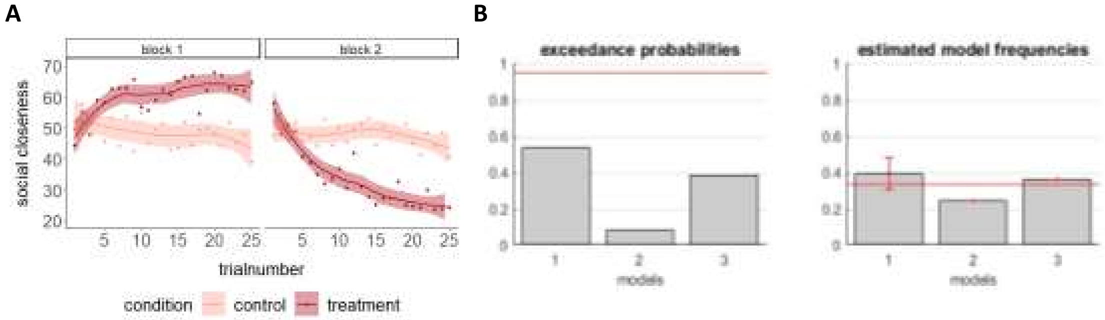


Figure 2.1.4 Behavioral pattern and Bayesian model comparison results of the behavioral control study. **A** Mean social closeness with model free trend line and pointwise 95% confidence interval (loess function) by block, condition, and trial number. Social closeness increased in block 1 of the treatment condition (acquisition phase) and starkly decreased in block 2 (extinction phase), demonstrating no sustainability of reciprocity-related social closeness. **B** Bayesian model comparison of three models (see **Figure 2.1.2** for model space) revealed that the basic model assuming simple updating directly based on the learning signal and individual recalibration of the learning signal associated with observing another's help vs. no help are equally likely to explain participants' reciprocity-related social closeness rating behavior.

Table 2.1.2 Results of the linear mixed models analysis with condition (treatment vs. control), trial number (1-25), block (block 1 vs. block 2) as fixed effects, participant as random intercept and trial number as random slope for participant. The dependent variable was participants' reciprocity-related closeness ratings in the behavioral control study (N = 27, 2700 observations, maximal VIF = 3.40). χ2 and P(χ2) are the type 3 Wald χ2 test statistics.

Factor	**beta**	**SE**	**t-value**	**χ2**	**P(χ2)**
(Intercept)	.034	.126	.27	.07	.789
Condition	**.53**	**.034**	**15.28**	**236.85**	**<.001**
trial number	**-.082**	**.024**	**-3.35**	**8.25**	**.004**
Block	-.029	.034	-.83	.70	.403
condition*trial number	**.229**	**.035**	**6.66**	**45.07**	**<.001**
condition*block	**-1.13**	**.049**	**-23.24**	**546.42**	**<.001**
trial number*block	.030	.034	.87	.77	.381
condition*trial number*block	**-.530**	**.049**	**-10.91**	**120.70**	**<.001**

Imaging results

Whole-brain results

The behavioural results revealed that empathy-related social closeness, in contrast to reciprocity-related social closeness, is robust against extinction, as individuals recalibrate the outcome value associated with observing the other person receive painful vs. non-painful stimulation. Moreover, results from computational modelling indicate that the outcome value of no-pain trials (non-reinforced trials) can lead to positive prediction errors, enabling an increase in empathy-related social closeness based on non-reinforced trials.

In a next step, we investigated the neural mechanisms underlying the observed sustainability of empathy-related closeness. As a manipulation check, first, we analysed the neural activation related to participants' emotion ratings after observing painful or non-painful stimulation in the treatment and the control partner. A parametric regressor contrasting trial type (painful/ non painful) revealed an increased activation for the processing of observed painful stimulation in regions that have been associated with empathy-for-pain (e.g., Beeney et al., 2011; Lamm et al., 2007; Naor et al., 2020; Shamay-Tsoory, 2011; Singer et al., 2004), including the anterior insula/ inferior frontal gyrus (peak coordinates: x = 38, y = 28, z = -4, P(whole-brain FWE-cluster-corrected) = .033, k = 143) and the bilateral temporo-parietal junction (TPJ, left hemisphere peak coordinates: x = -52, y = -52, z = 20, T(44) = 6.21, P <.001, k = 898; right hemisphere peak coordinates: x = 62, y = -48, z = 22, T(44) = 4.74, P <.001, k = 532, see **Figure S2.1.4** for visualization). Moreover, significant activation was observed in the right occipital pole (peak coordinates: x = 16, y = -92, z = 8, P = .005, k = 214). Contrasting the results of the parametric regression between the empathy and the control condition revealed no significant results, which is expected given that on average participants observed the same number of pain trials in both conditions.

However, in contrast to the control condition, the neural activation in the treatment condition should be modulated by the recalibration parameter, i.e., the parameter that prevented a decline of empathy-related closeness in the extinction phase. To test this, we inspected whether the response to pain vs. no pain trials is differentially modulated by ω in treatment compared to control conditions.

The results revealed significant neural activation in the bilateral superior temporal sulcus (STS)/temporo-parietal junction (TPJ) (left hemisphere peak coordinates: x = -66, y = -26, z =

0, T(44) = 5.62, P <.001, k = 517; right hemisphere peak coordinates: x = 60, y = -16, z = 10, T(44) = 6.56, P <.001, k = 471) and in the left inferior frontal gyrus extending into anterior insula (IFG/AI; peak coordinates: x = -32, y = 16, z = 18, T(44) = 4.73, P = 001, k = 269). According to these results, the recalibration of emotion ratings when observing painful or non-painful stimulation in others is associated with changes in activation strength in bilateral STS/TPJ and IFG/AI (see **Figure 2.1.5A** for visualization).

Connection between closeness ratings and neural activation during emotional reaction

According to previous results, activation in the STS/TPJ has been linked to cognitive empathy and activation in AI/IFG has been linked to affective empathy (Böckler et al., 2014; Dvash & Shamay-Tsoory, 2014; Walter, 2012). The effect of the recalibrated feedback signal on neural responses in the acquisition and extinction phase in STS/TPJ may hence be more strongly modulated by individual differences in trait perspective-taking, while the recalibration effect in AI/IFG may be more strongly modulated by individual differences in empathic concern. In this vein, we extracted the beta estimates from the left IFG/AI and bilateral STS/TPJ, i.e., the regions associated with the recalibration of the feedback signal, and entered them as predictors in linear mixed models. In a first model, we included block averaged beta estimates extracted from IFG/AI and the factors block (acquisition vs. extinction), empathy subscale (empathic concern vs. perspective-taking subscale of the IRI), trait score, and their interaction as fixed effects, participant and trial number as random intercept and social closeness as dependent variable.

Results revealed that the relationship between neural sensitivity to observed pain vs. no pain and social closeness was modulated by block, empathy subscale, and trait score (beta estimates from IFG/AI × block × empathy subscale × trait score interaction: $\chi2$ = 3.95, P = .047, β = -.21, SE = .10).

Inspection of the visualized estimates showed that for individuals with high trait scores in empathic concern (M+1SD), weaker neural activation in response to observed pain (vs non-pain) was associated with increased social closeness in block 1 (simple slopes: β = -.28, SE = .09, 95%CI = [-.45, -.10]) but the reversed trend in block 2 (β = .16, SE = .09, 95%CI = [-.03, .34]). Individuals with low trait empathic concern (M–1SD), however, did not show this pattern (block 1: β = -.10, SE = .08, 95%CI = [-.26, .07]; block 2: β = -.10, SE = .07, 95%CI = [-

.24, .04]). Trait scores in perspective-taking did not differentially modulate the relationship between neural sensitivity to observed pain vs. no pain and social closeness in acquisition (block 1) and extinction (block 2) periods. That is, in both blocks, individuals with low scores in perspective-taking showed increased social closeness with decreased neural activation in response to observed pain than non-pain (block 1: β = -.22, SE = .08, 95%CI = [-.37, -.07]; block 2: β = -.08, SE = .07, 95%CI = [-.22, .07]) relative to individuals with high trait scores in perspective-taking (block 1: β = -.11, SE = .07, 95%CI = [-.26, .03]; block 2: β = .05, SE = .09, 95%CI = [-.12, .23]).

The analogous analysis with beta estimates from STS/TPJ revealed a main effect of neural sensitivity to observed pain vs. no pain in STS/TPJ and social closeness (beta estimates from STS/TPJ: χ2 = 6.49, P = .011, β = -.06, SE = .03). Thus, decreased neural activation to observed pain than observed non-pain was generally linked to increased social closeness, independently of block number and subscales of trait empathy) Independently of neural activation, results showed an interaction between block number, empathy subscale, and trait score (block × empathy subscale × trait score interaction: χ2 = 5.37, P = .021, β = -.07, SE = .03).

Discussion

Here we present the results of two independent behavioral studies, and one fMRI study, showing that empathy-related social closeness is preserved under conditions when empathy for pain is only rarely reinforced (**Figure 2.1.3A** and **C**). In contrast, social closeness incited by a social norm such as reciprocity decreased significantly in an equivalent condition in which reciprocity-inducing experiences occurred less frequently (**Figure 2.1.4A**). Uncovering the mechanism, computational modelling showed that the preservation of empathy-related closeness was best captured by a model assuming recalibration of the feedback signal. Again, this finding was replicated with an independent sample (**Figure 2.1.3B** and **D**).

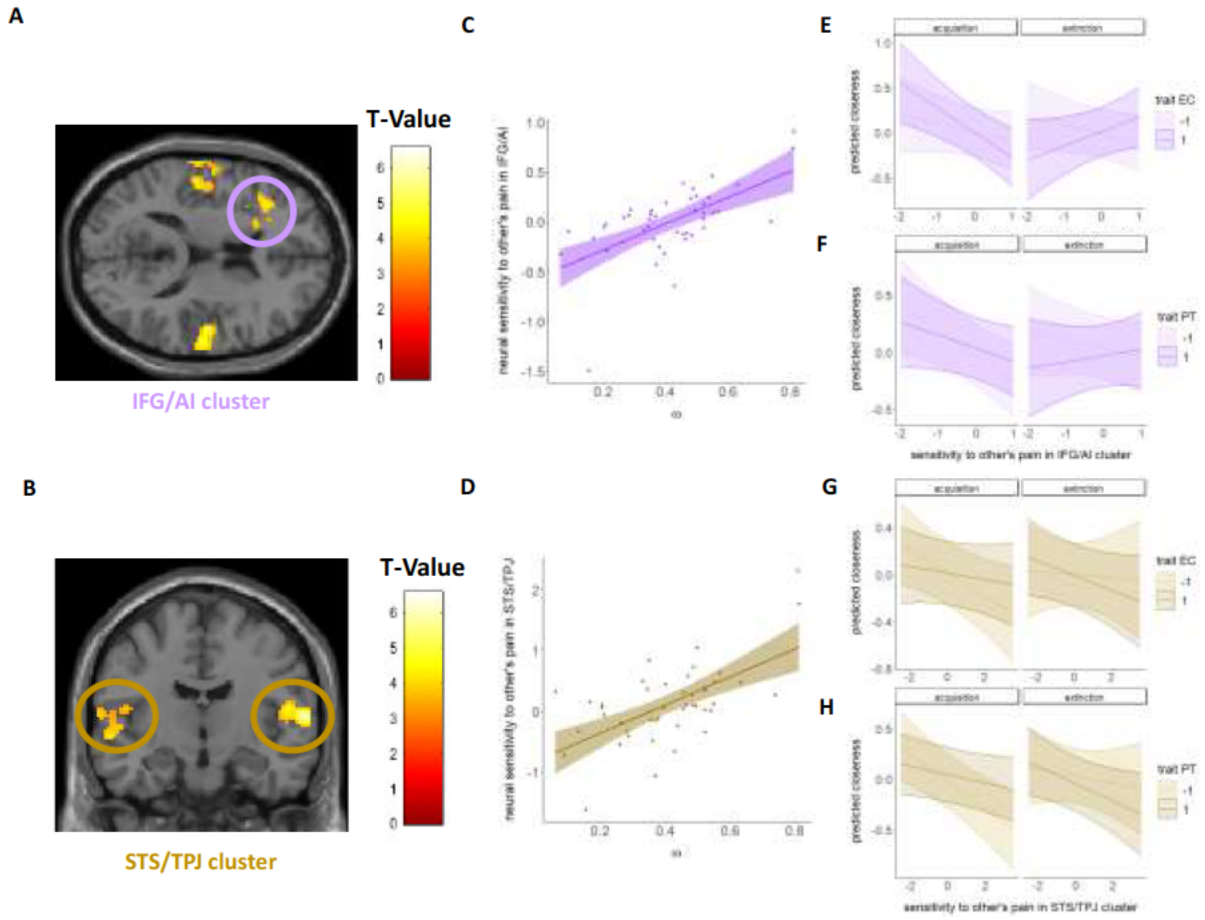


Figure 2.1.5 The extent of individual recalibration ω modulates the difference in neural tracking of trial type (painful stimulation of the other vs. non-painful stimulation of the other) between the treatment and the control condition in IFG/AI and STS/TPJ. The relationship between neural sensitivity to the other's pain is modulated by block (acquisition vs. extinction) and subscales of trait empathy. **A** Neural activation while participants evaluate their emotional reaction to the other's pain vs. no-pain: results of the second level regressions with the contrast parametric modulator trial type in the treatment condition > parametric modulator trial type in the control condition and recalibration parameter ω (extracted from the winning model) as covariate are shown. Individual recalibration modulated neural sensitivity to trial type more strongly in the treatment condition in contrast to the control condition in IFG/AI and **B** STS/TPJ. **C** Visualization of the whole-brain effect with individual recalibration values using the extracted beta values from IFG/AI and **D** STS/TPJ **E** Trait empathic concern modulated the relationship between neural sensitivity to trial type in IFG/AI and social closeness in acquisition and extinction. For high empathic individuals, larger responses to the observed non-pain vs. pain are associated with increased closeness during acquisition, whereas during extinction, larger responses to observed pain vs. non-pain are relatively more associated with increased closeness. This pattern does not hold for individuals low in trait empathic concern. **F** The analogous effect for trait perspective taking was less pronounced. **G** In STS/TPJ, increased activation in response to observed pain was generally associated with social closeness across acquisition and

extinction, a relationship not significantly modulated by trait empathic concern or H trait perspective-taking. For visualization purposes, maps were thresholded at P < .001 uncorrected with cluster size k ≥ 50.STS = superior temporal sulcus, TPJ = temporo-parietal junction, IFG = inferior frontal gyrus, AI = anterior insula, EC = empathic concern, PT = perspective-taking.

Follow-up modelling analyses (see supplementary computational results) showed that the formation of empathy-related social closeness was mainly driven by a learning signal that resulted from observing others in pain, whereas the maintenance of empathy-related closeness was driven by a learning signal resulting from observing no pain. Our computational results therefore suggest that initial social closeness is learned by frequently observing another in pain, while the maintenance of social closeness is accomplished via the positive associations of observing the relief of pain in others. In other words, social closeness based on empathy for pain may have been followed by social closeness based on empathic joy (Andreychik & Migliaccio, 2015; Batson et al., 1991), i.e., the joy of seeing less frequent pain in the other. This individual recalibration of the feedback was specific for empathy-related closeness as reciprocity-related social closeness did not show this pattern and declined if reciprocity was no longer reinforced – such behavioral pattern is best captured by a simple reinforcement model without a recalibration parameter (**Figure 2.1.4B**).

The recalibration of the learning feedback signal observed in our study is in line with previous studies that showed that the feedback value is susceptible to different learning contexts and can be individually adjusted (e.g., Bavard et al., 2018; Hunter & Daw, 2021; Pischedda, Palminteri, & Coricelli, 2020). For example, Hunter & Daw (2021) reported evidence that the uncertainty of reward in a given environment shapes the learning process. Extending such previous studies into the domain of social learning, our results show that these feedback recalibration mechanisms can preserve empathy-related closeness even if the other person is no longer suffering.

On the neural level, the recalibration of the feedback that resulted in the longevity of empathy-related social closeness was related to changes in activation in the anterior insula and adjacent inferior frontal gyrus, as well as the superior temporal sulcus, extending into the temporo-parietal junction. According to previous findings focusing on the processing of empathy-inducing events, both regions respond to the observation of another's pain (Lamm et al., 2007; Singer et al., 2004; Timmers et al., 2018). In more detail, the IFG/AI is part of a network that has been related to the processing of emotional empathy and the STS/TPJ is part of the network involved in the processing of cognitive empathy (Böckler et al., 2014; Naor et al., 2020; Shamay-Tsoory et al., 2009; Walter, 2012). In accordance with this previous evidence, our results show that closeness adaptation based on the recalibration of the closeness-preserving feedback signal in the IFG/AI was more strongly modulated by individual differences in trait empathic concern than trait perspective-taking. Specifically, high individual scores on empathic concern were related to increased social closeness based on stronger neural calibration in IFG/AI, indicated by stronger neural responses to observed lack of the other' s pain in the acquisition block (i.e., the block that is characterized by a high frequency of pain stimuli for the other). In the extinction block (i.e., the block that is characterized by a low frequency of pain stimuli for the other), stronger neural response to the other's pain were linked to increased social closeness. These regions hence appear to reflect a reversal in participants' learning signal from acquisition to extinction in individuals scoring high in empathic concern that is used to update social closeness. Interestingly, another recent study linked neural activation in the IFG to reappraisal of empathy for pain (Naor et al., 2020), with higher IFG activation being associated with stronger reappraisal. Together with these results, our findings indicate that the IFG and AI are implicated in the flexible adaptation of empathic responses and empathy-related learning.

In contrast, adaptation of social closeness based on the recalibration of the feedback signal in STS/TPJ was independent of empathic concern and perspective-taking. That is increased STS/TPJ activation in response to another's non-pain was generally linked to increased social closeness across acquisition and extinction (**Figure 2.1.5G** and **H**). This suggests that individual differences in empathic concern rather than perspective-taking are associated with a neural reversal that is linked to the recalibration mechanism supporting the persistence of empathy-related social closeness.

Together, the neural results add a novel aspect to existing findings, as they show that IFG/AI and STS/TPJ form a neural basis for a sustained effect of empathy on social behavior, here demonstrated with regard to social closeness.

Given evidence for gender differences in empathy (Christov-Moore et al., 2014), reciprocity (Dittrich, 2015), and prosocial behavior (Chowdhury, Jeon, & Saha, 2017), we recruited participants from the same gender (female) that were paired with same sex partners (confederates). Using a same-sex sample allowed us to control for potential unspecific gender effects. That said, we acknowledge that our results are based on a female sample which limits their generalizability. Future studies are required to replicate our results in male participants. Moreover, future study may test the longevity of empathy-related social closeness over longer periods of time and in every-day life settings.

Taken together, our studies demonstrate that empathy-related social closeness persists even when the other person is no longer suffering. Revealing the computational mechanism, we show that the longevity of empathy relies on a recalibration of the feedback value (i.e., the value associated with the information of whether a partner receives painful or non-painful stimulation), linked to neural responses in IFG/AI and STS/TPJ. These finding are important, because they show that once empathy is activated, empathy-related responses are a robust driver of social closeness.

Authors’ Contributions

G.H. and A.S. designed the research with input from J.B.E..; A.S. performed the research; A.S. programmed the experiment; A.S. and C.C.T. analyzed the data with input from G.H. and J.B.E.; G.H. and A.S. wrote the paper with input from C.C.T. and J.B.E..

Acknowledgments

We thank Kathrin Delius, Lea Brosig, and Larissa Böner for their invaluable support during data collection.

This work was supported by the German Research Foundation (HE 4566/2-1; HE 4566/5-1) and the German Academic Scholarship Foundation.

Implications for study 2

In study 1, we observed that empathy induces sustainable social closeness and explored the behavioral and neural mechanism underlying empathy-related social closeness sustainability. Results showed that when empathy is only rarely reinforced, social closeness still increases which can be explained by individual recalibration of what constitutes a reinforcing or a non-reinforcing stimulus. The extent of this individual recalibration was linked to increased neural sensitivity to observing painful vs. non-painful stimulation of the interaction partner in the temporo-parietal junction/superior temporal sulcus and inferior frontal gyrus/anterior insula. Moreover, results of the behavioral control study demonstrated that empathy induces social closeness more sustainably than the social motive reciprocity.

Whereas in study 1, we investigated empathy sustainability in terms of empathy-related social closeness, in study 2, we aimed to investigate how sustainably empathy leads to prosocial behavior. Using the confederate design in study 1, we tested how the process of making a prosocial as opposed to an egoistic decision is influenced by how strongly empathy is activated and whether the same neural regions are sensitive to this difference in empathy activation strength as were in the maintaining of social closeness. The prosocial decision process was characterized using drift-diffusion modelling and functional magnetic-resonance imaging.

2.2 Empathy incites a sustainable prosocial decisions bias

Anne Saulin[1*] & Grit Hein[1*]

[1] Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, 97080 Würzburg, Germany.

*corresponding authors:
Anne Saulin, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Saulin_A@ukw.de

Prof. Dr. Grit Hein, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Hein_G@ukw.de

Abstract

Empathy is a ubiquitous driver for prosocial behavior in daily life, especially towards those who are suffering. However, it is unclear whether empathy-related prosocial behaviour persists if the other person does not suffer anymore. Here we conducted two independent studies to investigate the longevity of empathy-related prosocial decisions and the underlying neural circuitries, and a third study to test the specificity of these effects. While undergoing functional magnetic resonance imaging (fMRI), participants performed a social decision task after observing pain in another person with high probability (high empathy block) and low probability (low empathy block). In a control condition, they performed the same task after observing others receiving pain stimulation at chance level. Drift-diffusion modelling results of two independent studies revealed an increased initial bias for making a prosocial decision after random and high empathy activation compared to baseline. Importantly, this bias was still evident in the low empathy block, i.e., after observing that the other person received pain with low frequency. A control study showed that the longevity of the prosocial decision bias was specific for empathy-based decisions and not evident when prosocial decisions were driven by a social norm like reciprocity. On the neural level, increased neural activation in the dorso-medial prefrontal cortex and temporo-parietal junction was predominantly linked to high trait empathic concern after frequent pain observation and to a larger individual general prosocial decision bias after rare pain observation. These results indicate that empathy leads to sustainable prosocial behavior, as indicated by a sustained prosocial decision bias and linked to differential neural responses in dmPFC and TPJ.

keywords:
empathy, prosocial behavior, drift-diffusion model, TPJ, dmPFC

Introduction

Empathy, i.e., sharing another person's affective state, is a principal driver for prosocial behavior (Batson et al., 2011, 1991; Hein, Morishima, et al., 2016). Especially empathy for the pain of another person has been consistently linked to an increase in prosocial decisions (e.g., Decety, Bartal, Uzefovsky, & Knafo-Noam, 2016; Hein, Morishima, et al., 2016; Hein, Silani, Preuschoff, Batson, & Singer, 2010).

On the neural level, the sharing of emotions (affective empathy) has been associated with neural responses in the anterior insula (AI), and the anterior and mid cingulate cortex (ACC/MCC; Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Fan, Duncan, de Greck, & Northoff, 2011; Preckel, Kanske, & Singer, 2018; Schurz et al., 2021; Stietz, Jauk, Krach, & Kanske, 2019). The sharing of other's intentions and thoughts (cognitive empathy) has been related to neural activation in the temporo-parietal junction (TPJ), the superior temporal sulcus (STS), the medial prefrontal cortex (mPFC), and the temporal poles (TP) (Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019).

The neural activation in regions associated with affective empathy (e.g., AI, ACC) and cognitive empathy (e.g., TPJ, mPFC), as well as decision circuitries (e.g., striatum) were linked to empathy-based social decision-making in contrast to social decision-making without explicit activation of empathy for pain (Hein, Morishima, et al., 2016; Krajbich, Hare, Bartling, Morishima, & Fehr, 2015; A. Tusche, Bockler, Kanske, Trautwein, & Singer, 2016).

Together, these previous studies show that empathy for pain increases the individual tendency to decide prosocially and characterized the underlying neural circuitries. However, it remains unclear whether the prosocial decision bias induced by empathy is stable, i.e., still evident if the other person is no longer suffering. Answering this question is important, because it addresses the longevity of response tendencies elicited by empathy, i.e., the motive that has been characterized as one of the strongest motivators of prosocial decisions.

To address this question, we conducted an fMRI study and an independent behavioural study in which we modelled and compared the prosocial decision-making process after the activation of high and subsequent low empathy and investigated the related changes in neural social decision circuitries. To induce empathy, we applied a well-established empathy for pain paradigm in which the participant observed how another person (confederate of the experimenter) received painful stimulation and rated how they felt when seeing the other in pain (e.g., Hein, Morishima, Leiberg, Sul, & Fehr, 2016; Saulin, Horn, Lotze, Kaiser, & Hein, 2022, see methods for details). In the treatment condition, to induce high empathy, participants observed pain in another person (confederate) with high probability (in 80% of all trials), to induce low empathy they observed pain in the other person with low probability (20% of all trials). After each of the empathy activation phases, participants performed a binary social decision task in which they could divide points in favour of themselves (egoistic choice) or in favour of another person (prosocial choice; **Figure 2.2.1A**). To investigate whether the prosocial response bias decays if empathy is reduced, the social decision task was divided in three comparable blocks. In a first block at the beginning of the study, participants performed the social decisions task before empathy activation (baseline block). In a second and a third block, they performed the same task after the activation of high and low empathy, respectively (**Figure 2.2.1B**). To compare potential changes in prosocial bias after high and subsequent low empathy activation, participants performed an analogous control condition, in which empathy was always activated at chance level (in 50% of the trials). To test whether our effects are specific for empathy-based prosocial decisions or also induced by other motivational states, in a third behavioural study, we used the identical design and social decision task, but induced a social norm (reciprocity) instead of empathy.

To assess the prosocial decision bias that was induced in the high and low empathy blocks, we used drift-diffusion modelling (DDM). In DDM, the decision process is conceptualized by continuous accumulation of evidence towards the different decision options which is characterized by three principal parameters (**Figure 2.2.1C**; Ratcliff, Smith, Brown, & McKoon, 2016; Voss, Rothermund, & Voss, 2004). Firstly, the drift rate or *v*-parameter, that indicates the speed of evidence accumulation. The *v*-parameter is hence a measure for the efficiency of the decision process. Secondly, the initial decision bias or *z*-parameter that captures potential decision biases prior to the evidence accumulation itself, and thirdly, the decision threshold or *a*-parameter that quantifies the relative amount of evidence needed in order to come to a decision and is an indicator of response caution (higher decision threshold indicates increased response caution). Previous works investigating different motivational propensities for making prosocial decisions suggest that the initial decision bias in particular may be sensitive to the strength of motive activation (Chen & Krajbich, 2018;

Saulin et al., 2022; Toelch et al., 2018). Moreover, changes in this parameter characterizing the motive-related increases in decision bias were associated with increased neural activation in striatum (Gluth, Rieskamp, & Büchel, 2012; Saulin et al., 2022) and anterior insula (Gluth et al., 2012).

Given that the *z*-parameter indicates an individual's response biases, we hypothesized that the activation of empathy increases the initial bias towards a prosocial decision compared to the baseline condition. With regard to the empathy activation, we assumed that observing pain in others with high frequency induces a stronger empathic reaction than observing pain in others with low frequency. In this case, empathy ratings in the low empathy block should decrease compared to the previous high empathy block. If the prosocial response bias decreases with decreasing empathy, the z-parameter capturing this bias should follow this pattern and should be smaller in the low empathy block than in the high empathy block. Alternatively, empathy may induce a sustainable prosocial decision bias that is still evident if the other person is no longer suffering. In this case, we should observe comparable z-parameters in the high and the low empathy block. In the control condition, we expected no analogous changes in prosocial bias. Hence, if empathy induces a sustainable prosocial decision bias, the results in the treatment and the control condition should be comparable.

On the neural level, an empathy-related increase in the initial response bias towards prosocial decisions may be associated with activational changes in neural regions related to affective empathy (AI, ACC), cognitive empathy (TPJ, dmPFC), and decision-making more generally (striatum, mPFC).

Material and methods

Participants

fMRI Study. Fifty-one right-handed female participants (mean age = 24.06, SD = 4.52) were recruited via online platforms and flyers on the university campus in Würzburg and two female confederates took part in the fMRI study. We chose female participants as well as female confederates in order to control for gender and avoid cross-gender effects. The confederates were students who had been trained to act as naïve participants. Participants did not know either of the confederates prior to the experiment. Before the experiment began, written informed consent was obtained from all the participants. The study was approved by the local ethics committee (268/18). Participants received monetary compensation.

We had to exclude 8 participants due to extreme behavior (these individuals always chose the egoistic choice option and hence did not react to the experimental empathy block 3), sleepiness (two participants), or technical problems (one participant). We thus analysed 43 data sets for the fMRI study. Post-hoc sensitivity analysis using GPower 3.1 with a = 0.05 and power (1-b) = 0.80 showed that using this sample sizes allowed for the detection of a true effect with an effect size of $f^2 = .16$ (F = 2.26).

Replication and control study. For the laboratory replication and the laboratory control study, fifty-six right-handed female participants (mean age = 22.98 years, SD = 3.39) were recruited via online platforms and flyers on the university campus in Würzburg and the same two female confederates as in the fMRI study took part in the laboratory studies. Participants were randomly assigned to the replication study or the control study. One participant in each study had to be excluded due to extreme and invariant behavior (these individuals always chose the egoistic choice option and hence did not react to the experimental empathy/reciprocity block 3). We thus analysed 54 data sets – 27 for each study. Post-hoc sensitivity analysis with a = 0.05 and power (1-b) = 0.80 showed that using these sample sizes allowed for the detection of a true effect with an effect size of $f^2 = .20$ (F = 2.29).

Procedure

fMRI study and laboratory replication study

Prior to the tasks, the individual thresholds for pain stimulation were determined for the participants and the confederates (for details, see section *Pain stimulation*). Next, the participants and confederates were assigned their different roles in a manipulated lottery of drawing matches. The participant always drew the last match in order to ensure she was assigned her designated role as the pain observer. The confederates were assigned the roles of pain recipients. In accordance with these roles, there were two parts of the experiment (corresponding to the two conditions treatment vs. control, see **Figure 2.2.1B** for an overview of the design).

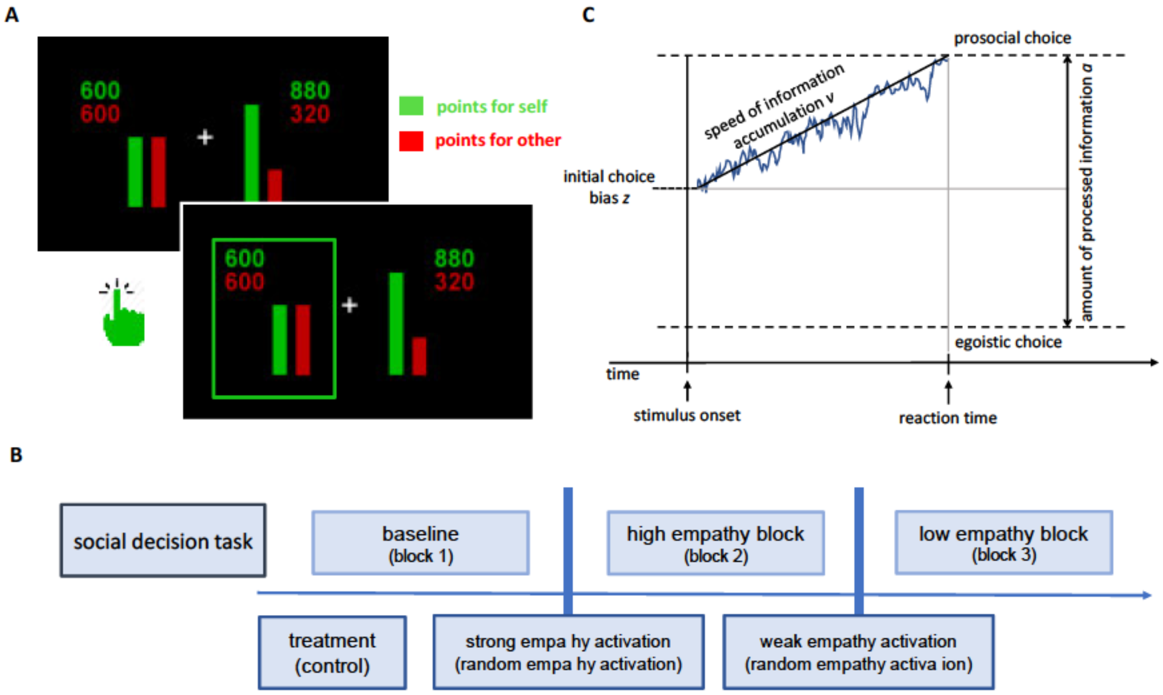


Figure 2.2.1 Overview of the social decisions task, visualization of the drift-diffusion model, and experimental design. **A** In each trial of the social decision task, participants chose between a prosocial and an egoistic way of dividing points between themselves (green bar and point values) and the respective partner (here red bar and point values). In this example, the participant chose the prosocial option (highlighted by a green rectangle). **B** Participants interacted with two different partners (confederates of the experimenter) in correspondence with the respective condition (treatment vs. control). After the first social decision task block (baseline block) and a first motive activation phase of strong empathy activation (participants observed painful stimulation of the partner in 80% of the activation phase), participants performed a second block of the social decision task (high empathy block). Behavior in this block should reflect the *initial* effect of empathy on prosocial behavior. After only weak activation of the empathy motive (20% of the trials of the motive task), participants performed a final block of the social decision task (low empathy block). Behavior in this final block should reflect a *sustained* effect of empathy on prosocial decision behavior. In the control condition, participants performed the same tasks with the other partner. However, the empathy motive was activated at chance level in both blocks of the motive activation phase (50 % of the activation phase, respectively). **C** The drift-diffusion model conceptualizes the decision process as noisy accumulation of information (squiggly blue line). The *v*-parameter describes the speed of information accumulation to choose one of the options, i.e., the efficiency of the decision process itself. The *z*-parameter reflects the initial decision bias, i.e., the degree to which an individual prefers one of the decision options prior to entering the decision process itself. The third component, the *a*-parameter, quantifies the amount of relative evidence that is required to choose one of the options. Once the accumulated information reaches either boundary, the decision is made (upper boundary = prosocial decision; lower boundary = egoistic decision).

In part 1, participants performed the interaction blocks with the first confederate while the other confederate ostensibly filled out questionnaires. In part 2, participants performed the interaction blocks with the second confederate and the first confederate ostensibly filled out questionnaires and then left. To ease identification during the interaction, the confederates were matched with a specific color (counterbalanced across participants).

Each interaction part comprised three blocks of the social decision task (baseline, high empathy block, low empathy block). In accordance with the condition specific activation strengths, participants underwent an empathy activation phase after the baseline block, as well as after the high empathy block of the social decision task (for visualization of this motive activation task and an overview of participants' behavior, see supplementary online **Figure S2.2.1**). The order of conditions was counter-balanced across participants. In the treatment condition, empathy was strongly activated after the baseline block (80% of the trials), and only weakly activated after the high empathy block (20% of the trials). In the control condition, empathy was activated at chance level after the baseline as well as after block 2 (50%, respectively). Importantly, on average, participants observed the empathy activating stimulus equally often in both conditions controlling for overall observations of painful stimulations.

In the fMRI study, the respective confederate was seated on a chair to the left of the participant with her hand visible by the participant. After both conditions were performed, the second confederate left and the participant remained in the scanner for anatomical image acquisition. Finally, participants filled out questionnaires measuring trait facets of empathy (IRI, Davis, 1983; Jordan et al., 2016) and reciprocity (Perugini et al., 2003) as well as how the ostensible other participants were perceived. Participants spent approximately 60 minutes in the scanner and the entire procedure took about 2.5 hours.

In the laboratory replication study, the respective confederate was seated next to the participant in a soundproof cabin facing the opposite direction such that no one could see the other's screen. After both conditions were completed, the confederate left and participants remained in the cabin to complete the same questionnaires as in the fMRI study assessing trait empathy and reciprocity and the participant's impression of the ostensible other participants. The whole procedure lasted approximately 2 hours.

To avoid possible reputation effects (e.g., Engelmann & Fischbacher, 2009; Gächter & Falk, 2002), which could influence participants' behavior, participants were informed that they would not meet the confederates after the experiment.

Laboratory control study (reciprocity motive)

Analogously to the fMRI study and the laboratory replication study, the individual thresholds for pain stimulation were first determined for the participants and the confederates (for details, see section *Pain stimulation*). Next, the participants and confederates were assigned different roles for the subsequent interaction tasks using a manipulated lottery of drawing matches. The participant always drew the last match in order to ensure she was assigned her designated role as the decider in the social decision task (in order to ensure that the participant is the one making the decisions) and pain recipient during the motive activation phase (in order to ensure that the participant is the one who can receive help from the interaction partner or not). The confederates were assigned the roles of potential helper during the motive activation phase who can decide to reduce the number of painful stimulation for the participant. In accordance with these roles, there were the same two parts as in studies 1 and 2 (corresponding to the two conditions treatment vs. control). In part 1, participants performed the interaction blocks with the first confederate while the other confederate ostensibly filled out questionnaires. In part 2, participants performed the interaction blocks with the second confederate and the first confederate ostensibly filled out questionnaires and then left. To ease identification during the interaction blocks, the confederates were matched with a specific color (counterbalanced across participants). Each interaction part comprised three blocks of the social decision task (baseline, high reciprocity block, low reciprocity block). In accordance with the condition specific activation strengths, participants performed a reciprocity activation after the baseline block, as well as after the high reciprocity block of the social decision task (for visualization of this motive task and an overview of participants' behavior, see supplementary **Figure S2.2.2**). The order of conditions was counter-balanced across participants. In the treatment condition, reciprocity was strongly activated after the baseline block (80% of the trials, i.e., in 80% of the trials the interaction partner decided to forgo a monetary reward in order to reduce the number of painful stimulations for the participant), and only weakly activated after block 2 (20% of the trials). In the control condition, reciprocity was activated at chance level after the baseline block as well as after block 2 (50%, respectively). Centrally, on average, participants observed the reciprocity activating stimulus equally often in both conditions.

Social decision task

The social decision task was a two-alternative-forced-choice adaptation of the commonly used Dictator Game (Forsythe et al., 1994), which has been successfully used in previous studies (e.g., Chen and Krajbich, 2018; Hein et al., 2016; Krajbich et al., 2015). In each trial of this decision task, participants allocated money to themselves and to the respective partner (**Figure 2.2.1A**) and could choose between maximizing the relative outcome of the other person by reducing their own relative outcome (prosocial decision) and maximizing their own relative outcome at a cost to the partner (egoistic decision). The outcome was relative to the outcome that the participant would have gained when choosing the other option. The initial number of points was always higher for the participant compared to the partners. This measure was inspired by previous behavioral economics research, showing that prosocial behaviors depend on the initial payoff allocation between the participant and the participant's partner (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Fehr and Schmidt, 1999). In particular, if subjects have a lower initial payoff than their partner (“disadvantageous initial inequality”), they are much less willing to behave altruistically toward the partner compared to a situation with advantageous initial inequality (i.e., when the participant has a higher initial payoff than the partner). The decision options used in the present study created advantageous inequality to optimize the number of prosocial decisions, which was the main focus of our study.

Participants sequentially performed two conditions (treatment condition and control condition). In the treatment condition, participants performed the task with a first interaction partner and in the control condition with a second interaction partner (both interaction partners were confederates of the experimenter). Importantly, the decision task blocks were identical in both conditions.

Each decision-trial started with a fixation cross shown for 1,000 ms, indicating the next interaction partner. After this cue, participants saw the two possible distributions of points in different colors, indicating the potential gain for the participant and for the current partner. Color designation of the partner was counterbalanced across participants. Participants had to choose one of the distributions within 4,000 ms. A green box appeared around the distribution that was selected by the participant at 4,000 ms after distribution onset. The box was shown for 1,000 ms. The length of the inter-trial interval, as indicated by a fixation cross, was jittered between 4,000 and 6,000 ms. At the end of the experiment, two of the distributions chosen by the participant were randomly selected for payment (100 points = 50 cents). We analyzed 25 (34 in the laboratory studies) decision trials in each condition, i.e., 150 (204 in the laboratory studies) trials in total per participant. For each condition and participant, the same distributions were used and presented in random order.

Pain stimulation

In the fMRI study, painful stimulation was applied using Digitimer DS7A constant current stimulator (Hertfordshire, United Kingdom) and a surface electrode attached to the left lower inner arm. Shock segments consisted of a single 1 ms square-wave pulses.

The criterion for painful stimulation was a subjective value of 8 on a pain scale ranging from 1 (no pain at all, but a participant could feel a slight tingling) to 10 (extreme, hardly bearable pain). The participants were told that a value of 8 corresponded to a painful, but bearable stimulus, and a non-painful stimulus corresponded to a value of 1 on the same subjective pain scale. These subjective pain thresholds were determined using a stepwise increase in current (stepsize of 0.05 mA), starting with a current of 0.00 mA, and increasing in stimulus intensity until it reached a level that corresponded to the individual's value of 8 (range = 0.25-1.50 mA).

For pain stimulation in the laboratory replication and the laboratory comparison studies, we used a mechano-tactile stimulus generated by a small plastic cylinder (612 g). The projectile was shot against the cuticle of the left index finger using air pressure (Impact Stimulator, Labortechnik Franken, Release 1.0.0.34). The same threshold procedure was used as in study 1, applying a stepwise increase of air pressure (stepsize of 0.25 mg/s), starting with the lowest possible pressure (0.25 mg/s), which caused the projectile to barely touch the cuticle, and increasing in stimulus intensity until it reached a level that corresponded to the individual's value of 8 (range = 2.00–6.00 mg/s).

Regression analyses

To test whether emotion ratings during the empathy activation phase and trait empathy influenced participants' frequencies of prosocial decisions, we first conducted a regression analysis with the predictors emotion ratings in the previous empathy activation phase, self-reported trait empathy as assessed with the empathic concern subscale of the IRI (Davis, 1983), and their interaction and the frequency of prosocial decisions as dependent variable. As control variables we included block (high empathy block/block 2 vs. low empathy block/block 3), condition (control vs. treatment), and study (scanner vs. laboratory). The analogous analysis was conducted for the laboratory control study using participants trait positive reciprocity (subscale of the PNR (Perugini et al., 2003)) instead of trait empathy.

In all studies, to test the influence of condition and block on participants' decision behavior, we conducted logistic mixed models with condition (treatment vs. control), block (baseline, initial response, sustained response), and their interaction as fixed effects, participant as random intercept and trial-by-trial binary decisions as dependent variable.

To test whether prosocial behavior was influenced by trial-by-trial point information, condition and their interaction, a logistic mixed model regression was conducted with the possible gain for the partner (i.e. difference in points between the two options for the partner, |partner's gain option 1 – partner's gain option 2|cf. Saulin et al. (2022), condition, block, and their interaction as fixed effects, participant as random intercept, and trial-by-trial binary decisions as dependent variable. Additionally, we estimated psychometric functions using the R-package "quickpsy" (Linares & López-Moliner, 2016) to model the relationship between the trial-by-trial possible gain for the partner and participants social decision behavior in each study.

With the extracted betas (see section *Second-level analyses* for details) from bilateral temporo-parietal junction (TPJ) and medial prefrontal cortex (mPFC), we conducted follow-up analyses to test the link between a sustained and initial neural empathy effect and trait and state characteristics of the participants in the fMRI study.

First, in an initial model, we included the predictors effect in prosocial decision bias (increase in *z*-parameter after the second empathy activation phase vs. increase in z after the second empathy activation phase), effect value (individual *z*-parameter beta weights of the DDM regression corresponding to the respective effect), neural region (TPJ vs. mPFC), and their interactions as predictors, and extracted neural betas as dependent variable. In an additional exploratory model, the same analysis was conducted but adding the general prosocial decision bias (i.e., the intercept of the *z*-parameter DDM regression) as additional level to the factor “effect in prosocial decision bias”.

Based on these results we conducted the main regression analysis with measure (affective empathy vs. general prosocial decision bias), measurement value (empathy trait score of the empathic concern subscale of the IRI (Davis, 1980) vs. *z*-intercept values resulting from the DDM), and neural region (TPJ vs. mPFC), and their interactions as predictors, and extracted neural betas as dependent variable. The analogous analysis was conducted for the perspective-taking subscale of the IRI instead of empathic concern. Please note that we conducted separate models for EC and PT due to high collinearity (spearman correlation: $\rho_{EC,PT}$ = .46, S = 7219, P = .002, N = 43).

Linear mixed model analyses were conducted in R (R version 4.0.4, R Core-Team, 2018) using the packages *lme4* (Bates et al., 2014) and *car* (Fox et al., 2018). For mixed models, we report the chi-square values derived from Wald chisquare tests using type 3 sum of squares from the ‘Anova()’ function (*car* package). We report the t-values derived from the summary() function and where of interest simple slopes as provided by the ‘emtrends’ function (*emmeans* package (Lenth et al., 2019)).

Drift-diffusion modelling

We used hierarchical regression drift-diffusion modeling (HDDM) (Vandekerckhove, Tuerlinckx, & Lee, 2011; Wiecki et al., 2013), which is a version of the classical drift-diffusion model that exploits between-subject and within-subject variability using Bayesian parameter estimation methods, because it is ideal for use with relatively small sample sizes and trial numbers. The analyses were conducted using the python implementation of HDDM regression version 0.8.0 (Wiecki et al., 2013).

Based on binary decisions, the HDDM approach provides detailed insights into the computation of egoistic and prosocial decisions, because it uses all the raw data that is available (trial-by-trial response times and decision outcome information of all decisions, irrespective of point distributions) to estimate sub-components of the underlying decision process. The v, z and a-parameters for each participant capture how each person maneuvers between the egoistic and the prosocial decision options, and finally approaches a decision boundary (i.e., the boundary for an egoistic or a prosocial decision, see e.g., Chen & Krajbich, 2018; Gallotti & Grujić, 2019, for comparable approach).

In the drift-diffusion modelling analyses, we initially compared how well five different models (see **Table 2.2.1** for overview of the model space) described participant's behavior. Since the partner's possible gain (*other possible gain*) significantly influenced individual *v*-parameters, this regressor was included as trial-by-trial influence on the *v*-parameter in all models (except the null model).

Table 2.2.1 Overview of the model space of the models tested to describe the empathy-based and reciprocity-based social decision process. The null model assumed no influences of manipulated factors on the decision process at all (M0) and V1 assumed that the trial-by-trial point information influences the *v*-parameter (V1). Three additional models assumed that the z-parameter is influenced by either condition (VZ1), condition and block (VZ2), or condition and block as well as their interaction (VZ3).

model	label	**specification**
null model	M0	-
test influence of point information on *v*	V1	*v ~ other possible gain*
influence on *v*-parameter and *z*-parameter models	VZ1	*v ~ other possible gain* *z ~ condition*
	VZ2	*v ~ other possible gain* *z ~ condition + block*
	VZ3	*v ~ other possible gain* *z ~ condition*block*

Apart from the parameter of interest for our research question (*z*), additional parameters are included in the estimation procedure. We also estimated the non-decision time t and allowed for trial-by-trial variations of the initial bias (*sz*), the drift rate (*sv*) and the non-decision time (*st*). These parameters were not estimated to vary by condition or block. They were however included based on results by Lerche & Voss (2016), who showed that in most cases, it is beneficial to include these parameters in order to improve model fit. In the estimation procedures we used the default settings for the priors and hyperpriors provided by the HDDM package. Specifically, the "informative group mean priors are created to roughly match parameter values reported in the literature and collected by (Matzke & Wagenmakers, 2009)" cited from Wiecki et al. (2013), Supplementary Material, page 1. Model convergence was checked by visual inspection of the estimation chain of the posteriors, as well as computing the Gelman-Rubin Geweke statistic for convergence (all values < 1.01) (Gelman & Rubin, 1992). To assess model fit, we conducted posterior predictive checks by comparing the observed data with 500 datasets simulated by our model (Wiecki et al., 2013). This approach allows for the computation of intervals within which the parameter falls with 95% probability. If the observed data falls within the 95% credibility interval of the simulated data, it can be assumed that the model can describe the data well enough.

To relate individual changes in the *z*-parameter to additional variables, we extracted each participant's regression weights for all effect included in the winning model. Please note that the HDDM regression procedure estimates regression weight for all effects included and an intercept, but not explicit parameter estimates for each factor level separately.

fMRI data acquisition

Imaging data was collected at a 3T MRI-scanner (Skyra syngo, Siemens, Erlangen, Germany) with a 32-channel head coil. Functional imaging was performed with a multiband EPI sequence of 42 transversal slices oriented along the subjects' AC-PC plane and distance factor of 50% (multi-band acceleration factor of 2). The in plane resolution was 2 x 2 mm^2 and the slice thickness was 2 mm. The field of view was 216 x 216 mm^2, corresponding to an acquisition matrix of 108 x 108. The repetition time was 1340 ms, the echo time was 25 ms, and the flip angle was 60°. Structural imaging was conducted using a sagittal T1-weighted 3D MPRAGE with 240 slices, and a spatial resolution of 1 x 1 x 1 mm^3. The field of view was 256 x 256 mm^2, corresponding to an acquisition matrix of 256 x 256. The repetition time was 2,300 ms, the echo time was 2.96 ms, the total acquisition time was 3:50 min, and the flip angle was 9°. We obtained, on average, 1,215 (SE = 5.07 volumes) EPI-volumes in the control condition and 1,208 (SE = 4.26 volumes) EPI volumes in the treatment condition for each participant. We used a rubber foam head restraint to avoid head movements.

fMRI Preprocessing

Preprocessing and statistical parametric mapping were performed with SPM12 (Wellcome Department of Neuroscience, London, UK) and Matlab version 9.2 (MathWorks Inc; Natick, MA). Spatial preprocessing included realignment to the first scan and unwarping and coregistration to the T1 anatomical volume images. Unwarping of geometrically distorted EPIs was performed using the FieldMap Toolbox. T1-weighted images were segmented to localize grey and white matter, and cerebro-spinal fluid. This segmentation was the basis for the creation of a DARTEL Template and spatial normalization to Montreal Neurological Institute (MNI) space, including smoothing with a 6 mm (full width at half maximum) Gaussian Kernel filter to improve the signal-to-noise-ratio. To correct for low-frequency components, a high-pass filter with a cut-off of 128s was used.

fMRI statistical analysis

First-level analyses

First-level analyses were performed with a general linear model (GLM), using a canonical hemodynamic response function (HRF). Regressors were defined from stimulus onset until the individual response was made by pressing a button (resulting in a time window of 1,000 ms + individual response time). The first-level model included as regressors of interest the decision phase (distribution onset until button press) and the possible gain for the partner as parametric modulator (see section *Regression analyses* for details). Participants also performed another task within the same scanning session. This task was modelled with three additional regressors of no interest, accounting for neural activation in this task. Additionally, invalid trials (i.e., trials in which the participant did not respond during the social decision task) were modelled as a separate regressor. The residual effects of head motions were corrected for by including the six estimated motion parameters for each participant and each session as regressors of no interest. To allow for modelling all the conditions in one GLM, an additional regressor of no interest was included, which modelled the potential effects of session.

Second-level analyses

Based on the first-level model, we computed a second level factorial model with the factors condition and block and computed the main effect contrasts of block. Specifically, in a first step, we tested the main effect of block, i.e., increase in neural activation from each block to the next. Second, we tested, in which regions neural activation more strongly decreased after the second empathy activation phase compared to after the first empathy activation phase.

We then tested in how far the initial and the sustained neural empathy effect in regions associated with stronger increases after the second empathy activation phase were specifically linked to (i) the individual prosocial decision bias and (ii) emotional and cognitive trait empathy (empathic concern subscale and perspective-taking subscale of the IRI (Davis, 1980)). Specifically, we extracted the beta values from the contrast treatment high empathy block > treatment baseline block (initial empathy effect) and from the contrast treatment low empathy block > treatment high empathy block (sustained empathy effect) in bilateral temporo-parietal junction and medial prefrontal cortex (Tzourio-Mazoyer et al., 2002) using marsbar (Matthew Brett et al., 2002). We then entered them as dependent variable in regression models with the predictors effect (initial vs. sustained) and indicators of prosocial decision bias and trait empathy (see section *Regression analyses* for details).

Results

Behavioral results - fMRI study and laboratory replication study

Empathy-based prosocial decision behavior

In order to test whether, empathic reactions during the empathy activation phase and trait empathy (as assessed using the empathic concern subscale of the IRI (Davis, 1983)) influenced participants' behavior, we conducted a regression analysis with emotional reaction during empathy activation and trait empathy and their interaction as predictors, study, block, and condition as control variable, and participants's probability for making a prosocial decision as dependent variable. Results showed that the higher participants' self-reported trait, empathy, the more strongly participants emotional reaction to observing that the interaction partner received painful stimulation was associated with higher frequencies in prosocial decisions after empathy activation (emotion rating × trait empathy interaction: $\beta = .12$, SE = .06, T(38) = 2.00, P = .048).

Mixed-models analysis of prosocial decision frequency revealed that the frequency of prosocial decisions was comparable across samples (main effect of study: $\chi^2 = 1.87$, P = .17, β

= -.06, SE = .05, T(66) = -1.37) as well as between blocks and conditions (all main effects and interaction χ^2 < 2.72, Ps >.25, see **Figure 2.2.2A** for visualization and **Table 2.2.2** for full results). Hence, prosocial decision frequencies did not decay after only weak empathy activation, i.e., from block 2 to block 3, indicating sustainability of empathy-related prosocial decision-making.

Response times

To test whether response times were influenced by block and condition, we conducted the analogous mixed model analysis with the predictors block and condition, and sample as control variable with the dependent variable response time. This analysis revealed that response times were significantly influenced by block (main effect of block: χ^2 = 8.88, P = .01, $\beta_{\text{initial response}}$ = -.06, SE = .03, T(66) = -2.31, $\beta_{\text{sustained response}}$ = -.07, SE = .03, , T(66) = -2.79) and condition (main effect of condition: χ^2 = 10.18, P = .001, β = 0.08, SE = .03, T(66) = 3.19). That is participants were faster in the high empathy block (block 2) and the low empathy block (block 3) as compared to the baseline block and faster in the control condition as compared to the treatment condition. Response times were comparable across studies (main effect of study: χ^2 = .97, P = .79, β = 0.04, SE = .15, T(66) = 0.27; see **Figure 2.2.2B** for visualization).

Table 2.2.2 Results of the linear mixed models analysis with block, condition, and their interaction as fixed effects, study as control variable, participant as random intercept and the frequency of empathy-based prosocial decisions in the fMRI study and the laboratory replication study as dependent variable (N = 70, maximal VIF = 7.98, 11759 observations).

Factor	Coefficicent	B	SE	T(63)	χ^2	P(χ^2)
	Intercept	.62	.04	16.62	276	
Block	Block 2	-.009	.01	-.65	1.12	.58
	Block 3	.006	.01	.40		
Condition	treatment condition	.006	.01	.45	.20	.65
Study	replication study	-.06	.05	-1.37	1.87	.17
condition:block	Block 2:treatment condition	.018	.02	.90	2.72	.27
	Block 3:treatment condition	-.015	.02	-.75		

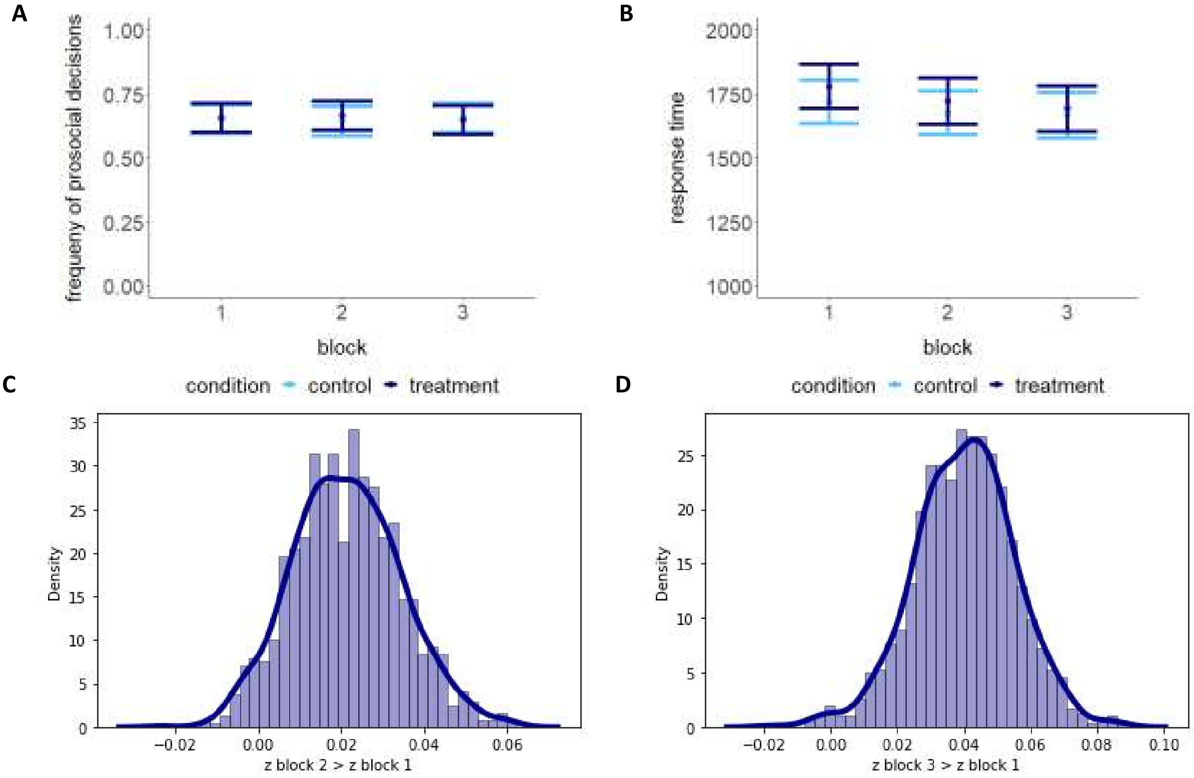


Figure 2.2.2 Mean relative frequencies and SEMs of empathy-based prosocial decisions, response times, and changes in initial prosocial decision bias across the fMRI study and the laboratory replication study. **A** Observable frequencies of empathy-based prosocial decisions were comparable across blocks and conditions. **B** Response times were faster in the control condition than in the treatment condition and faster in the high empathy block (block 2) and the low empathy block (block 3) as compared to the baseline block (block 1). **C** As indicated by the corresponding regression weights of the winning DDM (VZ3), the initial bias towards making a prosocial decision (*z*-parameter of the drift diffusion model) was increased from the baseline block (block 1) to the high empathy block/block 2 (probability = 94.9 %) **D** as well as comparing the baseline block (block 1) and the low empathy block/block 3 (probability = 98.8%) across both conditions.

Influence of point information

To test whether trial-by-trial point information influenced participants' behavior, we conducted the same analysis as above and added the trial-by-trial possible gain (see methods for computation of this variable) and its interactions as predictors. Results revealed a main effect of possible gain on the frequency of prosocial decisions (main effect of point information: χ^2 = 166.12, P < .001, β = .17, SE = .01, T(55) = 12.89), an effect that was tendentially larger in the laboratory replication study (point information × study interaction: χ^2 = 3.00, P = .08, β = .03, SE = .02, T(55) = 1.73). Thus, the more the other person could potentially gain from the participant choosing the prosocial distribution option, the more likely participants were to choose the prosocial option.

We further fitted psychometric functions to the data to model the relationship between trial-by-trial point information and the probability to make a prosocial decision. Results of this estimation showed that the values for the points of subjective equality based on the psychometric functions estimated were comparable across blocks and conditions (all Ps >.32, see supplementary **Figure S2.2.3** for visualization).

Drift diffusion modelling

Although, observable prosocial decision frequencies were not influenced by block or condition, the underlying decision process as characterized by drift-diffusion model components may be sensitive to empathy activation strength and may elucidate the mechanisms underlying the sustainability of empathy-based prosocial decision behavior.

In the drift-diffusion modelling analyses, we initially compared how well seven different models (see **Table 2.2.1** for overview of the model space) described participant's behavior. In all models (except for the null model), the partner's possible gain (other possible gain) was included as trial-by-trial influence on the *v*-parameter (cf. Saulin, Horn, Lotze, Kaiser, & Hein, 2022). Models varied with respect to whether condition, block, and their interaction influenced the *v*-parameter or the *z*-parameter.

The winning model for empathy-based decisions as indicated by the lowest DIC value was the most complex model (VZ3) (see supplementary **Table S2.2.1** for all DIC values). Thus, the model assuming an interaction effect between condition and block on the *z*-parameter best described the data compared to the other models included in the model space. Inspection of the posterior distribution of the effects' regression weights showed that the initial bias towards making a prosocial decision (*z*-parameter) was larger for the high empathy block (block 2) as compared to the baseline block (probability = 94.9%) and larger for the low empathy block (block 3) than the baseline block (98.8%). Hence, despite the comparable frequency of prosocial decisions across all blocks, the initial bias towards making this decision was increased in both conditions with respect to baseline (see **Figure 2.2.2B** and **C** for visualization).The probability for a larger initial prosocial decision bias in the low empathy block (block 3) than in the high empathy block (block 2) was 85.6 %, i.e., close to the threshold of 90% credibility. Overall, the probability for an increase of initial prosocial decision bias over the course of the three blocks is hence 90.3%. across treatment and control condition.

Further, the probability for an overall larger initial prosocial bias in the treatment condition as compared to the control condition was 87.7%, i.e., just below the threshold of 90% credibility. Regarding interaction effects, inspection of the posterior distribution for a larger increase from the baseline block to the high empathy block (block 2) in the treatment than in the control condition yielded a probability of 58.9%. Hence, the initial increase after baseline was comparable in the two conditions. Moreover, the probability for a larger increase in initial prosocial decision bias from the high empathy block (block 2) to the low empathy block (block 3) in the control condition than in the treatment condition was 86.8 %, again close to the threshold of 90% credibility. However, results showed that the increase from the baseline block to the low empathy block (block 3) was larger in the treatment condition compared to the control condition (90.4 %). This suggests that although empathy was only weakly activated in the second empathy activation phase (treatment condition), the overall increase with respect to baseline is at least as large as after two phases of empathy activation at chance level (control condition). Additionally, the trial-by-trial point information about the partner's possible gain influenced the *v*-parameter (100%).

Behavioral results - laboratory control study

Reciprocity-based prosocial decision behavior

In a separate control study conducted in the laboratory, we tested whether the effect of sustainable motive-based prosocial decision-making generalized to other social motives. Specifically, we conducted the same experiment as in the laboratory replication study, but activated the reciprocity motive, that is the norm to return a previously received favor.

In a first step, we again tested whether emotional reactions during the reciprocity activation phase and trait positive reciprocity (as assessed using the positive reciprocity subscale of the PNR (Perugini et al., 2003)) influenced participants' behavior. We hence conducted a regression analysis with emotional reaction during reciprocity activation and trait positive reciprocity and their interaction as predictors, study, block, and condition as control variable, and participants's probability for making a prosocial decision as dependent variable. Results showed that in this model, none of the predictors of interest was associated with prosocial behavior. However, running separate models for emotional reactions and trait reciprocity revealed that, the better participants felt in response to the partner's decision during reciprocity activation, the more frequently they made a prosocial decision in the social decision task (β = .16, SE = .07, T(19) = 2.23, P = .028). Likewise, the higher an individual's self-reported trait positive reciprocity, the more frequently they made a prosocial decision in the social decision task (β = .14, SE = .05, T(19) = 3.01, P = .003).

Linear mixed models analyses with the predictors block, condition, and their interaction revealed a main effect of block (χ^2 = 10.96, P = .004, $\beta_{\text{block 2}}$= -.05, SE = .02, T(24) = -2.24, $\beta_{\text{block 3}}$= -.07, SE = .02, T(24) = -3.23) and a block X condition interaction (χ^2 = 23.76, P < .001, $\beta_{\text{block 2}}$:treatment condition= 0.13, SE = .03, T(24) = 4.25, $\beta_{\text{block 3}}$:treatment condition= .002, SE = .03, T(24) = .07, for visualization, see **Figure 2.2.3A**). Hence, the changes of prosocial behavior across the three blocks depended on the condition. In the control condition, prosocial decisions continuously decreased from each block to the next, whereas in the treatment condition, the frequency of prosocial decisions increased after strong reciprocity activation and decreased after subsequent weak reciprocity activation.

Response times

Conducting the analogous mixed model with individual response times as dependent variable showed that response times of reciprocity-based social decisions were affected by block (χ^2 = 38.24, P < .001, $\beta_{\text{block 2}}$= -.15, SE = .04, T(24) = -4.06, $\beta_{\text{block 3}}$= -.22, SE = .04, T(24) = -6.07; **Figure 2.2.3B**). Thus, response times were faster in the high reciprocity block and the low reciprocity block as compared to the baseline block. Other main effects or interactions did not reach significance (all Ps >.39).

Influence of point information

To further test whether trial-by-trial point information influenced participants' reciprocity-based behavior, we conducted the same analysis as above and added the trial-by-trial possible gain (see methods for computation of this variable) and its interactions as predictors. Results revealed a main effect of possible gain on the frequency of prosocial decisions (main effect of point information: χ^2 = 118.85, P < .001, β = .15, SE = .01, T(20) = 10.90), independently of block or condition (for interaction effects, all Ps > .42).

To model the relationship between trial-by-trial point information and the probability to make a prosocial decision, we again fitted psychometric functions to the data. Results of this estimation showed that the values for the points of subjective equality based on the psychometric functions estimated were comparable across blocks and conditions (all Ps >.27, see supplementary **Figure S2.2.4** for visualization).

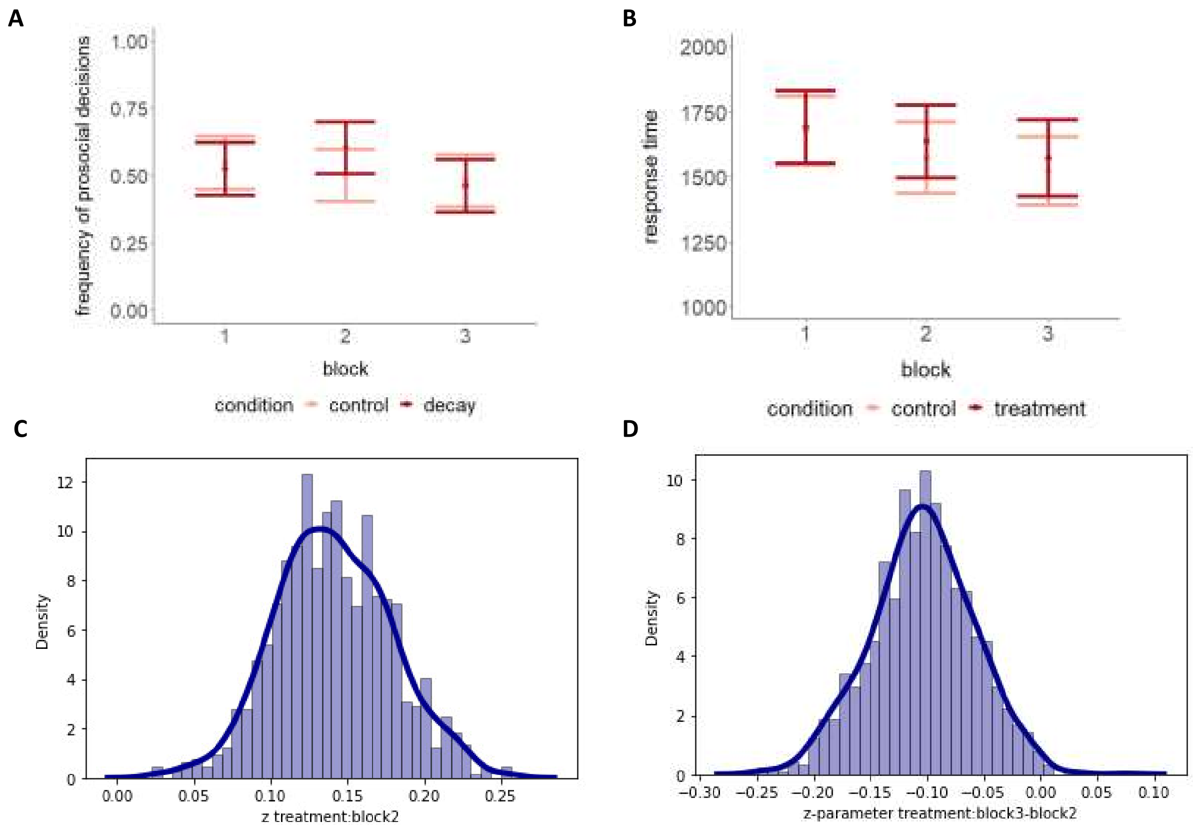


Figure 2.2.3 Mean relative frequencies with SEMs, mean response times with SEMs of reciprocity-based prosocial behavior, and changes in initial prosocial decision bias. **A** Observable frequencies of empathy-based prosocial decisions in the treatment condition (dark red) increased from before to after weak reciprocity activation and decreased after weak reciprocity activation. In the control condition (light red), the frequency of prosocial decisions gradually decreased from each block to the next. **B** Response times decreased from the baseline block (block 1) to the high reciprocity block (block 2) and the low reciprocity block (block 3) independently of condition. **C** The initial bias towards making a prosocial decision (*z*-parameter of the drift diffusion model) more strongly increased in the treatment condition than in the control condition from the baseline block (block 1) to the high reciprocity block (block2; probability = 100 %). **D** The initial bias towards making a prosocial decision (*z*-parameter of the drift diffusion model) more strongly decreased in the treatment condition than in the control condition from the high reciprocity block (block 2) to the low reciprocity block (block 3; probability = 99.4%).

Drift-diffusion modelling

Comparing the same seven models as for the empathy motive, the winning model as indicated by the lowest DIC value was the most complex model (VZ3) for empathy-based decisions (see supplementary **Table S2.2.1** for all DIC values). The trial-by-trial point information about the partner's possible gain influenced the *v*-parameter with over 95% probability (probability = 100%). For reciprocity-driven decisions, the initial bias towards making the prosocial decision was more strongly increasing from the baseline block to the high reciprocity (block 2) block in the treatment condition than in the control condition, (interaction effect condition × block effect: probability = 100%,). After the second phase of reciprocity activation, however, the initial bias decreased more strongly in the treatment condition as compared to the control condition (99.4%, **Figure 2.2.3C**). These findings mirror the pattern of observed prosocial decisions with an increase and a subsequent decrease in prosocial decisions in the treatment condition and a slight decrease over all blocks in the control condition.

fMRI results

Neural effects of block-wise increases in neural activation during the social decision process

DDM results of the empathy-related social decision process had revealed block-wise increases in prosocial decision bias across conditions. Based on these results, we next tested whether this block-dependent increase across the two conditions was also mirrored in changes of neural activation. Specifically, we tested in which regions neural activation increased after the first empathy activation phase and again after the second empathy activation phase, i.e., in which regions neural activation increased from each block to the next (baseline < block 2 & block 2 < block 3). Applying cluster-level whole-brain correction, this analysis revealed block-dependent increases in activation in bilateral striatum (left peak: x = -20, y = 10, z = -10, P < .001, k = 814; right peak: x = 18, y = 8 z = 6, P < .001, k = 487),and inferior frontal gyrus/anterior insula (left peak: x = -46, y = 8, z = 22, P = .022,k =174, right peak = x = 46, y = 6, z = 34, P = .043, k = 147), as well as left temporo-parietal junction (TPJ, peak: x = -38, y = 36, z = 18, P < .001), amongst other regions (see supplementary **Table S2.2.2**, for overview of full results at p<.001 uncorrected and k≥100). Hence, neural activation in these regions generally increased from each block to the next in both conditions.

Next, we aimed at more specifically testing whether increases in neural activation differed depending on whether this increase was in response to the first empathy activation phase compared to in response to the second empathy activation phase. In this vein, we contrasted the respective increases in neural activation ([baseline < block 2] vs. [block 2 < block 3]). The results revealed significant effects in the bilateral dorso-medial prefrontal cortex (dmPFC) and the right TPJ (dmPFC, left peak: x = -10, y = 50, z = 18, P (whole-brain cluster-corrected) = .001, k = 297, right peak: x = 18, y = 50, z = -28, P = .142, k = 103(TPJ, peak: x = 46, y = -60, z = 16, P = .022, k = 173, **Figure 2.2.4A** and **C**). On a lower, uncorrected threshold we also observed neural activation in the left temporal pole (**Table 2.2.3**). The reverse contrast revealed no significant effects (all Ps>.98 and ks <9). This indicates that neural activation in these regions increased more strongly after the second empathy activation phase than after the first empathy activation phase, making them potential candidate regions linked to a sustained effect of empathy.

Table 2.2.3 Results of the second-level analysis showing the regions with a larger increase in neural activation during the decision process after the second empathy activation phase as compared to after the first empathy activation phase. P<.001 uncorrected, k > 100.

Region	Hemisphere	T	FWE-P-value (cluster-level)	K	coordinates
temporo-parietal junction	Right	4.69	.022	173	46 -60 16
dorso-medial prefrontal cortex	Left	4.30	.001	297	-10 50 18
	Right	4.12	.142	103	18 50 28
temporal pole	Left	4.14	.124	108	-36 16 -28

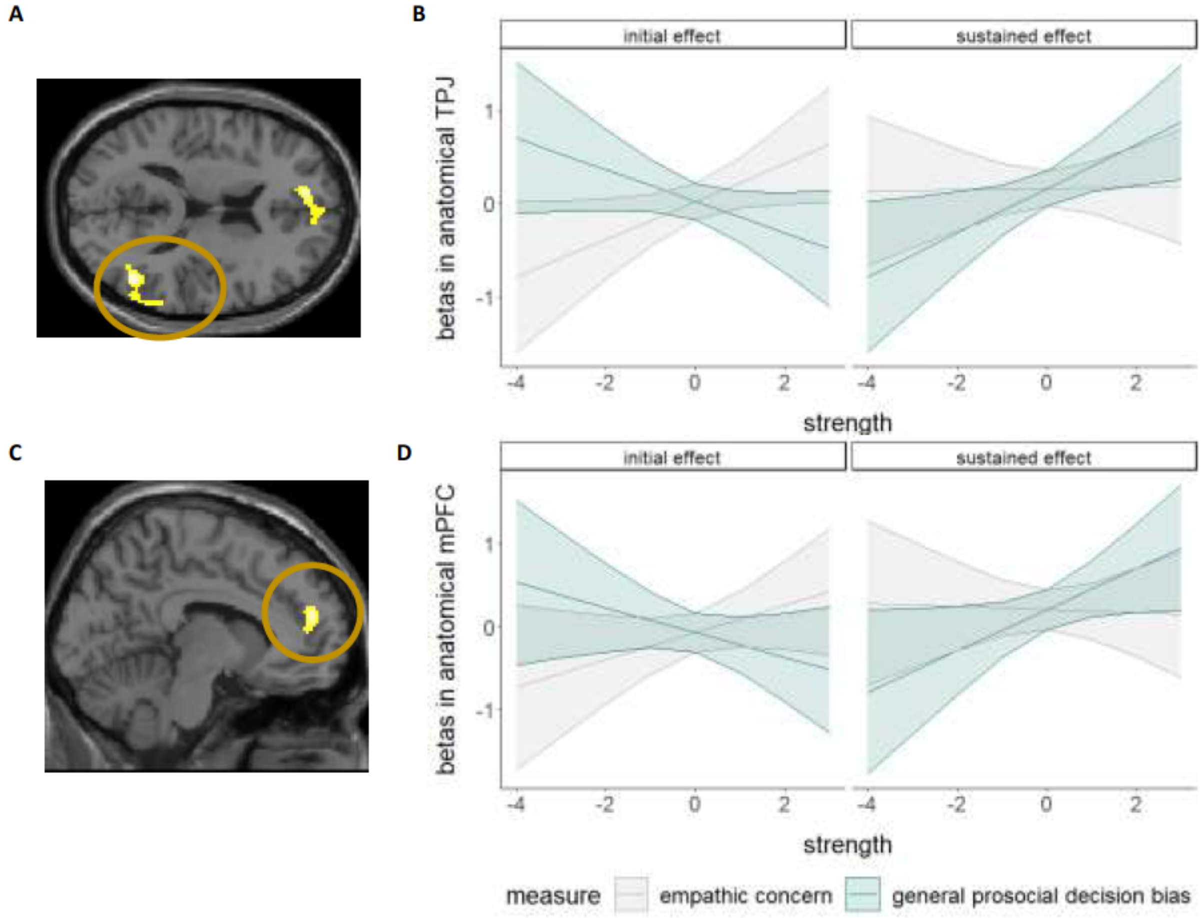


Figure 2.2.4 Regions of larger increase after the second empathy activation phase as compared to after the first empathy activation phase during the social decision process across conditions, and link to self-reported trait empathic concern and general prosocial decision bias in the treatment condition specifically. Across conditions (treatment and control), decision-related neural activation in **A** the right TPJ (peak: x = 46, y = -60, z = 16) and **C** the bilateral dmPFC (left peak: x = -10, y = 50, z = 18; right peak: x = 18, y = 50, z = -28) was significantly higher after the second as compared to after the first empathy activation phase. **B** Stronger activational changes in the TPJ and **D** the mPFC in response to an initial empathy effect during the empathy-based social decision process (i.e., increase after strong empathy activation compared to baseline) were positively associated with self-reported trait empathic concern (grey line, β = .20, SE = .112, 95% CI = [-.02, .42]; mPFC: β = .17, SE = .112,95% CI = [-.06, .38]). The sustained neural empathy effect during empathy-based social decision-making (i.e., increase after weak empathy activation compared to after previous strong activation) was positively associated with the general prosocial decision bias (green line, TPJ: β = .24, SE = .112, 95% CI = [.02, .46]; mPFC: β = .25, SE = .112,95% CI = [.03, .47]). TPJ = temporo-parietal junction, dmPFC = dorso-medial prefrontal cortex. Activation visualized at p<.001 uncorrected and k >100.

Relationship between neural responses after strong and weak empathy activation is selectively modulated by empathic concern and general prosocial decision bias

Behavioral results had shown that block-wise increases in prosocial decision bias were comparable across conditions. Hence, in the previous section, we first tested for analogous effects on a neural level and observed that neural activation in regions previously linked to empathy (IFG, AI, TPJ) and motivation (striatum) increased from each block to the next. Moreover, neural activation in dmPFC and TPJ yielded larger increases in response to the second empathy activation phase than in response to the first empathy activation phase. In line with our research question, we next aimed at understanding the modulators of a sustained effect of empathy activation in contrast to an initial effect of empathy activation in the treatment condition, i.e., after strong and subsequent weak empathy activation. To test this, we extracted the contrast beta estimates from anatomically defined temporo-parietal junction (TPJ) and medial prefrontal cortex (mPFC) corresponding to the initial effect of empathy in the treatment condition (treatment high empathy block > treatment baseline block) and the sustained effect of empathy in the treatment condition (treatment low empathy block > treatment high empathy block), to test whether these increases in neural activation were linked to participants' prosocial decision bias.

Participants' block-specific increases in prosocial decision bias, i.e., the regression weights indicating the block-specific increases of the *z*-parameter, were not related to a sustained neural empathy effect (simple slopes: TPJ: β = .05, SE = .11, 95% CI = [-.16, .26]; mPFC: β = .04, SE = .11,95% CI = [-.17, .25], see supplement for full results). This makes sense as the prosocial decision bias did not strongly increase after weak empathy activation. An additional exploratory analysis showed that the general prosocial decision bias (intercept of the *z*-parameter), which is an indicator of general state prosocial decision bias, was linked to a sustained neural empathy effect (simple slopes: TPJ: β = .24, SE = .11, 95% CI = [.02, .46]; mPFC: β = .26, SE = .11,95% CI = [.04, .48], see supplement for full results).

Hence, to test whether the sustained vs. initial neural empathy effect were differentially linked to this general prosocial decision bias in concert with affective and cognitive trait empathy (empathic concern and perspective-taking subscales of the IRI (Davis, 2006)), we conducted two final linear models.

The predictors in the first model were measurement (self-reported trait empathic concern vs. prosocial decision bias), measurement value (the respective normalized score/decision bias), region (TPJ vs. mPFC), effect type (initial vs. sustained), and their interactions as predictors and neural betas as dependent variable. Results showed a significant measurement × measurement value × effect type interaction (β = .60, SE = .22, t = 2.67, P = .008, see **Figure 2.2.4B** and **D** for visualization). Simple slope inspection revealed that in both regions, self-reported trait empathic concern was tendentially positively related to the neural increase in response to the initial empathy effect (TPJ: β = .20, SE = .112, 95% CI = [-.02, .42]; mPFC: β = .17, SE = .112,95% CI = [-.06, .38]), whereas the general prosocial decision bias was positively related to the neural increase in response to the sustained empathy effect (TPJ:β = .24, SE = .112, 95% CI = [.02, .46]; mPFC: β = .25, SE = .112,95% CI = [.03, .47]). This shows that, in TPJ as well as mPFC, neural activation increased the more after strong empathy activation, the more trait empathic concern an individual reported, hence reflecting a link between empathic concern and a neural response to an initial empathy effect. After subsequent weak empathy activation, however, neural increases during the social decision process were linked to a person's overall prosocial decision bias in the experiment, demonstrating a link between the prosocial decision bias and neural responses to a sustained empathy effect.

The second model was conducted analogously including trait perspective-taking as predictor instead of empathic concern. This model showed a similar pattern of results as the model testing the role of empathic concern, however only on a marginal level (measurement × measurement value × effect type interaction β = .49, SE = .22, t = 1.74, P = .08). No other effects reached statistical significance (all Ps > .07, Ts < 1.78).

Discussion

In the studies presented here, we investigated the sustainability of the empathy-based social decision process and its neural underpinnings. The directly observable frequency of empathy-based prosocial decisions was sustainable and remained on a high level throughout all three blocks. Drift-diffusion modelling analyses showed that an individual's initial bias towards making a prosocial decision was increased after strong and random activation. Decisively, it did not decrease after subsequent weak activation. Additionally, we compared the sustainability of the empathy-based social decision process with the same decision process based on reciprocity. Reciprocity-based prosocial decisions also became more frequent after strong activation. However, prosocial decisions starkly decreased after weak activation of reciprocity. This pattern was again reflected in an initial increase and subsequent decrease of the initial prosocial decision bias, showing that empathy more sustainably leads to a prosocial decision bias than reciprocity.

On a neural level, the increase of prosocial decision bias over the course of all three blocks was reflected in increases in neural activation from each block to the next in striatum, inferior frontal gyrus/anterior insula, as well as temporo-parietal junction (TPJ). Additionally, neural activation in dorso-medial prefrontal cortex and TPJ specifically more strongly increased from after the second empathy activation phase as compared to after the first empathy activation phase across both conditions. Moreover, the sustained increase in neural activation after weak activation of the empathy motive was positively associated with a person’s overall bias towards making a prosocial decision, whereas the initial neural response was associated with individual trait empathy, especially empathic concern.

The present results are in accordance with previous works that showed that experimental manipulations (Saulin et al., 2022), peer influences (Yu et al., 2021), as well as trait propensities for making prosocial decisions (Chen & Krajbich, 2018) increased initial prosocial decision biases. A study from our group, for example showed that the initial bias towards making a prosocial decision (*z*-parameter of the drift-diffusion model) was larger after the combined activation of empathy and reciprocity compared to after activation of reciprocity only (Saulin et al., 2022). Hence, this parameter appears to be sensitive to prosocial motive activation strength. More generally, changes in the initial decision bias have also been linked to changes in motivational strength (Gluth et al., 2012; Leong et al., 2019; Mulder et al., 2012). The findings observed here thus extend this general notion by showing that the initial prosocial bias increases after activation of a prosocial motive, indicating an increase in prosocial motivation based on empathy or reciprocity. The persistently high prosocial decision bias after only weak empathy activation further suggests that not only the prosocial decision bias was increased, but that this bias may be an indicator of the underlying empathy motive strength itself, hinting towards empathy as a sustainable social motive.

Neural activation during the social decision process increased over the course of the three blocks in both conditions in the inferior frontal gyrus/anterior insula (TPJ/AI), the temporo-parietal junction (TPJ), and the striatum. Whereas IFG, TPJ, and AI have frequently been associated with empathy and mentalizing and can be understood as part of the empathy-mentalizing network (Bellucci, Camilleri, Eickhoff, & Krueger, 2020; Schurz et al., 2021), striatal activation during the (social) decision process is linked to (Balleine, Delgado, & Hikosaka, 2007; Izuma et al., 2008) increased motivation towards a specific goal (Liljeholm & O 'Doherty, 2012; Reeve & Lee, 2012).

IFG/AI activation, in particular is increased upon observing others in pain (Beeney et al., 2011; Y. Li et al., 2021; Saarela et al., 2007), during empathy-based social decision behavior (Hein, Morishima, et al., 2016), as well as in individuals with higher trait empathy (Banissy, Kanai, Walsh, & Rees, 2012; Y. Li et al., 2020). Activation in the TPJ has frequently been associated with theory of mind or mentalizing (Böckler et al., 2014; Saxe & Kanwisher, 2003; Schurz, Tholen, Perner, Mars, & Sallet, 2017). While previous work has highlighted specific activation of IFG/AI linked to affective empathy and activation of the TPJ linked to cognitive empathy (Böckler et al., 2014), co-activation of these regions during social decision-making is in line with more recent approaches acknowledging that during social tasks with real world relevance, affective as well as cognitive facets of empathy are relevant for performing the task (Schurz et al., 2021).

Block-wise increase of neural activation in the regions observed hence further supports the interpretation of an increasing empathic motivation after each empathy activation block.

Moreover, neural activation in the TPJ and dmPFC was more strongly increased after weak empathy activation than after previous strong empathy activation. Hence these regions of the brain were more strongly activated in the block measuring a response to a sustained empathy effect, suggesting an especially important role for empathy sustainability for these two regions. This hypothesis was corroborated by follow-up analyses that demonstrated that the larger an individual's increase in neural activation in mPFC and TPJ after strong empathy activation, capturing an initial effect of empathy on social decision-making, the higher this individual scored on trait empathic concern. However, an individual's increase in neural activation in mPFC and TPJ after weak empathy activation, capturing a sustained effect of empathy on social decision-making, was related to an individual's general prosocial decision bias. As participants already knew before the first empathy activation phase that the interaction partner will receive painful stimulation, this general prosocial decision bias may reflect participants increased prosocial tendency based on this expectation of future observed pain. This suggests that already anticipated empathy can be associated with a sustained neural empathy effect, a hypothesis that needs to be empirically tested in future studies.

The dmPFC and TPJ are part of the social brain network and have frequently been associated with mentalizing processes (Adolphs, 1999; Eres, Decety, Louis, & Molenberghs, 2015; Frith & Frith, 2006; Park et al., 2017; Powers, Chavez, & Heatherton, 2015; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Saxe & Kanwisher, 2003). That is, increased activation of these regions in a certain task was related to higher demands for mentalizing processes afforded by the task (e.g., Rilling et al., 2004; Schurz et al., 2017). The present results hence suggest that during the social decision process after strong empathy activation, making the decision in favor of the other or in favor of the self is closely linked to the empathic experience of observing the other receive painful stimulation which is amplified by high trait empathic concern and tendentially perspective-taking. In the later decision block, however, i.e., after weak empathy activation, the general prosocial decision bias as experienced in the experiment becomes relatively more important.

In the third study, we tested whether motive sustainability extends to other prosocial motives such as reciprocity, i.e., the norm to return a previously received favor. Results showed that for the reciprocity motive the initial prosocial bias after only weak activation of the reciprocity motive as well as the frequency of prosocial decisions decreased. These findings indicate that reciprocity-based prosocial behavior is not sustainable within the present framework and may hence be itself not a sustainable motive once it’s only weakly activated.

Taken together, using fMRI and drift-diffusion modelling, the studies presented here demonstrate that activation of the empathy motive incites a sustainable prosocial decisions bias. Specifically, strong and random activation increased the initial bias towards making a prosocial decision and subsequent weak activation did not entail a decrease in prosocial bias. Repeatedly activating empathy was associated with increasing activation in the ‘social brain’ (anterior insula, TPJ, IFG) and striatum during the empathy-based social decision process. Sustained empathic neural responses, i.e., stronger increase after weak activation than after previous strong activation, were observed in the dmPFC and TPJ. The initial empathic neural response in these regions was linked to trait empathic concern whereas the sustained empathic neural response was associated with individual prosocial decision biases.

Authors' Contributions

A.S. and G.H. designed the research.; A.S. programmed the experiment; A.S. performed the research; A.S. analyzed the data with input from G.H.; A.S. and G.H. wrote the paper.

Acknowledgments

We thank Kathrin Delius, Lea Brosig, and Larissa Böner for their invaluable support during data collection.

This work was supported by the German Research Foundation (HE 4566/2-1; HE 4566/5-1) and the German Academic Scholarship Foundation.

Implications for study 3

Studies 1 and 2 explored how sustainably empathy induces social closeness and prosocial behavior and how this is linked to the social brain. Results showed that when empathy was only rarely reinforced, social closeness still increased and empathy-related prosocial behavior and prosocial decision bias persisted even after only weak empathy activation. Neurally, empathy sustainability with regard to prosocial decision-making was linked to increased neural activation in the TPJ, dmPFC, IFG, AI, and striatum. Studies 1 and 2 further advocate for empathy being more sustainable than the social motive of reciprocity as reciprocity-related social closeness diminished after rare reinforcement and prosocial behavior based on reciprocity also declined after only weak activation. These studies investigated social behavior related to each motive separately. However, human behavior is often driven by more than one motive.

In the next study, we thus addressed the question of whether empathy in combination with reciprocity may be able to boost the prosocial decision process and related neural activation that is only based on reciprocity or only based on empathy. In this vein, we compared the prosocial decision process and its neural underpinnings after activation of both empathy as well as reciprocity with that based on reciprocity alone and empathy alone.

2.3 The neural computation of prosocial decisions in complex motivational states

Anne Saulin[1*], Ulrike Horn[2], Martin Lotze[3], Jochen Kaiser[4], & Grit Hein[1*]

[1] Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, 97080 Würzburg, Germany.

[2] Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany.

[3] Functional Imaging Unit, Institute of Diagnostic Radiology and Neuroradiology, University of Greifswald, 17489 Greifswald, Germany.

[4] Institute of Medical Psychology, Faculty of Medicine, Goethe University, 60528 Frankfurt am Main, Germany.

*corresponding authors:
Anne Saulin, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Saulin_A@ukw.de

Prof. Dr. Grit Hein, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Hein_G@ukw.de

Published 2022 in NeuroImage

The neural computation of prosocial decisions in complex motivational states

Abstract

Motives motivate human behavior. Most behaviors are driven by more than one motive, yet it is unclear how different motives interact and how such motive combinations affect the neural computation of the behaviors they drive. To answer this question, we induced two prosocial motives simultaneously (multi-motive condition) and separately (single motive conditions). After the different motive inductions, participants performed the same choice task in which they allocated points in favor of the other person (prosocial choice) or in favor of themselves (egoistic choice). We used fMRI to assess prosocial choice-related brain responses and drift-diffusion modelling to specify how motive combinations affect individual components of the choice process. Our results showed that the combination of the two motives in the multi-motive condition increased participants' choice biases prior to the behavior itself. On the neural level, these changes in initial prosocial bias were associated with neural responses in the bilateral dorsal striatum. In contrast, the efficiency of the prosocial decision process was comparable between the multi-motive and the single-motive conditions. These findings provide insights into the computation of prosocial choices in complex motivational states, the motivational setting that drives most human behaviors.

keywords:

motivation, social decision-making, hierarchical drift-diffusion modeling, fMRI, social neuroscience

The neural computation of prosocial decisions in complex motivational states

Introduction

All choice behaviors are incited by motives, which can be complex. Documenting this motivational complexity, many animal (Jennings et al., 2013; Kennedy & Shapiro, 2009) and most human behaviors are driven by multiple motives that are active at the same time, and affect each other (Engel & Zhurakhovska, 2016; Hughes & Zaki, 2015; Jagers, Linde, Martinsson, & Matti, 2017; Kruglanski et al., 2018; Lewin, Cartwright, & Price, 1951; Takeuchi, Bolino, & Lin, 2015; Terlecki & Buckner, 2015). For example, the decision to help an elderly relative is often driven by empathy with her needs, and at the same time, by the wish to reciprocate help received by this person in the past, i.e, the social norm of reciprocity. Consequently, most choice behaviors are driven by combinations of different motives and cannot be explained by one "motivational force" alone. However, the combination of motives is not directly observable. Thus, to understand and predict choice behaviors, it is crucial to elucidate the neuro-computational mechanisms through which multiple simultaneously activated motives affect behavioral choice processes.

The processing of single-motive states and its impact on behavioral choices in animals (e.g., place preferences) (Jennings et al., 2013) have been linked to dopaminergic neurons in the striatum (Kim & Im, 2018; Robinson, Sotak, During, & Palmiter, 2006; Salamone & Correa, 2012). In line with these results, human neuroscience studies have shown that the striatum is involved in the processing of different individual motives, as well as motivated choice behaviors, both in the social (Báez-Mendoza & Schultz, 2013; Bhanji & Delgado, 2014) and non-social domain (Salamone et al., 2016; Shohamy, 2011). In more detail, the ventral striatum has been linked to the learning and encoding of values and the predictions of future rewards (Kable & Glimcher, 2007; Liljeholm & O ’Doherty, 2012; O'Doherty et al., 2004; Strait et al., 2015), whereas the dorsal striatum has been linked to initiating and optimizing choices based on these encoded values (Balleine et al., 2007; Liljeholm & O ’Doherty, 2012; O'Doherty et al., 2004; Palmiter, 2008; Robinson et al., 2006).Together, this previous work has provided insights into the neural underpinnings of individual motivational processes. However, the neural computation of behaviors that are driven by different motives remains unclear.

To address this issue, we developed a paradigm in which participants made the same choices (prosocial vs. egoistic) based on different, simultaneously activated motives (multi-motive condition), or based on each of these motives separately (single-motive conditions). Specifically, we studied the effect of simultaneously activated social motives in a social choice paradigm in which participants repeatedly had the choice between a prosocial and an egoistic option (**Figure 2.3.1A**). Inspired by an influential model of prosocial motivations (Batson et al., 2011), we induced two key motives that both incite prosocial behavior - the empathy motive, defined as the affective response to another person's misfortune (Batson, Turk, Shaw, & Klein, 1995; Hein, Morishima, et al., 2016; Lamm et al., 2011), and the reciprocity motive, defined as the desire to reciprocate perceived kindness with a kind behavior (Gouldner, 1960; Hein, Morishima, et al., 2016; McCabe et al., 2003).

In combination with fMRI and hierarchical drift-diffusion modeling (hierarchical DDM) (Forstmann et al., 2016; Ratcliff et al., 2016; Vandekerckhove et al., 2011; Wiecki et al., 2013), this paradigm allowed us to specify how the combination of different motives affects individual components of neural goal-directed (i.e., prosocial) choice computation, compared to computation of the same choice in a simple motivational state (i.e., driven by only one of the two motives).

Drift-diffusion models (DDMs) characterize how noisy information is accumulated to select a choice option (**Figure 2.3.1B**) based on three different parameters (the *v*, *z* and *a* parameters) (Forstmann et al., 2016; Ratcliff et al., 2016). The *v*-parameter describes the speed at which information is accumulated in order to choose one of the options, i.e., the efficiency of the choice process itself. The *z*-parameter reflects the initial choice bias, i.e., the degree to which an individual prefers one of the choice options prior to making the choice. Thus, in contrast to the *v*-parameter, which models the choice process itself, the *z*-parameter models the individual bias with which a person enters the choice process. For example, if a person has a strong initial bias towards prosocial choices (reflected by a large value of the parameter *z*), the starting point of the choice computation is located closer to the prosocial choice boundary, and thus, this person is more likely to choose the prosocial option. The third component, parameter *a*, quantifies the amount of relative evidence that is required to choose one of the options.

Previous neuroscience studies have identified brain regions that are associated with changes of these choice parameters. For example, it has been shown that reward-related improvement of perceptual discrimination is driven by changes in the *z*-parameter, related to changes of frontoparietal activation (Mulder et al., 2012). Another study using a task from the perceptual domain has shown that increased evidence accumulation under time pressure is linked to increased activation in premotor regions (preSMA) and the dorsal striatum (Forstmann et al., 2008). Other studies have used similar modelling approaches to investigate value-based decisions (Gluth et al., 2012; Hare et al., 2011), for example using a buying task in which participants could decide to accept or reject a stock after receiving probabilistic information about the stock from different rating companies (Gluth et al., 2012). As a main result, Gluth and colleagues showed that the amount of relative evidence that participants required for making a choice was related to the neural response in the anterior insula (AI) and the dorsal striatum. The finding in dorsal striatum resembled evidence from DDM studies obtained with perception paradigms (Forstmann et al., 2008) and indicates that'the striatum is a plausible neural candidate for tracking changes in choice components in different motivational settings (e.g., induced by time pressure, Forstmann et al., 2008, or by others' information, Gluth et al., 2012).

In our study, we modeled the three relevant choice parameters (*v, z*, and *a*) for choices that were driven by the combination of the two motives and for the same choices that were driven by each of the motives separately. It is important to note that the choice process may also be influenced by other motives than empathy and reciprocity, i.e., the motives that were experimentally induced. That said, our paradigm can provide insights into the multi-motive choice process even if other motives are potentially activated, because multi-motive choices are contrasted with the same choices that are driven by the respective single motives.

According to one hypothesis, the simultaneous activation of multiple motives may facilitate the computation of the choice option that is favored by the motives. In the present paradigm this means that computation of the prosocial choice option should be facilitated since empathy and reciprocity both drive prosocial behavior. In this case, we should observe an increase in prosocial behavior in the multi-motive condition (empathy and reciprocity motive active) compared to the single-motive conditions (only empathy or only reciprocity active) that cannot be explained by the difference between the single-motive conditions. Specifying the mechanism underlying such a facilitation, the DDM proposes that a facilitation of prosocial choices in the multi-motive condition may originate A) from an increased speed of information accumulation (*v*-parameter; **Figure 2.3.1C**, left panel (Flagan, Mumford, & Beer, 2017; Janczyk & Lerche, 2019; Krajbich et al., 2015)), B) from an enhancement of participants' initial bias to choose the prosocial option (*z*-parameter; **Figure 2.3.1C**, middle panel; (Chen & Krajbich, 2018; Mulder et al., 2012; Toelch et al., 2018), or C) from an enhancement of the *v*- as well as the *z*-parameter in the multi-motive condition, compared to the single-motive condition.

Alternatively, we may observe fewer prosocial decisions in the multi-motive condition compared to the single motive conditions, reflected by a decreased speed of information accumulation (lower DDM *v*-parameter) and/or decreased initial bias to choose the prosocial option (lower DDM *z*-parameter). Moreover, in the multi-motive condition, participants are required to process two motives simultaneously, in addition to the trial-by-trial choice option information (which was constant across all conditions because participants performed the identical choice task). This additional motive-related information may cause participants to make more careful responses in the multi-motive condition and thus increase the *a*-parameter in the multi-motive condition compared to the single-motive conditions (**Figure 2.3.1C**, right panel).

Regarding the underlying neural mechanisms, we hypothesized that changes in DDM choice parameters in the multi-motive compared to the single-motive conditions might be related to changes of activation in the ventral striatum, i.e., a region that is involved in the integration of different choice values (here the value of empathy-based and the value of reciprocity-based choices; (Kable & Glimcher, 2007; Liljeholm & O ’Doherty, 2012; O’Doherty et al., 2004; Strait et al., 2015), and/ or activation of the dorsal striatum, i.e., a region that is related to integration of choice preferences that derive from these different choice values (Balleine et al., 2007; Liljeholm & O ’Doherty, 2012; O’Doherty et al., 2004; Palmiter, 2008; Robinson et al., 2006).

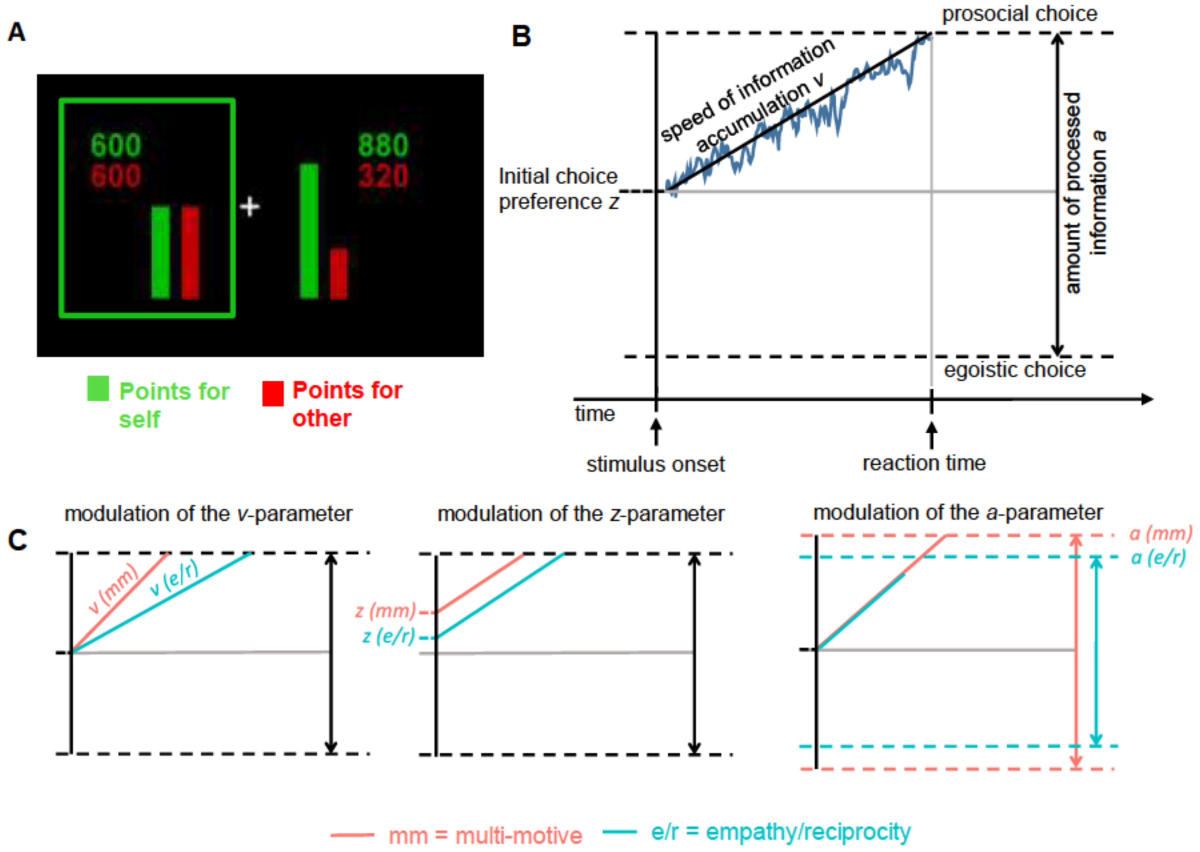

A
600
600
880
320
+
🟩 Points for self
🟥 Points for other
B
prosocial choice
speed of information accumulation v
Initial choice preference z
amount of processed information a
egoistic choice
time
stimulus onset
reaction time
C
modulation of the v-parameter
v (mm)
v (e/r)
modulation of the z-parameter
z (mm)
z (e/r)
modulation of the a-parameter
a (mm)
a (e/r)
mm = multi-motive
e/r = empathy/reciprocity

Figure 2.3.1. Example of point allocation during the choice task, schematic illustration of the drift-diffusion model and hypotheses regarding the impact of different drift-diffusion parameters on the choice process in multi-motive and single-motive conditions. **A** Participants chose between a prosocial and an egoistic option to allocate points to themselves (in this example shown in green) and a partner (in this example shown in red). Colors were counter-balanced across participants. In this example trial, the participant chose the prosocial option, which maximized the outcome of the partner at a cost to the participant (green box). **B** The drift-diffusion model conceptualizes the choice process as noisy accumulation of information (squiggly blue line). The v-parameter describes the speed at which information is accumulated in order to choose one of the options, i.e., the efficiency of the choice process itself. The z-parameter reflects the initial choice bias, i.e., the degree to which an individual prefers one of the choice options prior to making the choice. The third component, parameter a, quantifies the amount of relative evidence that is required to choose one of the options. Once the accumulated information reaches either boundary, the choice is made (upper boundary = prosocial choice; lower boundary = egoistic choice). **C** An enhancement of prosocial choice frequency in the multi-motive condition (red) compared to the single motive conditions (i.e., the empathy or the reciprocity condition; blue) may result from an increased speed of information accumulation (v-parameter; left panel), and/or an increased initial bias toward making a prosocial choice (z-parameter; middle panel). On average, the amount of required relative evidence (a-parameter) may be higher in the multi-motive condition compared to the single motive conditions (right panel).

Material and methods

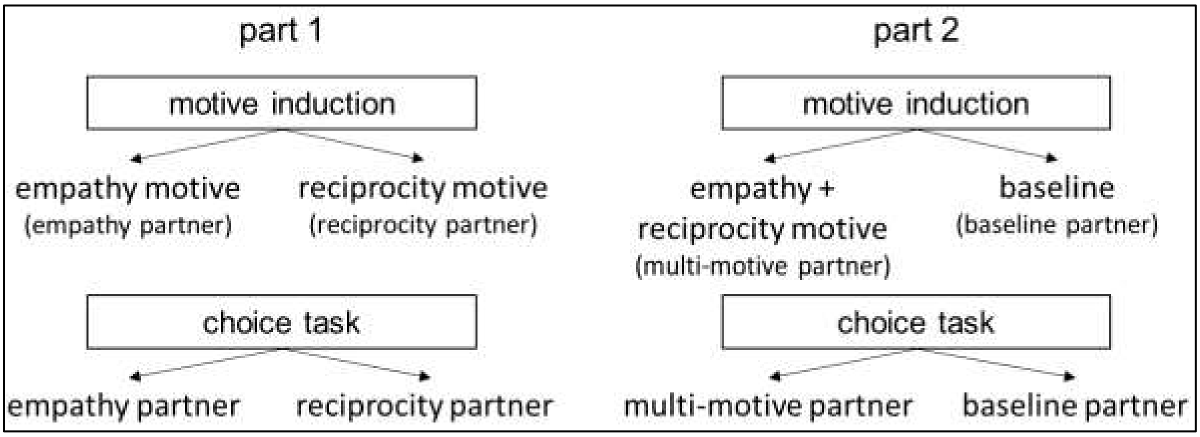
Participant details

Forty-two right-handed healthy female participants (mean age = 23.1 years, SD = 2.8 years) and four female confederates took part in the experiment. We chose female participants as well as female confederates in order to control for gender and avoid cross-gender effects. The confederates were students who had been trained to serve in all the different conditions counterbalanced across participants. Prior to the experiment, written informed consent was obtained from all the participants. The study was approved by the local ethics committee (BB 023/17). Participants received monetary compensation. Three participants had to be excluded due to technical problems and dropout. Another subject had to be excluded due to excessive head movements (more than 5% of the scans contained rapid head motion with more than 0.5 mm displacement per TR). Five participants had to be excluded as outlier based on their choices (less than ten prosocial choices across all condition; three standard deviations above the mean in central measures). Thus, we analyzed 33 data sets using a within-subjects design. We aimed for 34 data sets, which corresponds to the median sample size of neuroimaging studies determined in a recent review (N = 33; Yeung, 2018). We tested 40 participants to meet this target, accounting for a drop-out rate of about 15% which, based on our experience, is common in fMRI studies. Our final analyses includes 33 data sets, in accordance with the median sample reported by Yeung (2018). Given that it is difficult to collect large data sets with expensive and time-consuming methods like fMRI, the importance of stringent statistical thresholds is highlighted (C. S. Carter, Lesh, & Barch, 2016; Roiser et al., 2016; Woo et al., 2014; Yeung, 2018). To analyse the results of the second level regression we thus used cluster-level family wise error correction at the whole brain level after applying a threshold of $P < .001$ on an uncorrected level. Neural activations that are thresholded at this level are seen as valid and reliable (Eklund et al., 2016; Woo et al., 2014; Yeung, 2018).

Procedure

Prior to the motive induction and choice task, the individual thresholds for pain stimulation were determined for the participants and all the confederates (see section *Pain stimulation* for details). Next, the participants and confederates were assigned their different roles by a manipulated lottery (drawing matches). In order to ensure that each participant was always assigned her designated role as a participant (pain recipient during motive induction; decider during the decision task), the drawing of the matches was organized in such a way that she always drew the last match. The confederates were assigned the roles of the empathy partner, the reciprocity partner, the multi-motive partner or the baseline partner, and these roles were counterbalanced across participants. In accordance with these roles, two of the confederates first went to an ostensible other experiment and the other two waited to be seated in the scanner room. Each confederate was matched with a specific color and seating position (to the left vs. to the right of the fMRI scanner), and their color designation and seating positions were counter-balanced across participants.

Next, the first two confederates (the empathy partner, reciprocity partner, multi-motive partner, or baseline partner) were seated to the left and the right of the participant who was lying inside the fMRI scanner and the first motive induction took place (for overview of an example procedure, see Figure 2.3.2). After the motive induction, image acquisition for the choice task was started, during which the participant allocated points to her respective partners. This way the participant only had to remember interactions with two partners at any one time. After the choice task, the first confederates were replaced by the other two confederates and the second part commenced. Part 2 had the same structure as part 1: first, the participant underwent motive induction 2 followed again by the choice task. The order of motive inductions and the type of partner the confederates represented were counterbalanced across participants.



At the end of the experiment, all the confederates left and the participant remained in the scanner until anatomical image acquisition was completed. Finally, participants were asked to complete a questionnaire measuring trait aspects of empathy (IRI, Davis, 1983; Jordan et al., 2016) and reciprocity (Perugini et al., 2003). Participants spent approximately 60 min in the scanner and the entire procedure lasted approximately 2.5 hours. To avoid possible reputation effects, which could influence participants' behavior, participants were informed that they would not meet the confederates after the experiment.

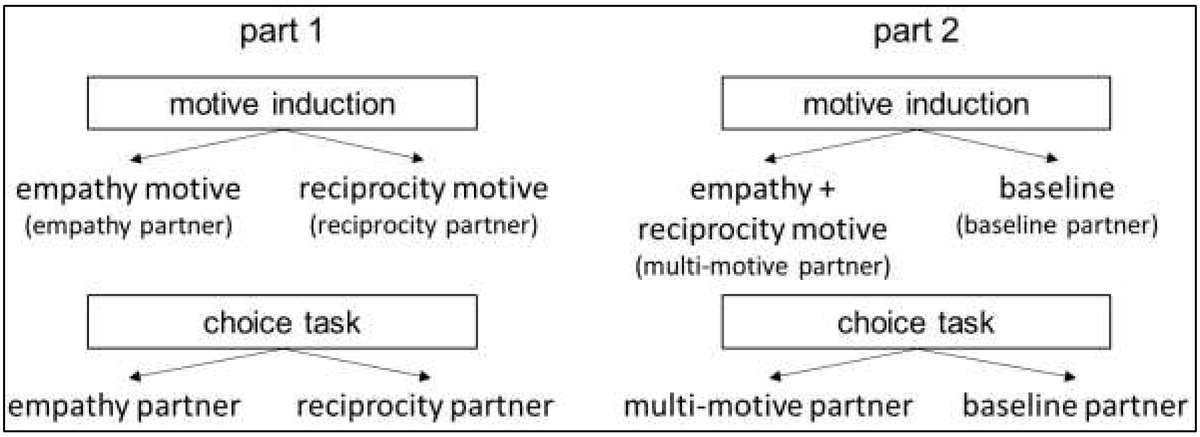


Figure 2.3.2. Overview of an example experimental procedure. The study consisted of two parts. In this example, in part 1, the empathy motive was activated towards one confederate (the empathy partner) and the reciprocity motive was activated towards the other confederate (the reciprocity partner). In the following choice task, participants allocated points to the empathy partner (i.e., driven by the empathy motive) or the reciprocity partner (i.e., driven by the reciprocity motive). Next, the confederates were replaced by two new individuals that served as partners for part 2. In part 2, the empathy and the reciprocity motive were activated simultaneously towards one confederate (multi-motive partner) and no motive was induced towards the other confederate (baseline partner). In the following choice task, participants allocated points towards the multi-motive partner (i.e., driven by two motives simultaneously) and towards the baseline partner (i.e., independently of any motive induction). The order of motive induction (empathy, reciprocity, multi-motive, baseline) was counterbalanced across participants and the four confederates. The respective partner was indicated by a cue in one of four counterbalanced colors.

Motive inductions

Empathy induction

During the study, participants were paired with four partners (confederates of the experimenter). Participants saw the hand of the respective partners with the attached pain electrode. In the empathy condition, the participants repeatedly observed one of the confederates (the empathy partner) receiving painful shocks in a number of trials, a situation known to elicit an empathic response (Batson et al., 1995; Hein, Morishima, et al., 2016; Lamm et al., 2011). Each empathy-induction trial started with a colored arrow shown for 1,000 ms, which indicated the empathy partner. After this cue and a jittered (1,000–2,000 ms) fixation cross, the same colored flash was displayed for 1,500 ms. Participants were informed that a dark-colored flash indicated that the corresponding partner received a painful stimulus at that moment; a light-colored flash indicated a non-painful stimulus. During (ostensible) stimulation of the respective partner, participants either saw a dark colored flash (painful stimulation) or a light colored flash (non-painful stimulation). Since all partners were confederates of the experimenter, they did not actually receive painful stimulations. Thus, the trials in which participants saw the dark colored flash were "ostensibly painful" for the partner. To assess the success of the empathy induction, a rating scale was shown for a maximum of 6 s. Participants reported how they felt after observing the partner receive painful or non-painful stimuli ("How do you feel?" in German). The scale ranged from -4 (labeled "very bad") to +4 (labeled "very good") with intervals of one and was visually displayed. Before analysis, the induction ratings were recoded such that high positive values reflect strong responses to the induction procedure (strong empathy motive). Participants had to respond within 6 s. The inter-trial interval was 1,500 ms. Empathy induction consisted of 12 trials: nine of which were ostensibly painful for their partner (i.e., the confederate).

Reciprocity induction

The reciprocity motive is defined as the desire to reciprocate perceived kindness with kind behavior (Gouldner, 1960; Hein, Morishima, et al., 2016; McCabe et al., 2003). Therefore, in the reciprocity condition, we induced the reciprocity motive by instructing one of the confederates (the reciprocity partner) to give up money in several trials to save the participant from painful shocks (Hein, Morishima, et al., 2016). Each reciprocity-induction trial also started with an arrow colored in the reciprocity partner's color, which pointed toward the seating position of the reciprocity partner (left or right) and was shown for 1,000 ms. Next, the participants were shown a flash displayed to the right and a crossed-out flash displayed to the left of a centered fixation cross. Participants were told that this was the decision screen, which the reciprocity partner also saw while making her decision to either save or not save the participant from painful stimulation. After a jittered interval of 2,000 to 4,000 ms, a box appeared around one of the flashes, indicating the ostensible choice of the reciprocity partner. Depending on where the box was displayed, the reciprocity partner had either decided to forego a monetary award of 2 € in order to save the participant from painful stimulation (a box around the crossed-out flash) or decided to take the money and not save the participant (a box around the flash that was not crossed-out). After 1000ms participants rated how they felt about the decision of the partner ("How do you feel?" in German). The ratings were recoded such that high positive values reflect a strong positive response to the decision of the partner, indicating a strong reciprocity motive.

After a jittered (1,000 to 2,000 ms) fixation cross, the participant saw an information on the screen, indicating whether the decision of the reciprocity partner would be implemented ("decision accepted") or not ("decision declined"), displayed for 1000 ms. This additional stage was included in order to ensure the same number of painful stimulations were administered across all conditions (50 %), while at the same time allowing for the high rate (75 %; 9 out of 12 trials) of the reciprocity partner's decisions to help. While instructing the participants, it was highlighted that the choice of the reciprocity partner reflected her willingness (or unwillingness) to help, while a computer algorithm decided about the implementation of the decision.

Thus, four types of reciprocity trials were possible. When the partner decided to save the participant from painful stimulation and this decision was accepted, the participant did not receive a painful stimulus, which was visually represented by a crossed-out flash (1,500 ms). However, when the reciprocity partner's decision to save the participant was declined, participants received a painful stimulus, which was accompanied by the display of a flash (1,500 ms). Similarly, when the partner decided not to save the participant and this decision was accepted, the participant received a painful stimulus accompanied by the display of a flash. Finally, when the partner decided to not save the participant and this decision was declined, the participant did not receive painful stimulation, which was visually represented by a crossed-out flash. The inter trial fixation cross was displayed for 1,500 ms before the next trial started.

Multi-motive induction

In the multi-motive condition, the participants repeatedly observed how one of the confederates (the multi-motive partner) received painful shocks and also gave up money to spare the participant from painful shocks. The multi-motive induction procedure combined the empathy- and reciprocity-induction procedures. As in the empathy-induction condition, it included 12 empathy induction trials, nine of which were ostensibly painful for the partner. As in the reciprocity-induction condition, it included 12 reciprocity trials, of which participants received help in nine out of 12 trials. The stimulation and trial structure were identical to the empathy- and reciprocity-induction trials described above, except that the relevant colors were replaced by the colors matched to the multi-motive partner (i.e., the color of the pain flash in the empathy trials and the color of the box highlighting the decision of the partner in the reciprocity trials).

Additional control trials for empathy and reciprocity induction

In order to equalize the number and types of trials (i.e., the length and structure of the interaction with each motive partner) across conditions, the empathy-induction procedure also included trials that were identical to the reciprocity trials, except that the computer decided whether the participant would be saved from a painful stimulus and not the empathy partner. This computer's decision was visually represented by a white-colored box appearing either around the crossed-out flash (saving the participant) or the normal flash (not saving the participant). It was clearly explained to each participant that the color white was not matched with any of the partners but indicated the computer's choice. The empathy-induction procedure consisted of 12 control trials, in addition to the 12 empathy trials described above, resulting in 24 trials, i.e., the identical number of trials as the multi-motive induction procedure.

Similarly, the reciprocity-induction procedure included trials that were identical to the empathy-induction trials, except that the reciprocity partner only received non-painful stimulation on these trials, as visually represented by a light-colored flash. In total, the reciprocity-induction procedure consisted of 12 of these control trials and 12 reciprocity trials (see above), i.e., 24 trials (identical to the other conditions).

Baseline induction

The baseline procedure consisted of 24 trials in total, 12 trials in which the baseline partner only received non-painful stimulation and 12 trials in which the computer decided whether the participant would be saved from a painful stimulus or not. This computer's decision was visually represented by a white-colored box either appearing around the crossed-out flash (saving the participant) or the normal flash (not saving the participant). It was clearly explained to the participant that the white box did not represent the decision of a person but indicated the computer's choice.

Choice task

After the motive inductions, participants performed a social choice task inside the fMRI scanner. The choice task was a two-alternative-forced-choice adaptation of the commonly used Dictator Game (Forsythe et al., 1994), which has been successfully used in previous studies (e.g., Chen & Krajbich, 2018; Hein, Morishima, et al., 2016; Krajbich et al., 2015). In each trial of this choice task, participants allocated money to themselves and one of the partners (**Figure 2.3.1A**) and could choose between maximizing the relative outcome of the other person by reducing their own relative outcome (prosocial choice) and maximizing their own relative outcome at a cost to the partner (egoistic choice). The outcome was relative to the outcome that the participant would have gained when choosing the other option. The initial number of points was always higher for the participant compared to the partners. This measure was inspired by previous behavioral economics research, showing that prosocial behaviors depend on the initial payoff allocation between the participant and the participant's partner (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Schmidt, 1999). In particular, if subjects have a lower initial payoff than their partner ("disadvantageous initial inequality"), they are much less willing to behave altruistically toward the partner compared to a situation with advantageous initial inequality (i.e., when the participant has a higher initial payoff than the partner). The choice options used in the present study created advantageous inequality to optimize the number of prosocial choices, which was the main focus of our study. The exact point distributions are provided in supplementary **Table A1**.

Depending on the type of partner the participants faced in the choice task, there were four conditions – the empathy condition, the reciprocity condition, the multi-motive condition, and the baseline condition. Importantly, the choice task was identical in all the conditions.

In more detail, participants were asked repeatedly to choose between two different distributions of points that each represented different amounts of monetary pay-offs for themselves and one of the partners (see **Figure 2.3.1A**). Each choice-trial started with a colored arrow shown for 1,000 ms, indicating the next interaction partner. After this cue, participants saw the two possible distributions of points in different colors, indicating the potential gain for the participant or for the current partner. Colors were counterbalanced across participants. Participants had to choose one of the distributions within 4,000 ms. A green box appeared around the distribution that was selected by the participant at 4,000 ms after distribution onset. The box was shown for 1,000 ms. The length of the inter-trial interval, as indicated by a fixation cross, was jittered between 4,000 and 6,000 ms. At the end of the experiment, two of the distributions chosen by the participant were randomly selected for payment (100 points = 50 cents). We analyzed 38 choice trials in each motive-induction condition, i.e., 152 trials in total. In addition to the 38 trials, each condition contained four trials in which the same choice option maximized the outcome of the participant and the partner (non-competitive trials). These trials were included to increase the variability of choices and thus to keep the participants engaged. They were excluded from the analyses because they could not be classified as prosocial or egoistic choice trial. For each condition and participant, the same distributions were used and presented in random order.

Pain stimulation

For pain stimulation, we used a mechano-tactile stimulus generated by a small plastic cylinder (513 g). The projectile was shot against the cuticle of the left index finger using air pressure (Impact Stimulator, Labortechnik Franken, Release 1.0.0.34). The criterion for painful stimulation was a subjective value of 8 on a pain scale ranging from 1 (no pain at all, but a participant could feel a slight touch of the projectile) to 10 (extreme, hardly bearable pain). The participants were told that a value of 8 corresponded to a painful, but bearable stimulus, and a non-painful stimulus corresponded to a value of 1 on the same subjective pain scale. These subjective pain thresholds were determined using a stepwise increase of air pressure (stepsize of 0.25 mg/s), starting with the lowest possible pressure (0.25 mg/s), which caused the projectile to barely touch the cuticle, and increasing in stimulus intensity until it reached a level that corresponded to the individual's value of 8 (range = 2.75–3.5 mg/s).

Experimental design and statistical analyses

Regression analyses

Regression analyses were conducted using the R-packages "lme4 and "car" (R Core-Team, 2018). For mixed models, we report the chi-square values derived from Wald chisquare tests using the "Anova" (car package) function. For predefined contrasts we report the t-values derived from the summary() function. When more than one predictor was included in the model, the function emmeans was used in order to compute contrasts between factor levels.

To test the differences in induction ratings and the relationship between induction ratings and frequencies of prosocial choice, the mean induction ratings and frequencies of prosocial choices were calculated for each participant for each condition (empathy, reciprocity, multi-motive, and baseline) and entered as a dependent variable into mixed models with conditions (empathy, reciprocity, multi-motive, and baseline) and induction ratings as fixed effects and participants as random effects. Additionally, in order to probe the specificity of the induction procedure, we tested whether trait empathy (empathic concern subscale of the IRI, Davis (1983)) and trait reciprocity (PNR, (Perugini et al., 2003)) differentially influenced choice behavior in the three motive conditions (empathy, reciprocity, multi-motive). Specifically, we conducted a linear mixed model regression with the frequency of prosocial choices as dependent variable, trait measure type (empathy/reciprocity), individual trait measure scores (empathy/reciprocity), motive induction condition (empathy, reciprocity, multi-motive), and their interactions as fixed effects, and participant as random intercept.

To test whether prosocial behavior was influenced by trial-by-trial point information, condition and their interaction, a logistic mixed model regression was conducted with the possible gain for the partner (i.e. difference in points between the two options for the partner, |partner’s gain option 1 – partner’s gain option 2|), the possible loss for the participant (i.e., the difference in points between the two options for the participant, |participant’s gain option 1 – participant’s gain option 2|) and condition as predictor variables. The binary choice outcome (prosocial vs. egoistic choice) was used as dependent variable. To investigate the differences in prosocial behavior between the social motives (multi-motive > reciprocity and multi-motive > empathy) more closely, contrasts were calculated using the *emmeans* function.

To specifically test whether prosocial behavior was differentially influenced by inequity aversion, a logistic mixed model regression was conducted with the predictor variables condition and the difference in point equality of the participant’s and the partner’s outcome between the two choice options. To compute this variable, we first calculated calculating the difference between the gains for each option (i.e., |partner’s gain option 1 – participant’s gain option 1| for each choice option). Second, these differences were subtracted from each other in order to obtain a measure of point equality for each choice trial. Again, the binary choice outcome (prosocial vs. egoistic choice) was used as dependent variable.

To test whether the frequency of prosocial choices and reaction times were equally distributed across conditions, we conducted pairwise Kolmogorov-Smirnov tests.

Additionally, we investigated whether the relationship between the possible gain for the partner and participants’ probability to make a prosocial choice can be described in terms of a psychometric function. For the estimation of the psychometric functions we used the R-package “quickpsy” which implements a Maximum-Likelihood-Estimation procedure to fit the cumulative normal distribution. To test whether the points of subjective equality (PSEs) differed between conditions, we conducted a linear mixed model with the condition as fixed effect, participant as random effect and PSE as dependent variable.

To test whether the relative difference between empathy and reciprocity in the *z*-parameter and *a*-parameter could explain the percent changes of these parameters in the multi-motive condition compared to the reciprocity condition, the percent change values ($\Delta z_{\text{multi-motive/reciprocity}}$ and $\Delta a_{\text{multi-motive/reciprocity}}$) were entered as dependent variables in a linear regression model. The respective relative differences ($\Delta z_{\text{empathy/reciprocity}}$ and $\Delta a_{\text{empathy/reciprocity}}$) and one regressor modeling the parameter type (*z*-parameter, *a*-parameter) were included as predictors.

Drift-diffusion modeling

We used hierarchical drift-diffusion modeling (HDDM) (Vandekerckhove et al., 2011; Wiecki et al., 2013), which is a version of the classical drift-diffusion model that exploits between-subject and within-subject variability using Bayesian parameter estimation methods, because it is ideal for use with relatively small sample sizes. The analyses were conducted using the python implementation of HDDM version 0.8.0 (Wiecki et al., 2013).

Based on binary choices, the HDDM approach provides detailed insights into the computation of egoistic and prosocial choices, because it uses all the raw data that is available (trial-by-trial reaction times and choice outcome information of all choices, irrespective of point distributions) to estimate sub-components of the underlying decision process. The v, z and a-parameters for each participant capture how each person manoeuvers between the egoistic and the prosocial choice options, and finally approaches a decision boundary (i.e., the boundary for an egoistic or a prosocial choice). In line with previous studies that have used a similar procedure in the realm of social decision making (e.g., Chen and Krajbich, 2018; Gallotti and Grujić, 2019), we believe that the HDDM results provide a sensitive and fine-grained proxy for individual differences in prosociality.

Since we did not have prior hypotheses about which and how many of the three central DDM parameters may reflect motive complexity, we estimated 11 possible variants of the DDM model, ranging from the most simple model (no parameter is modulated by condition) to the full model with *v*, *z*, and *a* possibly being modulated by our four conditions (baseline, empathy, reciprocity, and multi-motive). Since the point information varied between trials, which may influence drift-rate, we allowed the drift rate to vary by the trial-by-trial possible gain for the partner (see section *Regression analyses* for computation of this value). We performed model comparison based on the deviance information criterion (DIC) and extracted the parameters of the winning model (lowest DIC value). Apart from the three parameters of interest for our research question (*v, z, a*), additional parameters are included in the estimation procedure. We also estimated the non-decision time t and allowed for trial-by-trial variations of the initial bias (sz), he drift rate (sv) and the non-decision time (st). These parameters were not estimated to vary by condition. They were nonetheless included based on the results by Lerche and Voss (2016), who showed that in most cases, it is beneficial to include these parameters in order to improve model fit. In the estimation procedures we used the default values for the priors and hyperpriors provided by the HDDM package. In more detail, the "informative group mean priors are created to roughly match parameter values reported in the literature and collected by Matzke and Wagenmakers (2009)." cited from (Wiecki et al., 2013, Supplementary Material, page 1). Model convergence was checked by visual inspection of the estimation chain of the posteriors, as well as computing the Gelman-Rubin Geweke statistic for convergence (all values < 1.01) (Gelman & Rubin, 1992). To assess model fit, we conducted posterior predictive checks by comparing the observed data with 500 datasets simulated by our model (Wiecki et al., 2013). This approach allows for the computation of intervals within which the parameter falls with 95 % probability. If the observed data falls within the 95 % credibility interval of the simulated data, the model can describe the data well. The present results revealed a good match between the observed data and the modelled data. Parameters of interest from the winning model were extracted for further analysis. Specifically, for each participant, the condition-specific *v*-parameters, *z*-parameters, and *a*-parameters were extracted (resulting in 12 parameters per participant). In HDDM, the *z*-parameter is always relative to *a*. The reported values of *z* thus range between 0 and 1 and correspond to the absolute value of *z* divided by the a-parameter (*z/a*)

For closer investigation of processing differences in complex vs. more simple motivational states, we compared the posterior distributions of the conditions for each parameter by computing the probabilities for the multi-motive parameter being larger than the single motive parameters. This was done by calculating the densities of the differences distributions that are larger than zero (Wiecki et al., 2013). Additionally, we used the plausible value approach to estimate the corresponding t-value. This approach consists of repeatedly sampling participants' individual parameters from the winning model's posterior distribution. Extracting these parameters and comparing between the different conditions using frequentist statistics results in distributions of t-values whose means are a plausible proxy for the actual underlying t-value (Ly et al., 2017; Marsman, Maris, Bechger, & Glas, 2016).

fMRI data acquisition

Imaging data was collected at a 3T MRI-scanner (Verio, Siemens, Erlangen, Germany) with a 32-channel head coil. Functional imaging was performed with a multiband EPI sequence of

72 transversal slices oriented along the subjects’ AC-PC plane (multi-band acceleration factor of 6). The in plane resolution was 2.5 x 2.5 mm² and the slice thickness was 2.5 mm. The field of view was 210 x 210 mm², corresponding to an acquisition matrix of 84 x 84. The repetition time was 1 s, the echo time was 33.6 ms, and the flip angle was 54°. Structural imaging was conducted using a sagittal T1-weighted 3D MPRAGE with 176 slices, and a spatial resolution of 1 x 1 x 1 mm³. The field of view was 250 x 250 mm², corresponding to an acquisition matrix of 256 x 256. The repetition time was 1,690 ms, the echo time was 2.52 ms, the total acquisition time was 3:50 min, and the flip angle was 9°. For the T1-weighted images, GRAPPA with a PAT factor of 2 was used. We obtained, on average, 1,911 (SD = 5.6 volumes) EPI-volumes during the choice task of each participant. We used a rubber foam head restraint to avoid head movements.

fMRI Preprocessing

Preprocessing and statistical parametric mapping were performed with SPM12 (Wellcome Department of Neuroscience, London, UK) and Matlab version 9.2 (MathWorks Inc; Natick, MA). Spatial preprocessing included realignment to the first scan, and unwarping and coregistration to the T1 anatomical volume images. Unwarping of geometrically distorted EPIs was performed using the FieldMap Toolbox. T1-weighted images were segmented to localize grey and white matter, and cerebro-spinal fluid. This segmentation was the basis for the creation of a DARTEL Template and spatial normalization to Montreal Neurological Institute (MNI) space, including smoothing with a 6 mm (full width at half maximum) Gaussian Kernel filter to improve the signal-to-noise-ratio. To correct for low-frequency components, a high-pass filter with a cut-off of 128 s was used.

fMRI statistical analysis

Our participants made prosocial choices in the majority of the trials (Mean = 74 %, SD = 19 %) with more than half of the participants making prosocial choices in 80 % or more of the trials in at least one of the four conditions (see **Table A2**). Given the lack of egoistic choices and given that our study focused on the computation of prosocial choices, egoistic choices trials were not included in the imaging analyses and we also refrained from computing direct contrasts between prosocial and egoistic choices.

First-level analyses were performed with the general linear model (GLM), using a canonical hemodynamic response function (HRF) and its first derivative (time derivative). Regressors were defined from cue onset until the individual response was made by pressing a button (resulting in a time window of 1,000 ms + individual response time). For each of the four conditions (the three motive conditions and baseline condition), the respective regressors of prosocial choice trials were included as regressors of interest. The respective regressors of all the other trials (e.g., egoistic choice trials and trials with missed button presses) were included as regressors of no interest. Given that more than half of our participants (64 %) made fewer than five egoistic decisions in at least one of the conditions, we refrained from computing direct contrasts between prosocial and egoistic choices and included egoistic choices in this regressor of no interest (see **Table A2** for the number of trials per participant and condition). The residual effects of head motions were corrected for by including the six estimated motion parameters for each participant and each session as regressors of no interest. To allow for modeling all the conditions in one GLM, an additional regressor of no interest was included, which modeled the potential effects of session.

For the second-level analyses, contrast images for comparisons of interest (empathy > reciprocity, multi-motive > empathy, reciprocity > empathy, and multi-motive > reciprocity) were initially computed on a single-subject level. In the next step, the individual images of the main contrast of interest (multi-motive > reciprocity) were regressed against the percent change in the *z*-parameter ($\Delta z_{\text{multi-motive/reciprocity}}$) and *a*-parameter ($\Delta a_{\text{multi-motive/reciprocity}}$) in the multi-motive condition, relative to the reciprocity condition, using second-level regressions. Second-level results were corrected for multiple comparisons, using cluster-level family wise error (FWE) correction on a whole brain level. We also report results at a threshold of $P_{\text{uncorrected}} < 0.001$ and a cluster threshold of $k > 10$ in the supplementary online material.

To test if the neural response in the dorsal striatum was related to the relative difference in *z* between empathy and reciprocity ($\Delta z_{\text{empathy/reciprocity}}$), the (multi-motive > reciprocity) contrast was regressed against the empathy vs reciprocity *z*-differences ($\Delta z_{\text{empathy/reciprocity}}$) and the multi-motive z-enhancement ($\Delta z_{\text{multi-motive/reciprocity}}$) in the same model. Additionally, the individual beta-estimates of the neural multi-motive condition > reciprocity and empathy > reciprocity contrasts were extracted from an independent anatomical ROI of bilateral putamen based on the aal nomenclature (Tzourio-Mazoyer et al., 2002), using MarsBaR (M

Brett, Anton, Valabregue, & Poline, 2002) and the WFU PickAtlas Tool (Maldjian, Laurienti, Kraft, & Burdette, 2003).

In order to clarify the commonly shared influence of the partner's possible gain on the neural prosocial choice process, we added the partner's gain as trial-by-trial parametric modulator of the decision phase to a first level GLM in which all conditions are collapsed into one single regressor. On the second level, we conducted a one sample t-test on this parametric modulator corrected for multiple comparisons, using cluster-level family wise error (FWE) correction on a whole brain level.

The reported anatomical regions were identified using the SPM anatomy toolbox (Eickhoff et al., 2005).

Data and code availability

Behavioral data and scripts are available at github.com (https://github.com/AnneSaulin/complex_motivations).

Imaging data are available at neurovault.org (https://www.neurovault.org/collections/5879/).

Results

Motive induction

During the empathy induction, participants indicated how they felt after observing the person in pain. During the reciprocity induction, they indicated how they felt after receiving a favor from the other person. In the multi-motive condition, participants provided both of these ratings. Strong empathy is indicated by negative feelings when seeing the partner in pain, indicated by negative ratings. Strong reciprocity is indicated by positive feelings when observing the decision of the partner, indicated by positive ratings. To allow the comparison of the ratings in all conditions, empathy ratings were recoded such that positive ratings now reflect strong empathy, i.e., multiplied by -1. The results of linear mixed models (lmms) showed that the induction ratings in the motive conditions were significantly higher than those in the baseline condition (χ^2 = 515.15, P < .000001, β = 1.61, SE = 0.071, $\text{rating}_{\text{baseline}}$= -1.02 ± 1.00, $\text{rating}_{\text{empathy}}$= 1.57 ± 0.77, $\text{rating}_{\text{reciprocity}}$= 1.50 ± 0.89, $\text{rating}_{\text{multi-motive}}$ = 1.54 ± 0.91, (*M* ± *SEM*)). There were no significant differences in the induction ratings between the motive conditions (χ^2 = 0.14, P = .93, $\beta_{\text{reciprocity}}$ = -0.07, SE = 0.20, $\beta_{\text{multi-motive}}$ = -0.02, SE =

0.17). The induction ratings in the motive conditions were significantly associated with the frequency of prosocial choices (χ^2 = 6.38, P = .01). This effect held to a comparable extent across all three motive conditions (motive condition × rating interaction, χ^2 = 3.61, P = .16, see **Table A3** for full results). Specifically, the two single-motive conditions yielded similar induction ratings (χ^2 = 0.23, P = .64, $\beta_{\text{reciprocity}}$ = -0.07, SE = 0.15) and had a comparable effect on the frequency of prosocial choices (χ^2 = 4.77, P = .03, condition × rating interaction, χ^2 = 2.06, P = .15, see **Table A4** for full results). These results show that the strength of motive induction and the link to prosocial choices was comparable for the empathy and the reciprocity motives (**Figure A1**). Further, supporting that the induction procedure specifically influenced empathy and reciprocity motivations, trait empathy and trait reciprocity differentially influenced the frequency of prosocial behavior in the three motive conditions (trait measure type × trait measure value × motive condition interaction, χ^2 = 6.08, P = .047, see **Table A5** for full results).

Frequency of prosocial choices

Pairwise Kolmogorov-Smirnov tests revealed that the frequency of prosocial decisions was comparably distributed across conditions (all Ds < 0.24, all Ps > 0.29, for detailed statistics, please see **Table A6**) as were the reaction times (all Ds < 0.27, all Ps > .17, **Figure 2.3.3C**).

The frequency of prosocial choice was significantly influenced by condition (χ^2 = 56.99, P < .0001, see **Figure 2.3.3A** and **Figure A2**, $\text{prosoc}_{\text{baseline}}$= 67.7 ± 20.6 %, $\text{prosoc}_{\text{empathy}}$= 77.0 ± 16.8, $\text{prosoc}_{\text{reciprocity}}$= 73.1 ± 18.4, $\text{prosoc}_{\text{multi-motive}}$= 77.4 ± 18.0 (*M* ± *SEM*), see **Table A7** for full results), indicating that the motive inductions had a differential effect on later prosocial choices. Moreover, prosocial choices were influenced by the possible gain for the partner (χ^2 = 668.64, P < .0001). However, model comparison revealed that neither including the possible gain for the participant (χ^2 = .23, P = .63), nor its interaction with condition significantly improved the model fit (χ^2 = .86, P = .84). Thus, for the analyses reported below condition and possible gain for the partner were used as additional predictors.

The frequency of empathy-based prosocial choices was increased compared to reciprocity-based choices (z ratio = 2.94, P = .02), whereas the frequency of prosocial choices between the multi-motive condition and the empathy condition was comparable (z-ratio = .56, P =

.94). However, the multi-motive condition yielded significantly more prosocial choices compared to the reciprocity condition (z-ratio = 3.49, P = .003).

To clarify this effect, we calculated the percent change in prosocial choices in the multi-motive condition relative to each single motive condition

$$\Delta prosoc_{\text{multi-motive/reciprocity}} = \frac{prosoc_{\text{multi-motive}} - prosoc_{\text{reciprocity}}}{prosoc_{\text{reciprocity}}} \times 100$$

$$\Delta prosoc_{\text{multi-motive/empathy}} = \frac{prosoc_{\text{multi-motive}} - prosoc_{\text{empathy}}}{prosoc_{\text{empathy}}} \times 100$$

where $prosoc_{\text{multi-motive}}$ equals the frequency of the prosocial choices in the multi-motive condition, $prosoc_{\text{reciprocity}}$ equals the frequency of prosocial choices in the reciprocity condition, and $prosoc_{\text{empathy}}$ equals the frequency of prosocial choices in the empathy condition.

The percent change of the multi-motive condition relative to reciprocity was significantly positive (T(32) = 2.07, P = .047, $\Delta prosoc_{\text{multi-motive/reciprocity}}$ = 8.61 ± 4.17 (*M* ± *SEM*)), demonstrating that prosocial choices were enhanced when reciprocity was combined with empathy, relative to reciprocity alone. The percent change in the multi-motive condition relative to the empathy condition was not significantly different from zero (T*(*32) = 0.42, P = .674, $\Delta prosoc_{\text{multi-motive/empathy}}$= 1.05 ± 2.47 (*M* ± *SEM*)), indicating that the simultaneous activation of the reciprocity motive did not enhance the empathy motive.

Reaction times

Reaction times were significantly influenced by conditions (χ^2 = 27.89, P < .0001, see **Table A8** for full results). That is, participants were faster in the motive conditions compared to the baseline condition (baseline vs. empathy: T(32) = 5.03, P < .0001, baseline vs. reciprocity: T(32) = 3.62, P = .002, baseline vs. multi-motive: T(32) = 3.70, P = .001). There were no differences in reaction times for prosocial choices between the motive conditions (all Ps > .49) (**Figure 2.3.3B**), and the reaction time distributions were comparable (**Figure 2.3.3C**).

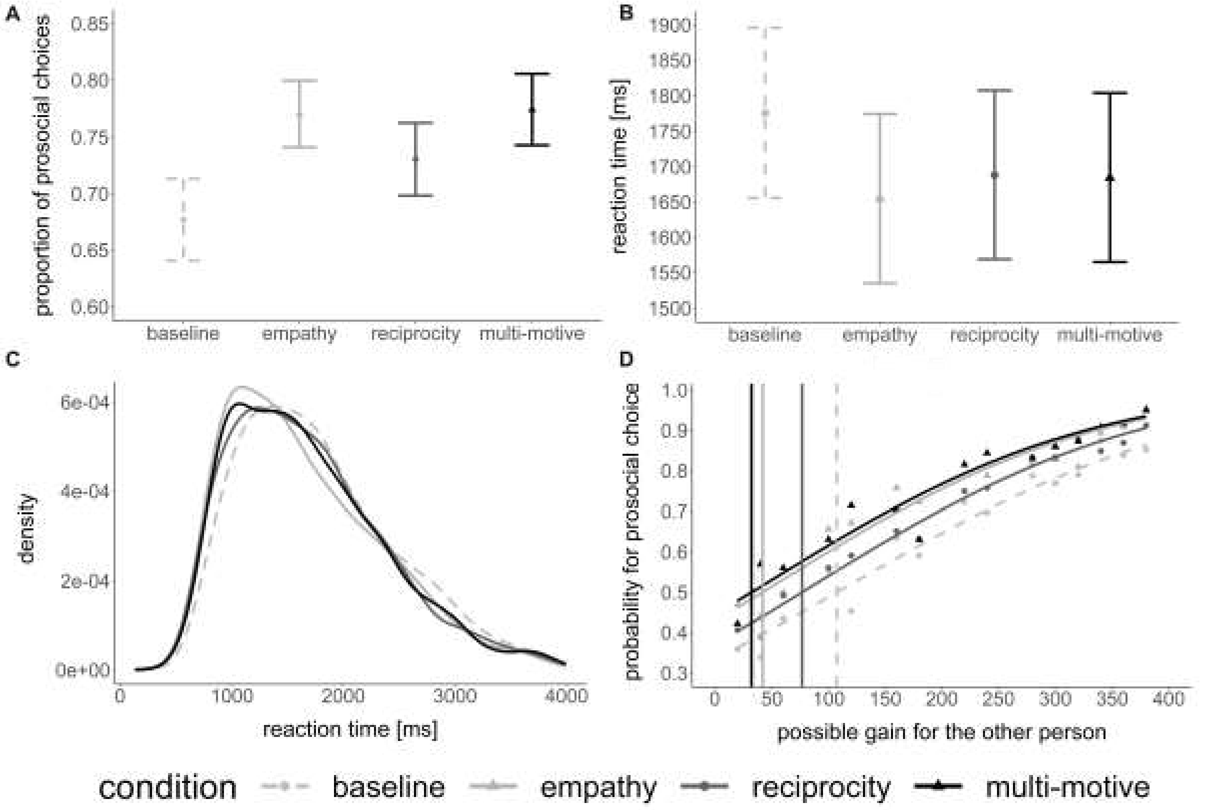


Figure 2.3.3. Descriptive statistics, distributions and psychometric function of the choice and reaction time data. **A** Mean proportion of prosocial choices per condition. Error bars denote standard errors of means. **B** Mean reaction times per condition. Error bars denote standard errors of means. **C** Distribution of reaction times across participants per condition. **D** Psychometric functions for the different conditions of the probability to make a prosocial choice depending on the amount of points the participants' partner could possibly gain in each trial, that is, the point value for the partner in case of a prosocial choice minus the point value for the partner in case of an egoistic choice. Vertical dashed lines indicate the points of subjective equality in the different conditions (for exact values and spread, see **Table A9**). Please note that the probability of making a prosocial decision never reached 0 because of the high frequency of prosocial choices in our data (participants made prosocial choices even if the gain for the other person was low). Thus, our results do not yield data points much lower than the respective points of subjective equality (PSEs).

Point equality and prosocial behavior

In a next step, we tested whether considerations of equity differentially influenced participants' prosocial behavior in the different conditions.

We considered the difference in point equality of the participant's and the partner's outcome between the two choice options to test whether inequity aversion differentially influenced participants' prosocial choice behavior. The results showed a main effect of difference in point equality (χ^2 = 65.87, P < .0001) and a main effect of condition (χ^2 = 46.91, P < .0001). However, no interaction effect was observed (χ^2 = 0.19, P = .98, see **Table A10** for full results). Based on these results we conclude that inequity aversion does not differentially affect the different motive conditions and thus cannot explain behavioral differences in the different conditions.

In line with these results, the relationship between the frequency of prosocial choices and the partner's possible gain was comparable between the different motive conditions, as reflected by comparable values for the points of subjective equality based on the psychometric functions estimated for the different conditions (χ^2 = 2.89, P = .41, **Figure 2.3.3D** and **Table A9**).

Since we were most interested in the underlying prosocial decision processes in more complex as compared to simpler motivational states, we used hierarchical drift-diffusion modeling (HDDM) (Vandekerckhove et al., 2011; Wiecki et al., 2013) to understand prosocial choice behavior in the multi-motive condition relative to the reciprocity condition and relative to the empathy condition.

Hierarchical drift-diffusion modeling

We estimated the three aforementioned DDM parameters (*v, z, a*) for every condition and participant. We also estimated the non-decision time *t* (0.94 ± 0.04 (*M* ± *SEM*)). However, this parameter was not estimated to vary by condition and was thus not further analyzed. Comparing the observed data with 500 datasets simulated by our model (Wiecki et al., 2013) showed that the winning model fit the data with 95% credibility (see **Table A11** for overview of all models and DIC values, and see **Table A12** for quantile comparison and 95% credibility). Based on the hypotheses depicted in **Figure 2.3.1C**, we tested whether the observed percent change in the multi-motive condition can be explained by an increase in the speed of information accumulation (*v*-parameter, **Figure 2.3.1C**, left panel), and/or an increase in initial prosocial bias (*z*-parameter, **Figure 2.3.1C**, middle panel). Additionally, we tested whether the induction of both motives enhanced the amount of relative evidence that participants required during the choice process, relative to the two single-motive conditions (*a*-parameter, **Figure 2.3.1C**, right panel).

Testing the first hypothesis (**Figure 2.3.1C**, left panel), we observed no significant differences between the motive conditions ($p(v_{multi\text{-}motive} > v_{empathy}) = 46.38\ \%$, plausible T(32) = -.31, P =

.76; $p(v_{multi\text{-}motive}>v_{reciprocity}) = 72.25$ *%,* plausible T(32) = 1.03, P = .31*,* see **Figure A4** for distribution of t-values). Further, there was a slight percent change in *v*-parameters in the multi-motive condition relative to the reciprocity condition ($\Delta v_{\text{multi-motive/reciprocity}} = \frac{v_{\text{multi-motive}} - v_{\text{reciprocity}}}{v_{\text{reciprocity}}} \times 100$ = -1.54 ± 0.92 % (*M* ± *SEM*), *T(*32) = -1.71, P = .09) and no percent change relative to the empathy condition ($\Delta v_{\text{multi-motive/empathy}} = \frac{v_{\text{multi-motive}} - v_{\text{empathy}}}{v_{\text{empathy}}} \times 100$ = -0.13 ± 0.83 % (*M* ± *SEM*), T(32) = -0.16, P = .88). This result showed that the speed of information accumulation, i.e., the efficiency of the choice process itself, was mainly unaffected by the combination of the two motives, relative to the single-motive conditions.

Testing the second hypothesis (**Figure 2.3.1C**, middle panel), we observed an increase in initial prosocial bias (*z*-parameter) in the multi-motive condition compared to the reciprocity condition ($p(z_{multi\text{-}motive}>z_{reciprocity}) = 93.55$ *%,* plausible T(32) = 3.66, P < .001) (**Figure 2.3.4A**), but not compared to the empathy condition ($p(z_{multi\text{-}motive}>z_{empathy}) = 70.33$ *%*, plausible T(32) = 1.67, P = .10). The percent change in the *z*- parameter of the multi-motive condition was significantly positive relative to the reciprocity condition ($z_{\text{multi-motive/reciprocity}} = \frac{z_{\text{mulit-motive}} - z_{\text{reciprocity}}}{z_{\text{reciprocity}}} \times 100$ = 6.40 ± 1.21 % (*M* ± *SEM*), (*T(*32) = 5.36, P < .001) and marginally larger than zero relative to the empathy condition ($\Delta z_{\text{multi-motive/empathy}} = \frac{z_{\text{multi-motive}} - z_{\text{empathy}}}{z_{\text{empathy}}} \times 100$ = 2.54 ± 1.39 % (*M* ± *SEM*), (T(32) = 1.85, P = .07).

In addition, we had hypothesized that the combination of the two motives may increase the amount of relative evidence that participants required in order to reach a decision (captured by the *a*-parameter; **Figure 2.3.1C**, right panel). The *a*-parameter was not significantly higher in the multi-motive condition compared to the reciprocity condition ($p(a_{multi\text{-}motive}>a_{reciprocity}) =$ 84.70 %, plausible T(32) = 1.73, P = .09) and the empathy condition ($p(a_{multi\text{-}motive}>a_{empathy}) =$ 82.35 %, plausible T(32) = 1.43, P = .16). However, there was a significantly positive relative percent change in *a*-parameters in the multi-motive condition relative to the reciprocity condition ($\Delta a_{\text{multi-motive/reciprocity}} = \frac{a_{\text{multi-motive}} - a_{\text{reciprocity}}}{a_{\text{reciprocity}}} \times 100$ = 9.77 ± 4.36 % (*M* ± *SEM*), *T(*32) = 2.28, P = .03) and also relative to the empathy condition ($\Delta a_{\text{multi-motive/empathy}} = \frac{a_{\text{multi-motive}} - a_{\text{empathy}}}{a_{\text{empathy}}} \times 100$ = 9.57± 4.63 % (*M* ± *SEM*), *T(*32) = 2.10, P = .04).

Taken together, the DDM results showed that the combination of the two motives enhanced participants' bias for choosing the prosocial option, relative to the initial prosocial choice bias biases induced by the reciprocity motive (captured by the percent change in the *z*-parameter). The combination of empathy and reciprocity also led to a relative increase in the amount of relative evidence that people required to make a choice relative to the reciprocity motive, and also relative to empathy (captured by the percent change in the *a*-parameter). In contrast, the speed of information accumulation, i.e., the efficiency of the choice process itself, was comparable between multi-motive and single-motive conditions (no change in *v*-parameter).

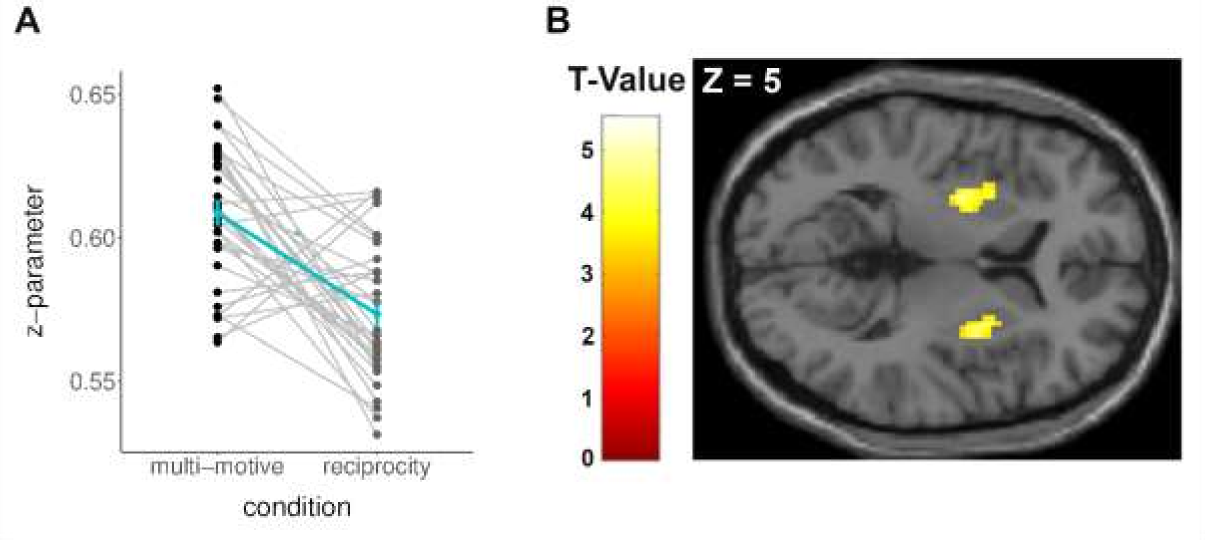


Figure 2.3.4. Increase in initial prosocial bias in the multi-motive condition relative to the reciprocity condition and related neural activity. **A** Initial prosocial bias (z-parameter) were significantly stronger in the multi-motive compared to the reciprocity condition (plausible T(32) = 3.66, P < .001). Individual values are depicted for the multi-motive condition (red) and the reciprocity condition (blue). Means and standard errors of the mean are depicted in black. **B** The individual changes of initial prosocial choice biases in the multi-motive condition relative to the reciprocity condition were tracked by an increase in neural responses in the bilateral dorsal striatum (P(whole-brain FWE$_{\text{cluster-corrected}}$) = .001; MNI peak coordinates; right hemisphere: x = 30, y = 2, z = -2, left hemisphere: x = -28, y = -9, z = 1; visualized at P < .001 uncorrected; **Table A14** and **Figure A6**).

These results may indicate that the observed percent changes in the multi-motive condition relative to the reciprocity condition (in the *z*- and the *a*-parameters) originate from the simultaneous activation of the two motives in the multi-motive condition. Alternatively, as we observed no significant difference between the multi-motive condition and the empathy-condition, it is also conceivable that the empathy motive replaced the reciprocity motive when the two motives were activated simultaneously. In this case, the observed percent changes in the multi-motive condition would reflect the dominance of empathy over reciprocity, instead of a multi-motive effect. If in fact empathy replaced the co-activated reciprocity motive, the relative difference in the *z*-parameters and *a*-parameters between the empathy and the reciprocity conditions should predict the individual extent of the percent changes in the multi-motive condition relative to the reciprocity condition. To test this explanation, we calculated the relative differences in the *z*-parameters and *a*-parameters between empathy and reciprocity ($\Delta z_{\text{empathy/reciprocity}} = \frac{z_{\text{empathy}} - z_{\text{reciprocity}}}{z_{\text{reciprocity}}} \times 100$ and $\Delta a_{\text{empathy/reciprocity}} = \frac{a_{\text{empathy}} - a_{\text{reciprocity}}}{a_{\text{reciprocity}}} \times 100$), entered them as predictors in a regression analysis, and tested their effects on the observed percent changes in the multi-motive condition ($\Delta z_{\text{multi-motive/reciprocity}}$; $\Delta a_{\text{multi-motive/reciprocity}}$). This analysis revealed no significant effects (β = -0.20, SE = 0.255, P = .42; interaction with parameter type (*z* vs *a*): β = .002, SE = 0.38, P = 1.00, main effect of empathy dominance: β = 0.08, SE = 0.136, P = .59). These results demonstrate that the difference between the two single motives cannot account for the changes in choice parameters in the multi-motive condition relative to the reciprocity condition, bolstering the claim that the observed effects are driven by the simultaneous activation of the two motives. The three DDM parameters of interest for each condition and the relative differences between the baseline condition and the motive conditions are provided in **Figure A3** and **Table A13**.

Imaging results

Next, we investigated the neural underpinnings of the prosocial decision process comparing the multi-motive and the single motive conditions. The main contrasts of mean neural activation during the prosocial decision phase did not show significant neural activations (neither whole-brain nor small-volume corrected).

We investigated how the simultaneous activation of the two motives, and the resulting changes in initial prosocial bias and amount of required relative evidence affected the neural computation of prosocial choices. To do so, we regressed participants' individual percent change in initial prosocial biases ($\Delta z_{\text{multi-motive/reciprocity}}$) and the amount of relative evidence ($\Delta a_{\text{multi-motive/reciprocity}}$) on the neural contrast in prosocial choices between the multi-motive condition and the reciprocity condition, using second-level regressions. As a main result, the first analysis revealed activations in the bilateral dorsal striatum that were related to the individual change in prosocial bias (right hemisphere: P(FWE$_{\text{cluster-corrected}}$) = 0.001; center co-ordinates: x = 30, y = 2, z = -2; k = 143 voxels, T(31) = 5.49; left hemisphere: P(FWE$_{\text{cluster-corrected}}$) = 0.003; center co-ordinates: x = -28, y = -9, z = 1; k = 121 voxels, T(31) = 5.36; **Figure 2.3.4B, Figure A6, Table A14**). The stronger the percent increase in initial prosocial bias in the multi-motive condition relative the reciprocity condition, the stronger the neural response in bilateral dorsal striatum.

To test the alternative hypothesis that the increase in dorso-striatal activity may reflect the dominance of empathy (captured by the relative difference in *z*-parameters between empathy and reciprocity, $\Delta z_{\text{empathy/ reciprocity}}$), instead of a multi-motive effect, we also compared the relationship between $\Delta z_{\text{multi-motive/reciprocity}}$ and $\Delta z_{\text{empathy/reciprocity}}$ on extracted beta values of the multi-motive vs reciprocity contrast using an independent anatomical mask of bilateral putamen based on the aal nomenclature (Tzourio-Mazoyer et al., 2002). The results showed that neural activation in dorsal striatum is associated with $\Delta z_{\text{multi-motive/reciprocity}}$, but not with $\Delta z_{\text{empathy/reciprocity}}$ (significant interaction between index type and neural activation: β = 0.69, SE = 0.225, P = .003, no main effect of beta values: β = -0.11, SE = 0.159, P = .52, marginal main effect of index type: β = -0.41, SE = 0.223, P = .07, **Figure A5B**). Thus, empathy dominance is not likely to explain the results.

To test whether the increase in striatal activation in the multi-motive compared to the reciprocity condition is driven by outliers, we extracted the individual beta-estimates of the multi-motive vs reciprocity contrast from bilateral dorsal striatum and plotted its relationship with the percent change in the z-parameter in the multi-motive condition relative to the reciprocity condition (**Figure A5A**). The inspection of the plot shows that the relationship between the percent signal change in the z-parameter in the multi-motive condition relative to the reciprocity condition was not driven by outliers.

The respective second-level regression with the percent change in the *a*-parameter revealed neural activity in bilateral anterior insula on a lower, uncorrected threshold ($P_{\text{uncorrected}} <$ 0.001; center co-ordinates right hemisphere: x = 33, y = 32, z = 1, P(FWE$_{\text{cluster-corrected}}$) = .970, k = 9 voxels; center co-ordinates left hemisphere: x = -30, y = 27, z = -2, P(FWE$_{\text{cluster-corrected}}$) = .902, k = 13 voxels).

Additionally, we tested whether trial-by-trial changes in the partner's gain modulate neural activation during the prosocial choice process. In line with the behavioral results, the partner's gain did not differentially influence neural activation in the different conditions. However, neural activation during the prosocial choice process in bilateral insula was significantly associated with trial-by-trial changes in the partner's gain across all four conditions (right insula peak-coordinates: x = 43, y = -6, z = 18, k = 108 voxels, P(FWE whole-brain cluster corrected) = .009; left insula peak-coordinates: x = -38, y= -9, z = 16, k = 517 voxels, P(FWE whole-brain cluster corrected) < .001; see **Table A15** for all clusters k > 10). The same analysis including prosocial as well as egoistic choice trials replicated this result (right insula peak-coordinates: x = 43, y = -6, z = 16, k = 109 voxels, P(FWE whole-brain cluster corrected) = .009; left insula peak-coordinates: x = -38, y= -9, z = 18, k = 294 voxels, P(FWE whole-brain cluster corrected) < .001). Hence, the insular activation appears to track trial-by-trial changes of the partner's gain across all conditions.

Discussion

Many behaviors derive from complex motivational states that are characterized by different, simultaneously activated motives (Engel & Zhurakhovska, 2016; Hughes & Zaki, 2015; Jagers et al., 2017; Takeuchi et al., 2015; Terlecki & Buckner, 2015). However, the mechanisms through which combinations of motives affect behaviors, e.g., the computation of prosocial choices, are poorly understood.

Our results showed that the simultaneous activation of two motives changes participants' choices compared to activation of a single motive condition. In more detail, a combination of two prosocial motives (the empathy and the reciprocity motive) elicited more prosocial choices than the reciprocity motive alone (**Figure 2.3.3A**). This multi-motive increase occurred although the two single motives were activated with comparable strength (indicated by the induction ratings, **Figure A1**). Moreover, the different motive conditions had no effect on reaction times (**Figure 2.3.3B** and **C**), inequality aversion, or the subjective value assigned to the partner's gains (**Figure 2.3.3D**). Furthermore, the partner's gain was associated with neural activation in bilateral insula across all conditions, which adds to the observation that insular activation is also sensitive to other-regarding experiences such as vicarious reward (Morelli, Sacchet, & Zaki, 2015), avoiding risk for others (Shenhav & Greene, 2010), and making fair (Dawes et al., 2012) or altruistic decisions (Cutler & Campbell-Meiklejohn, 2019).

Specifying the change in prosocial choice behavior in the multi-motive condition, the drift-diffusion-modeling analyses showed that the combination of the two motives enhanced participants' initial bias for making a prosocial choice, compared to the reciprocity condition (reflected by the increase in *z*-parameter; **Figure 2.3.4A**) and with a similar trend, relative to the empathy condition. Moreover, the combinations of the two motives increased the relative amount of relative evidence that participants required during the choice process (reflected by the relative increase of the *a*-parameter). This indicates that participants assessed their choices more carefully if they made them based on two different motives. In contrast, the speed of information accumulation, i.e., the efficiency of the decision process itself (reflected by the *v*-parameter), remained unchanged. The observed change in initial prosocial bias (the *z*-parameter) is in line with previous findings that reported a shift of choice biases due to the prior likelihood of one of the choice options or a higher reward value associated with one option (Mulder et al., 2012), personal predispositions (Chen & Krajbich, 2018), or prior information about how other people decided (Toelch et al., 2018). Extending these results, our findings reveal that initial choice biases are altered by simultaneously activated motives, and thus characterize how complex motivational states change the choice process compared to single-motive states.

We hypothesized that changes in DDM choice parameters in the multi-motive compared to the single-motive conditions might be related to changes in activation in the ventral or dorsal striatum, inspired by evidence associating the ventral striatum with the processing of choice values (Kable & Glimcher, 2007; Liljeholm & O 'Doherty, 2012; O'Doherty et al., 2004; Strait et al., 2015), and/or the dorsal striatum with encoding of choice preferences (Balleine et al., 2007; Liljeholm & O 'Doherty, 2012; O'Doherty et al., 2004; Palmiter, 2008; Robinson et al., 2006).Our results show that the combination of different motives is associated with an increase in activation in bilateral dorsal striatum, reflecting an enhancement of individual prosocial choice biases in the multi-motive condition relative to the reciprocity condition (**Figure 2.3.4**), and, based on extracted beta-values from an independent anatomical region, also relative to the empathy condition (**Figure A5**). The increase in activation in dorsal striatum is in line with previous neuroscience studies showing that motivation-related changes in decision parameters are captured by dorsal striatal responses (Forstmann et al., 2008; Gluth et al., 2012). Extending this previous evidence, we show that the dorsal striatum integrates choice biases that are elicited by multiple motivational forces, and thus provides a plausible neural candidate for the generation of complex motivational states.

We found that the simultaneous activation of the empathy motive and the reciprocity motive in the multi-motive condition enhanced the participants' initial prosocial biases relative to the reciprocity condition. This indicates that the empathy motive enhanced the reciprocity motive, but not vice versa. Given this result, we argued that the observed changes in the multi-motive condition may reflect the dominance of one motive over the other motive (i.e., a dominance of empathy over reciprocity). If this were true, the multi-motive induced changes in the choice process would reflect a motivation that is similar to the state induced by the dominant motive, instead of a more complex motivational state that was incited by the combination of different motives. Our results show that the multi-motive induced changes in the choice process (i.e., DDM and neural choice parameters) are in fact related to differences between the multi-motive condition and the reciprocity condition and cannot be explained by mere dominance of the empathy motive over the reciprocity motive. This finding supports the conclusion that the simultaneous activation of two motives alters the prosocial choice process compared to single motive states.

Because we were mainly interested in participants' choices under the different motive conditions, our paradigm was designed to optimize the number of trials in the allocation task. The motive induction procedures only included the minimal number of trials required for inducing the different motives (twelve trials per motive induction). Due to the small number of trials, an analysis of neural responses during the multi-motive and single-motive induction procedures would not be meaningful.

Participants made prosocial choices also in the baseline condition which indicates that participants are motivated to behave prosocially without experimental activation of empathy and reciprocity. It is thus important to note that prosocial choices can be driven by other motives in addition to empathy and reciprocity. However, these additional motives should be the same across conditions since participants perform the same social choice task. Hence, contrasting behavior between the different conditions should carve out the effects that were experimentally manipulated, i.e., how the combination of empathy and reciprocity influences the prosocial choice process relative to empathy or reciprocity alone.

Likewise, because the “basic” social choice tasks were identical in all experimental conditions, task-specific effects should average out if the different conditions are contrasted. This means that the observed effects are driven by the different motive inductions and should be independent of the social choice task that was used in the present study. In other words, our behavioural and neural findings should generalize to other behaviours that are elicited by the combination of the empathy and the reciprocity motives. However, in how far the present effects are scaled depending on the exact task affordances (e.g., relative importance of the single motives for the respective task, how much time participants have to deliberate their decision) is a question for future research. Likewise, future studies need to test if the observed increase in striatal activation due to a multi-motive alteration of initial choice bias also applies to other (e.g., non-social) motivational states. Food choices, for example, are often driven by more than one motive such as the motive to eat healthy and the motive to eat sweet food, maximizing calorie intake. The dorsal striatum has previously been associated with food choice preferences in healthy (Small, Jones-Gotman, & Dagher, 2003; Wallace et al., 2014) as well as pathological participants (Foerde, Steinglass, Shohamy, & Walsh, 2015). It is thus possible that the interplay between the non-social motives during food choices is associated with neural activation in the dorsal striatum.

To avoid cross-gender effects, which are likely to occur if female participants are paired with male confederates and vice versa, we only tested females. Future studies are required to show if our results generalize to male participants.

Conclusions

Based on our current findings we conclude that the simultaneous activation of two different prosocial motives changes the computation of prosocial choices. According to our results, choices that are made in a more complex motivational state, i.e., driven by multiple motives, are characterized by a change in initial choice bias, which is associated with an increased neural response in dorsal striatum. Moreover, choices were made more carefully relative to simple motivational states. Together, these findings show how the human brain combines different prosocial motives, and how this motive combination affects the computation of prosocial choices.

Authors' Contributions

G.H. and A.S. designed the research with input from J.K., M.L., and U.H.; A.S. and U.H. performed the research; A.S. programmed the experiment; A.S. and U.H. analyzed the data with input from G.H., M.L., and J.K.; G.H. and A.S. wrote the paper with input from U.H., M.L., and J.K..

Acknowledgments

We thank Martin Domin, Jörg Pfannmöller, and Kai Klepzig for their technical support, and Vassil Iotzov for helpful discussions about the data.
This work was supported by the German Research Foundation (HE 4566/2-1; HE 4566/5-1) and the German Academic Scholarship Foundation.

Implications for study 4

In study 3, we investigated how the combination of empathy with another social motive that incites prosocial behavior shapes the prosocial decision process compared to the prosocial decision process only based on empathy and only based on that other prosocial motive, namely reciprocity. Behavioral results showed that the combination of empathy and reciprocity increased participants' probability to make prosocial decisions as well the initial decision bias towards making a prosocial decision. Additionally, the larger an individual's increase in initial prosocial decision bias, the larger the increase in neural activation in bilateral dorsal striatum. This behavioral and neural increase, however, was only observable relative to the reciprocity-related prosocial decision process but not relative to the empathy-related prosocial decision process. Thus, empathy can boost prosocial behavior based on a social norm that has been shown to motivate prosocial behavior.

In study 4, we tested whether empathy is also sustainable with respect to being resilient to the combination with the motive of outcome maximization. Previous works have shown that paying people to act prosocially does not necessarily boost prosocial behavior but may also undermine other social motives motives to do so and hence impede prosocial behavior. In study 4, we tested empathy sustainability in this regard by activating empathy for pain as well as paying participants a bonus for making prosocial decisions, hence offering a financial incentive to act prosocially (empathy-bonus condition). We compare the prosocial decision process based on this combined motivation with the prosocial decision process based on empathy alone (empathy alone condition). If empathy is resilient to an undermining effect of financial incentives, prosocial decision-making in the empathy-bonus condition should be comparable to or even increase compared to the empathy alone condition.

2.4 Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females

Iotzov, Vassil*[1,2]; Saulin, Anne[1]; Kaiser, Jochen[2]; Han, Shihui[3]
and Hein, Grit*[1]

[1]Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Wuerzburg, 97080 Würzburg, Germany.
[2]Institute of Medical Psychology, Faculty of Medicine, Goethe University, 60528 Frankfurt am Main, Germany.
[3]School of Psychological and Cognitive Sciences, PKU-IDG/ McGovern Institute for Brain Research, Peking University, Beijing, 10008, China.

*corresponding authors:
Vassil Iotzov, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Iotzov_V@ukw.de

Prof. Dr. Grit Hein, Translational Social Neuroscience Lab, Department of Psychiatry, Psychosomatic and Psychotherapy, University Hospital of Wuerzburg, Margarete-Höppel-Platz 1, 97080 Würzburg / Germany, E-mail: Hein_G@ukw.de

Published 2022 in Social Neuroscience

Reprinted in this dissertation with permission

Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females

Abstract

Financial incentives are commonly used to motivate behaviors. However, there is also evidence that incentives can impede the behavior they are supposed to foster, for example, documented by a decrease in blood donations if a financial incentive is offered. Based on these findings, previous studies assumed that prosocial motivation is shaped by incentives. However, so far, there is no direct evidence showing an interaction between financial incentives and a specific prosocial motive. Combining drift-diffusion modeling and fMRI, we investigated the effect of financial incentives on empathy, i.e., one of the key motives driving prosocial decisions. In the empathy-alone condition, participants made prosocial decisions based on empathy. In the empathy-bonus condition, they were offered a financial bonus for prosocial decisions, in addition to empathy induction. On average, the bonus enhanced the information accumulation in empathy-based decisions. On the neural level, this enhancement was related to the anterior insula, the same region that also correlated with empathy ratings. Moreover, the effect of the financial incentive on anterior insula activation was stronger the lower a person scored on empathy. These findings show that financial incentives enhance prosocial motivation in the absence of empathy.

keywords:
empathy, prosocial behavior, incentives, drift-diffusion modelling, fMRI

Introduction

Financial incentives are frequently used to motivate people. Such measures are based on empirical evidence showing that financial incentives increase the frequency of the rewarded behavior (Garbers & Konradt, 2014; Wei & Yazdanifard, 2014), including cooperative and prosocial behaviors (Balliet et al., 2011; Stoop, van Soest, & Vyrastekova, 2018). For example, in a meta-analysis, Balliet and colleagues found that reward positively affects cooperation (Balliet et al., 2011). Consequently, financial incentives may increase the motivation to behave prosocially (Ariely, Bracha, & Meier, 2009). However, there is other evidence that incentives may undermine the very behavior they are meant to strengthen (Bénabou & Tirole, 2006; Besley & Ghatak, 2018; Deci, Koestner, & Ryan, 1999; Murayama et al., 2010; Niza, Tung, & Marteau, 2013; Rode, Gómez-Baggethun, & Krause, 2015; Titmuss, 1970). The most classic example in the realm of prosocial behaviors is the observation that people donate less blood if they are paid to do so, compared to the amount of blood that they donate without payment, i.e., only motivated by wanting to help others (Niza et al., 2013; Titmuss, 1970). In line with these observations, other studies have shown that adding financial incentives can reduce prosocial behaviors (Ariely et al., 2009; Bowles, 2008; Holmås, Kjerstad, Lurås, & Straume, 2010). In sum, the evidence regarding the effects of incentives on prosocial decisions is inconsistent and mainly based on behavioral observations that do not provide insights into the underlying motivational processes. As a result, it remained unclear whether and how financial incentives interact with a specific prosocial motive.

In social psychology models of prosocial behavior, financial incentives play a role because they can motivate prosocial behaviors based on an egoistic motive (Batson & Shaw, 1991). Incentivized prosocial behavior that is driven by an egoistic motive can benefit the other, but the benefit for others is only a byproduct and the ultimate goal is the increase of the decider's welfare. In contrast, in case of an empathic motivation, the decider ultimately strives to increase the wellbeing of the other, irrespective of a potential reward (Batson, 1994; Batson et al., 2011, 2004). Within this (Batson, 1994; Batson et al., 2011, 2004) and in other recent motivation models (Engel & Zhurakhovska, 2016; Hughes & Zaki, 2015; Saulin et al., 2022), it is proposed that different motives influence each other and that most behaviors are driven by an interaction between these different motives. Previous social psychology work has investigated how empathy is shaped by selfish motives, such as the motive to withdraw from a stress-inducing situation (Batson, Duncan, Ackerman, Buckley, & Birch, 1981). However, to the best of our knowledge, there are no previous studies that tested how financial incentives affect the components of empathy-based prosocial decisions.

Empathy itself is a multidimensional construct (Timmers et al., 2018). Commonly, researchers distinguish between so called cognitive empathy or theory of mind (ToM) and emotional empathy – a distinction that is even mirrored on a neural level (Cox et al., 2012; Cutler & Campbell-Meiklejohn, 2019; Decety et al., 2016; Dvash & Shamay-Tsoory, 2014; Fan et al., 2011; Kanske et al., 2015; Preckel et al., 2018; Shamay-Tsoory et al., 2009; Stietz et al., 2019; Zaki & Ochsner, 2012). Cognitive empathy has often been associated with neural activation of the medial prefrontal cortex (mPFC), the superior temporal sulcus (STS), the temporal poles (TP), and the temporo-parietal junction (TPJ; Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019), while emotional empathy is often associated with the anterior insula (AI), and the anterior and mid cingulate cortex (ACC/MCC; Cutler & Campbell-Meiklejohn, 2019; Dvash & Shamay-Tsoory, 2014; Fan et al., 2011; Preckel et al., 2018; Schurz et al., 2020; Stietz et al., 2019). That said, there is recent evidence pointing to an involvement of the AI and the ACC in tasks requiring cognitive and emotional empathy, in addition to task requiring emotional empathy only (Cutler & Campbell-Meiklejohn, 2019; Schurz et al., 2021).

Previous work has established a reliable link between the individual strength of the empathy motive and the propensity to act prosocially, e.g., decisions that maximize the outcome of another person at costs to oneself (Batson et al., 1995; Decety et al., 2016). The stronger the empathy motive, the stronger the propensity to decide in favor of the other person. It is assumed that many prosocial decisions are driven by both, cognitive and emotional empathy (Kanske et al., 2015; Stietz et al., 2019; Zaki & Ochsner, 2012).

In the present study, we induced empathy using a well-established empathy for pain paradigm in which participants observed two interaction partners receiving painful shocks (Hein, Engelmann, et al., 2016; Hein, Morishima, et al., 2016; Lamm et al., 2011). This procedure has been shown to induce empathy as a motive that incites prosocial behavior based on the affective response to another person's misfortune (Batson et al., 1995; Decety et al., 2016; Hein, Morishima, et al., 2016; Lamm et al., 2011; Marsh, 2018). Nevertheless, in the light of previous evidence that emotional and cognitive empathy also work in concert (Kanske et al., 2015; Preckel et al., 2018; Stietz et al., 2019; Zaki & Ochsner, 2012), cognitive empathy may also play a role in the prosocial decision process.

To investigate the effect of financial incentives on empathy-based prosocial decisions, participants allocated points to the partners at a cost to themselves (**Figure 2.4.1B**). The allocation of points towards the one partner (empathy partner) should be based on the previously activated empathy motive (empathy-alone condition). The allocation of points towards the other partner (empathy-bonus partner) was also based on the previously activated empathy motive. However, in this condition, participants were additionally informed that they would receive a bonus for choosing the prosocial option in the majority of trials in the subsequent allocation task (empathy-bonus condition). Importantly, achieving the bonus criterion in the empathy bonus condition did not result in a financial loss for participants. To control for other motivations that might play a role besides empathy (self-image concerns; reciprocity), the incentive was offered in private, the decisions were kept anonymous, and the participants knew that they would not meet the other players after the study. This measure is important because it minimizes participants' motivation to maintain a positive public image, i.e., a different motive that may affect participants' prosocial decisions besides empathy (Ariely et al., 2009; Bénabou & Tirole, 2006; Besley & Ghatak, 2018; Exley, 2018).

To specify how incentives modulate empathy-related decisions, we used drift-diffusion modeling (DDM). DDMs assume that during binary decisions, noisy information is accumulated to select a decision option mainly based on three different parameters (the *v*-, *z*- and *a*-parameters; **Figure 2.4.1C**) (Forstmann et al., 2016; Ratcliff et al., 2016). The *v*-parameter describes the speed of the evidence accumulation, i.e., the efficiency of the decision process itself. Thus, in our task, a larger *v*-parameter indicates faster information accumulation regarding the prosocial option. The individual decision bias is reflected by the *z*-parameter. In contrast to the *v*-parameter, the *z*-parameter models the individual preferences with which a person starts the decision process. For example, if a person has a strong prior preference for prosocial decisions, the starting point of the decision process is closer to the prosocial decision boundary, and therefore less evidence has to be accumulated regarding the prosocial option.

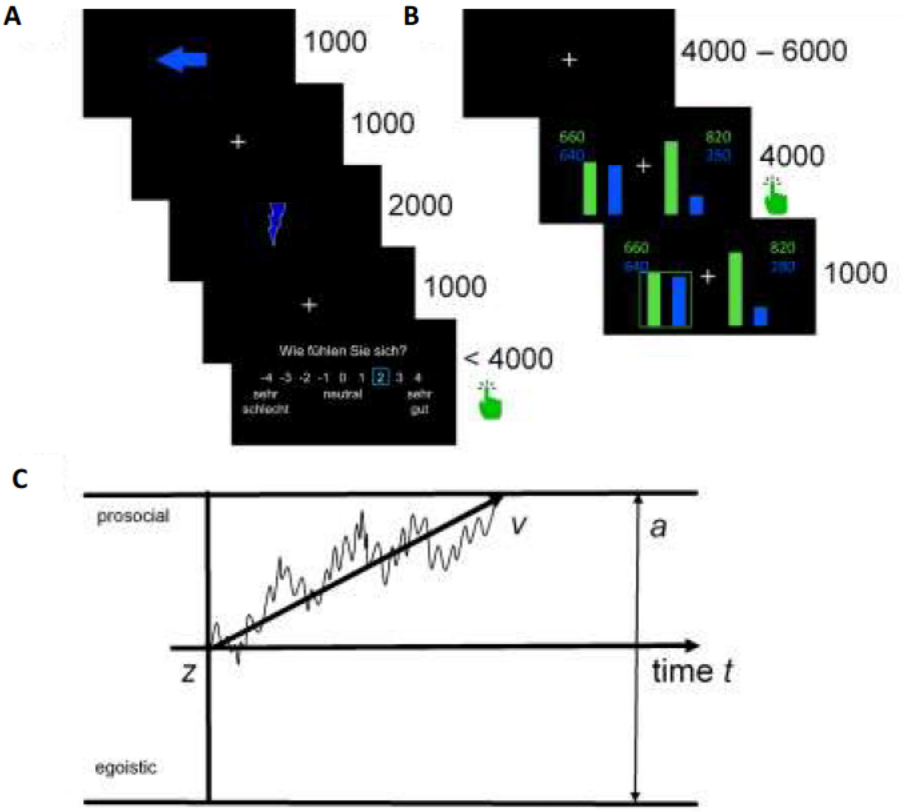


Figure 2.4.1 Examples of induction and decision trials and schematic overview of the drift-diffusion model (DDM). **A** Example trial of the empathy induction. The arrow cue indicated the receiver of the stimulation (self, the empathy-alone partner in one condition, or the empathy-bonus partner in the other condition). The lightning bolt indicated pain stimulation. Participants rated how they felt after observing the stimulation of the partner or receiving it themselves (-4 = very bad; +4 = very good). **B** Example trial of the allocation task. Participants chose between a prosocial option that maximized points for the partner or a selfish option that maximized points for themselves. In this example trial, the participant chose the prosocial option, which maximized the outcome of the partner at a cost to the participant (green box). **C** Schematic overview of the drift-diffusion model. According to the drift-diffusion model, the decision process is a noisy accumulation of information (jagged black line). From the distributions of both prosocial and selfish decisions, a set of parameters is estimated that allows drawing conclusions about the underlying cognitive processes. These are mainly the speed of information accumulation (v-parameter), the starting point of the decision process (z-parameter), and the amount of information to be processed (a-parameter). As soon as the accumulated information reaches one of the two boundaries, the decision is made (upper boundary = prosocial option; lower boundary = selfish option).

The amount of evidence that needs to be accumulated to distinguish between the two options is reflected by the a-parameter. We modeled these three parameters (*v*, *z*, and *a*) for decisions that were driven by the empathy motive alone and that were driven by the combination of the empathy motive and the financial incentive, based on the raw data from the entire data set (i.e., including trial-by-trial information of all decisions). Additionally, the non-decision time (*t0*) was estimated across conditions (see Drift-Diffusion Modeling for details).

Extending the classical DDM approach, a recent model has proposed that the evidence in favor of one or another choice alternative might be shaped by affective and motivational states (Roberts & Hutcherson, 2019). Supporting this assumption, affective states have been found to change central parameter of the choice process, such as the drift rate (*v*-parameter) (Aylward, Hales, Robinson, & Robinson, 2019; Lerche, Neubauer, & Voss, 2018; Thompson & Steinbeis, 2021) and the starting point (*z*-parameter) (White, Liebman, & Stone, 2018). Inspired by these results, we assumed that the evidence in favor of a prosocial choice might be different in different motivational states (i.e., induced by empathy and its potential interaction with the incentive), reflected by changes in the drift rate and/ or the starting point.

Influential social psychology models propose that incentive-induced egoism and empathy can incite prosocial behavior (Batson, 1994; Batson et al., 2011, 2004) and that different motives interact with each other (Engel & Zhurakhovska, 2016; Hughes & Zaki, 2015). One assumption is that financial incentives may enhance empathy-related prosocial decisions, inspired by findings of reward-related increases of prosociality (Garbers & Konradt, 2014; Wei & Yazdanifard, 2014). If this was true, the frequency and efficiency of prosocial decisions should be higher in the empathy-bonus compared to the empathy-alone condition. Specifying the potential effect of the incentive on the prosocial choice process, the DDM proposes that incentive-related facilitation of prosocial choices may originate A) from an increased speed of information accumulation, i.e., an increased drift rate (*v*-parameter, **Figure 2.4.2A**) (Aylward et al., 2019; Lerche et al., 2018; Roberts & Hutcherson, 2019; Thompson & Steinbeis, 2021), B) an enhancement of participants' initial preference to choose the prosocial option, i.e., a shift of the starting point towards the prosocial decision boundary (*z*-parameter, **Figure 2.4.2C**) (White et al., 2018), or C) from an enhancement of the v- as well as the z-parameter in the empathy-bonus compared to the empathy-alone condition (**Figure 2.4.2E**).

Alternatively, it is possible that financial incentives undermine empathy-related prosocial decisions, in line with previous findings that showed an incentive-related decrease in prosocial behavior (Bénabou & Tirole, 2006; Murayama et al., 2010; Rode et al., 2015; Titmuss, 1970). In this case, prosocial decisions should be more frequent in the empathy-alone compared to the empathy-bonus condition. According to the DDM, such an undermining effect may be reflected A) by a reduced speed of information accumulation (*v*-parameter; **Figure 2.4.2B**), B) a shift of the starting point away from the prosocial decision boundary (*z*-parameter; **Figure 2.4.2D**), or C) a reduction in both parameters in the empathy-bonus compared to the empathy-alone condition (**Figure 2.4.2F**).

Finally, it is possible that the effect of financial incentives depends on the strength of the empathy motive, i.e., might be influenced by how highly empathic individuals are. If this is true, the individual difference between the empathy-bonus vs. empathy-alone condition and changes in the drift rate and/or the starting point should be related to the individual empathy ratings, i.e., the measure that captures the strength of the empathy motive during the first part of the study.

In line with the notion that DDM-based analyses provide an elegant approach to relate individual differences in cognitive processes to neural activity (White, Curl, & Sloane, 2016), previous studies have started to link changes in DDM parameters to changes in neural processing (Ulrike Basten, Biele, Heekeren, & Fiebach, 2010; De Hollander, Forstmann, & Brown, 2016; de Lange, Rahnev, Donner, & Lau, 2013; Domenech, Redouté, Koechlin, & Dreher, 2018; Forstmann & Wagenmakers, 2015; Gluth et al., 2012; Mulder et al., 2012; Pedersen, Endestad, & Biele, 2015; Peters & D'Esposito, 2020; White et al., 2016). For example, an increase in the *z*-parameter has been linked to an increase in frontoparietal activation (specifically superior frontal gyrus, right middle frontal gyrus, left inferior frontal gyrus, left intraparietal sulcus, medial frontal gyrus, anterior cingulate gyrus;(Mulder et al., 2012) and motor cortex (de Lange et al., 2013).

An increase in drift rate (*v*-parameter) has been associated with increased activity in presupplementary motor area, caudate nucleus, and anterior insula (Gluth et al., 2012) and the dorsomedial prefrontal cortex, right inferior frontal gyrus, and bilateral insula (Pedersen et al., 2015; for reviews, see De Hollander et al., 2016; Forstmann & Wagenmakers, 2015).

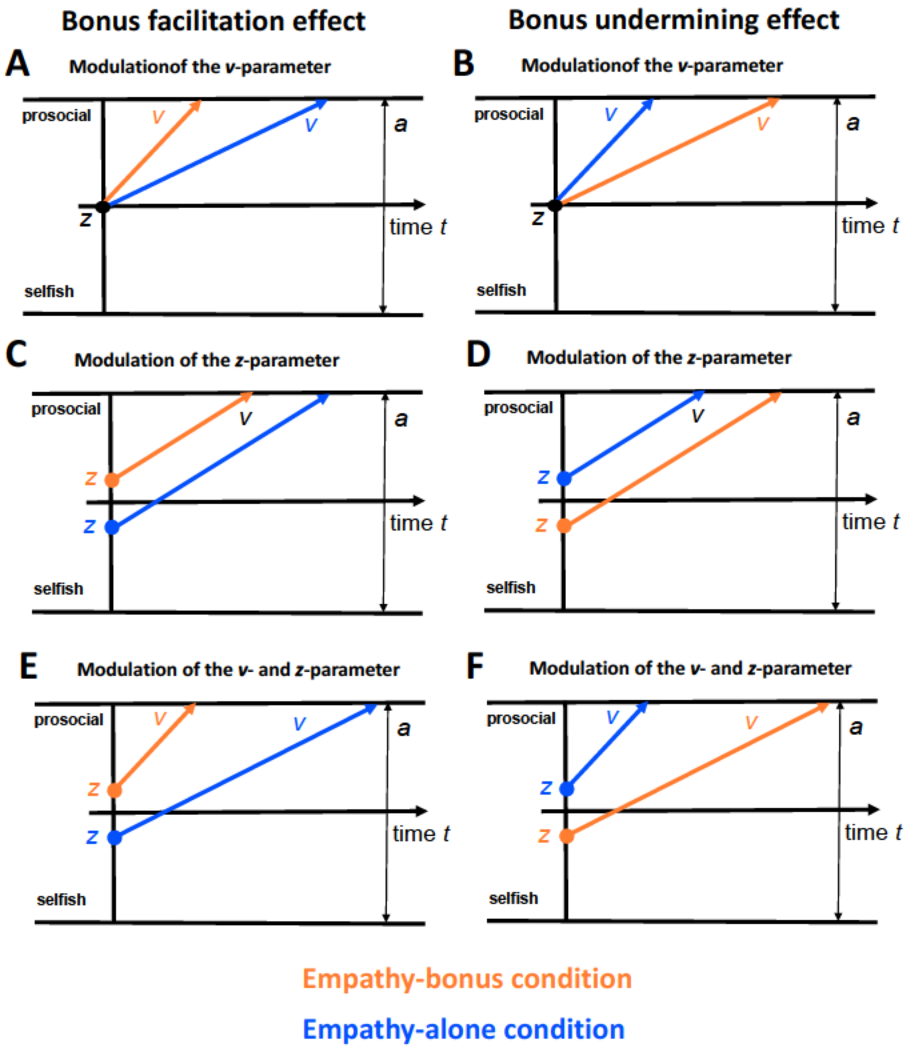

Bonus facilitation effect
Bonus undermining effect
A
Modulationof the v-parameter
B
Modulationof the v-parameter
prosocial
v
v
a
z
time t
selfish
C
Modulation of the z-parameter
D
Modulation of the z-parameter
E
Modulation of the v- and z-parameter
F
Modulation of the v- and z-parameter
Empathy-bonus condition
Empathy-alone condition

Figure 2.4.2. Example of hypotheses regarding the effects of the bonus on the drift-diffusion parameters v and z during empathy-based prosocial decisions. A facilitation effect of the bonus and thus an increase in prosocial decisions frequency in the empathy-bonus condition (orange) compared with the empathy-alone condition (blue) may result from an increased speed of information accumulation (v-parameter; **A**), an increased initial bias toward prosocial decisions (z-parameter; **C**) or by a modulation of both parameters (v- and z-parameter; **E**. An undermining effect of the bonus and thus a reduction in prosocial decisions frequency in the empathy-bonus condition (orange) compared with the empathy-alone condition may result in a reduced speed of information accumulation (v-parameter; **B**), a reduced initial bias toward prosocial decisions (z-parameter; **D**), or by a modulation of both parameters (v- and z-parameter; **F**).

Some of the regions that capture changes in individual components of the decision process have also been related to prosocial decisions. There is well-established evidence that prosocial decisions correlate with brain activations in regions that are also associated with individual differences in empathy, such as the anterior insula (AI) cortex and the anterior and mid cingulate cortex (ACC/MCC) (Hein, Morishima, et al., 2016; Hein et al., 2010; Marsh, 2018; Masten, Morelli, et al., 2011). Moreover, prosocial decisions were found to involve medial prefrontal regions and temporo-parietal regions that have been associated with cognitive empathy (Dvash & Shamay-Tsoory, 2014; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019), as well as reward-related regions such as the striatum (Preckel et al., 2018). Based on this previous evidence, the processing of an incentive-related increase in the *v*- and/or *z*-parameter (reflecting facilitation of empathy-related decisions) between the empathy-bonus and the empathy-alone condition may increase the neural activation in a network consisting of the AI, the ACC/MCC, medial prefrontal, temporo-parietal and striatal regions. In contrast, an incentive-related decrease in the *v*- and/or *z*-parameter (reflecting a potential undermining effect) may be related to a decrease of activity in this network.

Methods

Materials and Methods

Participant details

33 healthy women (mean age M = 25.05 years, SEM = 0.74, min = 18, max = 35) participated in the study. Females were invited to participate irrespective of race and ethnicity, but the final sample was 100% Caucasian. All of them had the German “Abitur” (the diploma required for admission to college studies). Participants were recruited via flyers distributed at the Frankfurt University. They were required to master German on a C1 level in order to ensure understanding of the instructions, to have normal or corrected-to normal vision, to be right-handed, and to have no history of mental disorders and regular drug consumption. The sample size was estimated based on a meta-analysis, which showed that rewards have a positive effect on cooperation of d = 0.51 (Balliet et al., 2011). A post hoc sensitivity analysis was conducted using G*Power version 3.1.9.2 (Erdfelder, FAul, Buchner, & Lang, 2009). According to the estimation, a minimal sample size of N = 33 is required to detect effects of incentive on prosocial decisions, comparing two dependent means with α = .05 and power (1-β) = .80. The power of .80 was chosen based on the recommendations from Ellis (2012, p. 53). We chose a female instead of a gender-mixed subject group because it allowed us to choose female confederates and thus to avoid the potential complications of gender-mixed

pairing of participants and confederates. The confederates were two female students trained to play their roles in counterbalanced order. The data from two participants had to be discarded as outlier (frequency of prosocial decisions, 3.42 SDs below the mean ($M_{empathy-alone}$ = 44.35, SD $_{empathy-alone}$ = 12.97). Thus, we analyzed 31 data sets. We obtained ethics approval (EK 458122014) for conducting the study and written informed consent from our participants. The experiment was conducted following the Helsinki guidelines. Participants received monetary compensation (show up fee plus payout from two randomly chosen trials of the allocation task).

Procedure

Overall procedure

The study consisted of two parts (**Figure 2.4.3**). In part 1, the empathy motive was activated towards one partner (a confederate). In the following allocation task, participants allocated points to the respective partner (here driven by empathy; empathy-alone condition). Next, the confederate was replaced by a new individual that served as a partner for part 2. In part 2, the empathy motive was activated again. However, before starting the decision task, the participant was told that she would receive a bonus if she decided prosocially in the clear majority of the decision trials. In the following allocation task, participants again allocated points to the respective partner (here driven by empathy and the financial incentive; empathy-bonus condition). The order of the two conditions (empathy-alone and empathy-bonus) was counterbalanced across participants and the two confederates.

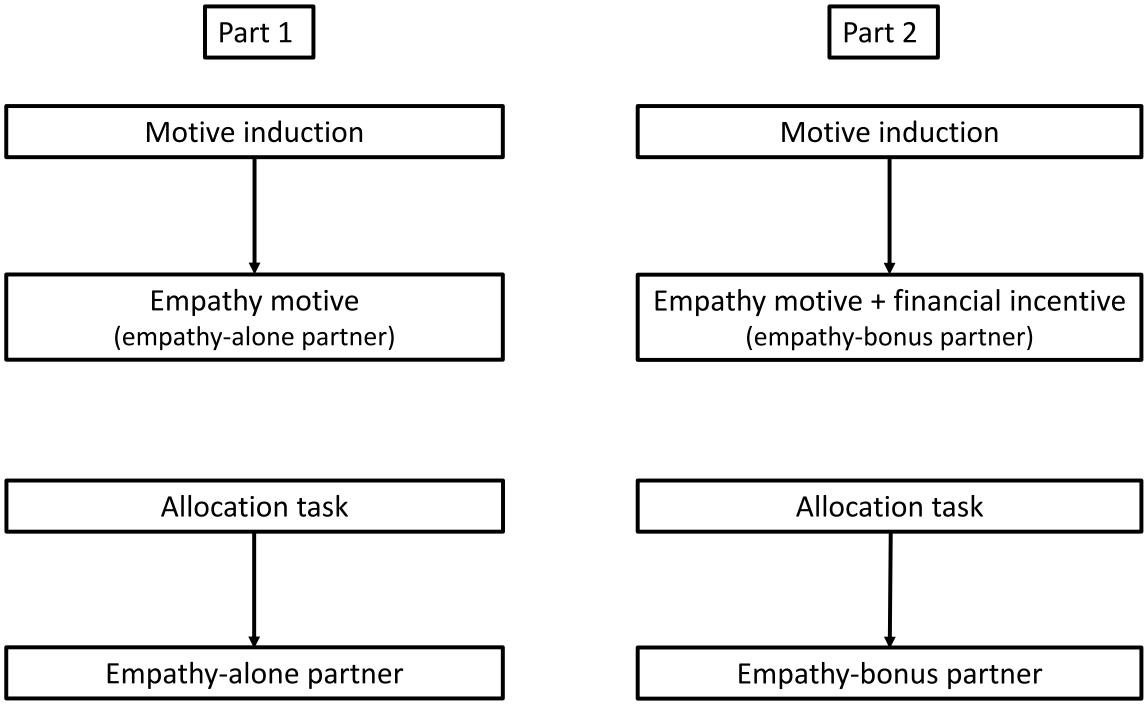


Figure 2.4.3. Overview of an exemplary experimental procedure. The study consisted of two parts. In this example, in part 1, the empathy motive was activated towards one confederate (the empathy-alone partner). In the following allocation task, participants allocated points to the empathy partner (i.e., driven by the empathy motive). Next, the confederate was replaced by a new individual that served as partner for part 2. Again, the empathy motive was activated towards this second confederate. After the empathy motive induction, additionally, a bonus for choosing the prosocial option in the majority of trials in the subsequent allocation task was offered (empathy-bonus partner). Thus, in the following allocation task, participants allocated points towards the empathy-bonus partner (i.e., driven by the empathy motive and the additionally offered bonus). The order of motive induction (empathy-alone, empathy-bonus) was counterbalanced across participants and both confederates. The respective partner was indicated by a cue in one of two counterbalanced colors.

Outside the fMRI scanner, we attached pain electrodes to the back of the participants' and the confederates' hands and determined the individual thresholds for painful and painless stimulation using a standard procedure (Hein, Engelmann, et al., 2016; Hein, Morishima, et al., 2016). Next, the participant and the confederates played a manipulated lottery (drawing matches) that ostensibly determined the amount of pain the person would receive in the following task. Because the empathy induction required saliently more pain for the confederates, the drawing of the matches was organized in such a way that the participant always drew the last match and thus was assigned to receive only a few painful stimuli.

The participant was placed inside the fMRI scanner, and one of the confederates was placed on a chair next to the participant in the scanner room. The confederate's hand with the pain electrode was placed on a tilted table over the participants' knee. Through a mirror in the head coil, participants could see the hand of the other, together with the visual stimulation on a screen that was positioned at the end of the fMRI bed. During the empathy induction, participants either saw a dark-colored flash (painful stimulation) or a light-colored flash (non-painful stimulation), indicating the intensity of the stimulation of the confederate. In a small portion of trials (five from fifteen), they received pain stimulation themselves, indicated by a dark-colored flash of a different color. During the decision task, participants were presented two options to allocate points between themselves and the other person. Colors were counterbalanced across participants.

The study started with the empathy induction, followed by the allocation task towards the first confederate. After replacing this confederate, the same procedure (empathy induction followed by the allocation task) was repeated with the second confederate (**Figure 2.4.3**). In the empathy-alone condition, the allocation task started immediately after the empathy induction. In the empathy-bonus condition, after the empathy induction, participants were told that they would receive a bonus (additional 5 Euro payment) if they chose the prosocial option in the majority of trials. We deliberately refrained from specifying the percentage of prosocial decisions that were required to win the bonus to avoid strategy effects. The bonus was equal to the maximally possible outcome in the allocation task (i.e., the outcome that a participant would gain if she always chose the selfish option). To minimize reputation effects, participants received the bonus information in private without the partner's knowledge.

Apart from the bonus in the empathy-bonus condition, the experimental procedure was identical in both conditions. The order of the conditions and the assignment of the confederates was counterbalanced across participants. At the end of the experiment, both confederates left, and the participants stayed in the scanner until anatomical image acquisition was completed. Finally, participants were asked to complete the Interpersonal Reactivity Index (IRI) (Davis, 1980) and a scale that assessed their impression of both confederates (Hein, Engelmann, et al., 2016). The impression ratings were comparable between confederates (lmm $\chi^2(1) = 0.36$, $p = .55$, $B = -0.10$, $SEM = 0.16$).

Participants spent approximately 60 min inside the scanner, and the entire procedure lasted about 2 hours. In addition to the show-up fee, participants received the payout from two randomly chosen allocation trials and the bonus of five Euros if they made prosocial decisions in 75% of the trials in the empathy-bonus condition.

All ratings during the induction phase and all decisions in the allocation task were kept anonymous. Particular care was taken to ensure that this was clear to participants by pointing out the following: Inside the scanner room, the partner had a separate visual display, such that the participant viewed stimuli via back-projection from a mirror onto a screen, while the confederates beside the scanner viewed stimuli via cardboards/video glasses with a built-in display (Hein, Engelmann, et al., 2016). Thus, all ratings and decisions were private and could not be observed by the other participants (Hein, Engelmann, et al., 2016). Moreover, participants knew that they would not meet after the experiment because the scanned participant needed to stay longer for an anatomical scan. The experimenter was outside the scanner room, and it was pointed out that he could not see the ratings and decisions either.

Empathy induction

In each empathy-induction trial, first, we presented a colored arrow indicating the person who will receive the following electric stimulation for 1000 ms. After this cue, a fixation cross was presented for 1000 ms, followed by a colored lightning bolt shown for 2000 ms. Participants were informed that a blinking dark-colored lightning bolt indicates a painful stimulus, whereas a blinking light-colored lightning bolt indicates a non-painful stimulus. After receiving or observing the electric stimulation, we showed a 9-point rating scale with the question "How do you feel?". The scale ranged from -4 (labeled "very bad") to +4 (labeled "very good"). Participants had to respond within 4000 ms (**Figure 2.4.1A**). The empathy induction consisted of 30 trials: 10 that were ostensibly painful for the partner (other-pain trials), 5 that were not painful for the partner (other-no-pain trials), 5 painful trials for the participant (self-pain trials), and 10 non-painful trials (self-no-pain trials) for the participant. The self-pain trials were added to allow participants to simulate the state (pain) of the other person. To test their potential influence on empathy changes, we compared the ratings in other-pain trials that were preceded by a self-pain trial (i.e., empathy ratings under the condition of self-pain experience) with the ratings in other-pain trials that were preceded by an other-pain trial (i.e., empathy ratings without preceding self-pain experience). The results showed no difference between the other-pain ratings after self-pain and the other-pain ratings without prior self-pain (T(61) = 0.34, p = .73). Based on these results, the self-pain experience had no significant effect on empathy changes during empathy induction.

To further account for the potential effect of self-pain experiences on empathy ratings, individual empathy ratings (i.e., ratings for others’ pain) were divided by individual self-pain ratings. This quotient reflects the feeling for others pain relative to self-pain and was used as a continuous measure of state empathy in all analyses.

Allocation task

The allocation task was identical in both conditions and based on a well-established paradigm (Hein, Morishima, et al., 2016). In each trial, participants allocated points to themselves and the respective partner (**Figure 2.4.1B**) and could choose between maximizing the relative outcome of the other person by reducing their own relative outcome (prosocial choice) and maximizing their own relative outcome at a cost to the partner (selfish choice). The outcome was relative to the outcome that the participant would have gained when choosing the other option. The initial number of points was always higher for the participant compared to the partners. This measure was inspired by previous behavioral economics research, showing that participants make more prosocial decisions if their initial payoff is higher than the partner's payoff ("advantageous inequality") (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Schmidt, 1999). The choice options used in the present study created advantageous inequality to optimize the number of prosocial choices, which was the main focus of our study.

For the point distributions, we used values between 900 and 1200. The respective value was divided into a self:other ratio of 60:40 or of 90:10. Each trial of the allocation task contained a prosocial and a selfish option. The prosocial option was always the more egalitarian option, with a point distribution of 60% (self) to 40% (other). In contrast, in the selfish option, points were allocated with a ratio of 90% (self) to 10% (other). Participants' losses were symmetrical to the partner's gains. For example, a total of 1000 points were distributed with self:other ratios of 60:40 (600:400 points), 90:10 (900:100 points). Thus, the participant's loss was 900 - 600 = 300 points, which corresponded exactly to the gain of the partner (400 - 100 = 300 points). We used these fixed and symmetrical ratios to minimize unspecific effects of loss aversion.

Each decision trial started with an inter-trial interval indicated by a fixation cross presented for a period jittered between 4000 and 6000 ms (**Figure 2.4.1B**). Next, participants saw the two possible distributions of points in different colors, indicating the potential gain for the participant and the potential gain for the current partner. Participants had to choose one of two distributions within 4000 ms by pressing the left button on a response box to select the distribution on the left side and the right button to select the distribution on the right side. The position of the two allocation options was randomized across trials to minimize response biases due to motor habituation. A green box appeared around the distribution that was selected by the participant at 4000 ms after distribution onset. The box was shown for 1000 ms. At the end of the experiment, two of the distributions chosen by the participant were randomly selected for payment (100 points = 50 cents). Participants performed 60 decision trials in each motive-induction condition, i.e., 120 trials in total. Participants were not informed about the exact number of trials to avoid confounding effects (e.g., counting trials).

Pain stimulator

For pain stimulation, we used electrical stimulation (bipolar, monophasic; output range 5Hz, 0-10 mA) from a single-current stimulator (Neurometer CPT/C; Neurotron Inc.). After attaching the electrodes at the index finger of the right hand and connecting them to the single-current stimulator, the respective person was asked to press the button for defining the current threshold and deciding when she is feeling the stimulation – the value of this threshold was used as painless stimulation. In a second run, the participant was asked to press the same button, but now to hold it pressed until the pain was at an unacceptable level and then to release – this threshold was used for the painful stimulation.

Experimental design and statistical analyses

The aim of our study was to compare prosocial decisions driven by empathy alone with prosocial decisions driven by a combination of empathy and a financial bonus. Therefore, we used a within-subject design in which each participant performed the identical social decision task under two different conditions: the empathy-bonus and the empathy-alone condition. Behavioral data were analyzed with R-Studio Version 1.1.463, R Version 3.6.0 (R Core Team, 2019), and Python (HDDM 0.8.0; Python Version 3.7.6; Jupiter notebook server 6.0.3 (Van Rossum, 2007; Wiecki et al., 2013).

Regression analyses

All regression analyses were performed with the R-packages "stats" (R Core Team, 2019) using "lme4" (Bates, Mächler, Bolker, & Walker, 2015), "car" (Fox & Weisberg, 2019) and “MuMIn“ (Barton, 2019). We used linear models within condition, and mixed models with participants as random effect between conditions as the data have a hierarchical structure that violates the independent assumptions of standard regression models. In the analyses with a continuous variable as dependent variable, linear mixed models were applied. For significant results, the marginal R^2m was calculated using the R-package “MuMin” (Barton, 2019). For the analysis with a dichotomous dependent variable, a logistic mixed model was chosen.

Empathy ratings showed a right-skewed distribution (Shapiro-Wilk W = .94, P < .01), so the data was log-transformed to normal distribution. Pearson correlation was computed between the empathy ratings and the empathic concern scale (EC) from the Interpersonal Reactivity Index (IRI) as well as between the empathy ratings and the personal distress scale (PD) from the (IRI) (Davis, 1980). Results were visualized with the "tidyverse" package (Wickham et al., 2019) and the "ggeffects" package (Lüdecke, 2018). All continuous predictors in our regressions are z-scored.

In addition to the collected data, we also used data from the baseline condition (without motive induction) of a previous study with a similar paradigm and the same assignment task (Hein, Morishima, et al., 2016).

Drift-Diffusion Modelling

We chose the DDM because of its small but trackable number of key parameters and because it is relatively easy to reduce other sequential sampling models (SSMs) to the DDM given specific parameter constraints (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). We used hierarchical drift-diffusion modelling (HDDM) (Vandekerckhove et al., 2011; Wiecki et al., 2013), which is a version of the classical drift-diffusion model that exploits between-subject and within-subject variability using Bayesian parameter estimation methods and thus is ideal for use with relatively small sample sizes. The analyses were conducted using the python implementation of HDDM (Wiecki et al., 2013). Based on previous studies showing changes in drift rate (Aylward et al., 2019; Lerche et al., 2018; Roberts & Hutcherson, 2019; Thompson & Steinbeis, 2021) and the starting point (White et al., 2018), if decisions are made in different affective states, we assumed that these two parameters might also be affected by motivational states. However, given that the modulation of affect and motivation is not the same, effects on the third parameter (the *a*-parameter) are also possible. Therefore, we estimated the full model with *v*, *z*, and *a* possibly being modulated by our two conditions. In addition, we estimated two further models in which both conditions were modulated by the *v*- and *z*-parameters (*a*-parameter estimated across both conditions) or by the *v*- parameter only (*z*- and *a*-parameters estimated across conditions). The best model fit was obtained for a model that allowed for modulation of all three parameters in both conditions (DIC = 4347), followed by a model that allowed variations of the *v*- and *z*-parameters (DIC = 4427), and a model that only allowed a variable *v*-parameter (DIC = 4508). Moreover, we estimated the non-decision parameter (*t0*), which indicates the duration of all extradecisional processes like basic encoding or motor processes (Voss et al., 2004). In paradigms like ours that used an identical experimental setting across conditions, it was recommended to estimate the *t0*-parameter across conditions (Nunez, Vandekerckhove, & Srinivasan, 2017; Servant, Montagnini, & Burle, 2014; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). Following this recommendation, we estimated the t0-parameter across the empathy-bonus and the empathy-alone conditions (mean *t0* = 0.59, SE = 0.02) (see supplementary **Table S2.4.2** for full HDDM results).

We conducted the same DDM analyses with two different inputs. In the first analysis, parameters were estimated based on the standard HDDM condition-wise inputs (reaction time, participants' choices, condition, and participant). In the second analysis, we added the trial-by-trial point difference (self-loss or other-gain) as covariate effecting the drift rate to estimate a hierarchical random intercept model (see Chen & Krajbich (2018) for a similar approach). It should be pointed out that the self-loss and the other-gain were always identical, i.e., the points that were gained by the participant corresponded to the loss of the partner (see section Allocation task for details).

To evaluate the model fit, we conducted posterior predictive checks by comparing the observed data with 500 datasets simulated by our model (a method that has also been recommended for HDDMs) to obtain quantile comparison and 95% credibility (supplementary **Table S2.4.1**) (Wiecki et al., 2013). Moreover, model convergence was checked by visual inspection of the estimation chain of the posteriors, as well as computing the Gelman-Rubin Geweke statistic for convergence (all values < 1.01) (Gelman & Rubin, 1992). Parameters of interest from the model were extracted for further analysis. Specifically, for each participant, the condition-specific *v*-parameters, *z*-parameters, and *a*-parameters were extracted (resulting in 6 parameters per participant). For the parameter comparison, we directly analyzed the posteriors, as recommended by Wiecki et al. (2013). Specifically, we tested the probability of larger *v*-, *z*-, or *a*-parameters are larger in the empathy-bonus compared to the empathy-alone condition. To do so, for each of the three DDM parameters considered, we examined the proportion of posteriors in which the respective parameter is larger for one condition than for the other (Wiecki et al., 2013). A value of 50% corresponds to the chance level, which means that values of over 90% and 95% indicate very high probabilities.

Image Acquisition and Analyses

The experiment was conducted on a 3-T Siemens Magnetom Prisma whole-body MR scanner (Siemens Healthineers), equipped with a one-channel Siemens head coil. Scanner noise was reduced with soft foam earplugs, and head motion was minimized with foam pads. Stimuli presented in the induction phase and in the allocation task were projected onto a rear projection screen located in the front of the scanner. Behavioral responses were recorded with a five-key fiber-optic response box placed on the right hand, and when necessary, vision was corrected using MRI-compatible lenses that matched the dioptre of the participant. Structural image acquisition consisted of 176 T1-weighted transversal images (voxel size of 1 mm). Functional imaging data were collected during the allocation task, using T2*-weighted echo-planar imaging (32 slices, slice thickness of 3 mm, ascending acquisition; repetition time, 2100 ms; echo time, 30 ms; flip angle, 80°; field of view, 240 mm; matrix, 80 × 80). In every decision session, 300 images were acquired - a total of 600 Images for both sessions.

Preprocessing and statistical model

The images were analyzed with SPM12 (Functional Imaging Laboratory, 2019) and Matlab version 8.6 (Matlab Inc, 2015). Images were preprocessed following the standard procedure recommended in the SPM manual (Functional Imaging Laboratory, 2019), including realignment, slice time correction, coregistration, segmentation, normalize, smoothing.

First-level analyses were performed with the general linear model (GLM), using a canonical hemodynamic response function (HRF). For each of the conditions (empathy-alone and empathy-bonus condition), the respective regressors of prosocial choice trials were included as regressors of interest. The prosocial decisions regressor spanned the period from the onset of the decision screen until the participants' reaction (average of 1146.4 ms). Regressors of no interest included the period from the participants' reaction to decision offset (average of 2853.6 ms) and the immediately following period showing the participants' decision (1000 ms).

Sixteen of our participants made less than five selfish decisions in at least one condition. To avoid empty cells in the model, we refrained from computing direct contrasts between prosocial and selfish choices, and selfish choices were included as regressor of no interest.

For the second-level analyses, contrast images for comparisons of interest (empathy-bonus > implicit baseline, empathy-alone > implicit baseline, empathy-bonus > empathy-alone, and empathy-alone > empathy-bonus) were initially computed on a single-subject level. In the next step, the individual images of the main contrast of interest (empathy-bonus > implicit baseline) were regressed against the v-parameter. Results were thresholded using 5% family wise error (FWE) corrected voxel-based inference. To provide insights into larger networks, additionally, we also conducted explorative analyses with p = 0.001 cluster-forming threshold using 5% FWE cluster-based inference (**Table 2.4.1**) and no correction for multiple comparisons (**Table S2.4.5**). Note that peak-coordinates derived from cluster-wise inference only provide information about activated brain components, but not the exact brain region (Eklund et al., 2016; Yeung, 2018). Beta estimates were extracted from the entire clusters of activation in the anterior insula obtained from 5% FWE cluster-based inference with P < .001 cluster-forming threshold, k = 50, using MarsBaR (Matthew Brett et al., 2002). Moreover, we created an independent region of interest based on a recent meta-analysis on empathy for pain experiments (Jauniaux, Khatibi, Rainville, & Jackson, 2019) by creating a 20mm sphere around the reported peak coordinates (x = -43; y = 14; z = 7).

Code and data availability

Behavioral data and scripts are available at github.com (https://github.com/Vassil-Iotzov/empathy_incentives). Imaging data are available at neurovault.org (https://identifiers.org/neurovault.collection:7568).

Results

Empathy was induced with comparable strength in both conditions.

To quantify the strength of the induced empathy, we calculated the participants' trial-by-trial ratings while observing the partner in pain relative to their self-pain ratings. Comparing the ratings between the empathy-alone and the empathy-bonus condition revealed no significant differences between conditions (lmm $\chi^2_{(1)}$ = 0.0001, *P* < .99, *B* = -0.002, *s.e.* = 0.22), indicating that empathy was induced with comparable strength in the empathy-alone and the empathy-bonus condition. To test if the empathy ratings were related to empathic concern, cognitive empathy (perspective taking) or personal distress, we conducted a regression analysis with the individual scores of the empathic concern, the perspective taking and the personal distress subscales of the Interpersonal Reactivity Index (IRI; Davis, 1980) as predictors and the empathy ratings (empathy-alone condition) as dependent variable. The results revealed a significant effect of empathic concern, (*B* = 0.47, *s.e.* = 0.18, *P* = .01, R^2 = .24), but not of personal distress (*B* = -0.23, *s.e.* = 0.18, *P* = .21) and perspective taking (*B* = 0.30, *s.e.* = 0.17, *P* = .10). This finding indicates that the induced motivation, captured by the empathy ratings, mainly reflected empathic concern.

The financial incentive increased the frequency of prosocial decisions, in particular, if empathy was low.

The frequency of prosocial decisions was significantly higher in the empathy-bonus condition (M = 85.65%, s.e. = 0.03%) compared to the empathy-alone condition (*M* = 73.92%, *s.e.* = 0.04%, **Figure 2.4.4A**, lmm $\chi^2_{(1)}$ *= 14.35*, *P* < .01, *B* = -0.57, *s.e.* = 0.15, R^2_m = .08).

Next, we tested whether empathy ratings were related to the probability of prosocial decisions. A logistic mixed model with the participants decisions (prosocial/selfish) as dependent variable and empathy ratings, condition (empathy-alone / empathy-bonus) and empathy ratings × condition as predictors revealed a significant positive effect of empathy ratings (lmm $\chi^2_{(1)}$ = 3.99, *P* = .05, *B* = 0.41, *s.e.* = 0.11), a significant positive effect of condition (lmm $\chi^2_{(1)}$ = 95.10, *P* < .001, *B* = 0.88, *s.e.* = 0.10) and a significant condition x empathy rating interaction (lmm $\chi^2_{(1)}$ = 10.23, *P* = .001, *B* = -0.43, *s.e.* = 0.13; R^2_m = .05; **Figure 2.4.4B**). These results indicate that the probability of prosocial decisions increases with increasing empathy ratings in the empathy-alone condition, but not in the empathy-bonus condition (**Figure 2.4.4B**).

An additional regression analysis with the difference in prosocial decisions (empathy-bonus minus empathy-alone) as dependent variable and empathy ratings as predictor revealed a significant negative relationship (*B* = -0.36, *s.e.* = 0.17, *P* = .05, R^2 = .13). The lower an individual's empathy ratings, the stronger the increase in the frequency of prosocial decisions in the empathy-bonus condition relative to the empathy-alone condition.

Comparing the reaction times of prosocial decisions in the empathy-bonus and the empathy-alone condition revealed no significant difference, (lmm $\chi^2_{(1)}$ *= 2.24*, *P* = .13, *B* = 0.27, *s.e.* = 0.18). There was also no difference when only selfish decisions were considered (lmm $\chi^2_{(1)}$ *= 0.14*, *P* = .71, *B* = -0.08, *s.e.* = 0.22) and when all decisions were included (lmm $\chi^2_{(1)}$ *= 1.99*, *P* = .16, *B* = 0.26, *s.e.* = 0.19).

Furthermore, a linear mixed model with reaction times of the prosocial decisions as dependent variable and empathy ratings, condition (empathy-alone / empathy-bonus) and empathy ratings × condition as predictors was conducted. The results revealed a significant negative effect of empathy ratings (lmm $\chi^2_{(1)}$ = 6.61, *P* = .01, *B* = -0.36, *s.e.* = 0.17), which was comparable in both conditions, condition (lmm $\chi^2_{(1)}$ = 2.17, *P* = .14, *B* = 0.27, *s.e.* = 0.18), condition x empathy rating interaction (lmm $\chi^2_{(1)}$ = 0.02, *P* = .89, *B* = -0.02, *s.e.* = 0.18; R^2_m = .15). According to these results, higher empathy ratings predicted faster prosocial decisions.

As an additional analysis we also compared the number of prosocial decisions in the empathy-alone condition with the number of prosocial decisions in a baseline condition (without any motive induction) from a previous study of our working group using the same allocation task in a similar paradigm (Hein, Morishima, et al., 2016). The results revealed significantly more prosocial decisions in the empathy-alone condition compared to the baseline condition, empathy-alone (*M* = 73.92%, *s.e.* = 0.39), baseline condition (*M* =

49.37%, *s.e.* = 0.32), ($T_{59.963}$ = 4.85, P < .01), suggesting that the induced empathy increase the frequency of prosocial decisions compared to a baseline condition without motive induction.

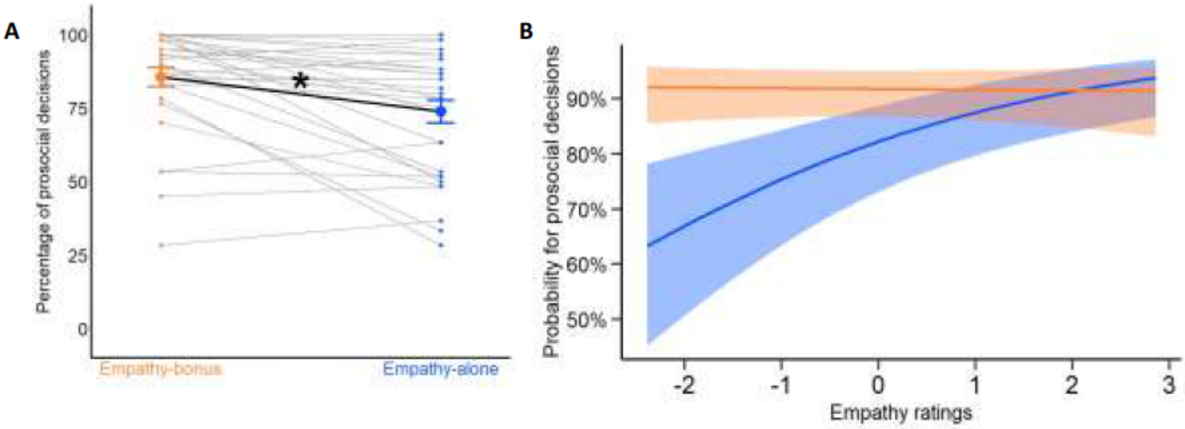


Figure 2.4.4. Individual percentage of prosocial decisions and the relationship between empathy ratings and prosocial decisions in both conditions. **A** Individual percentage of prosocial decisions in the empathy-bonus (orange) and the empathy-alone condition (blue). **B** Positive relationship between the individuals' probability for a prosocial decision and the individuals' empathy ratings in the empathy-alone condition (blue). The higher the participants' empathy ratings, the higher the probability for prosocial decisions. In contrast, in the empathy-bonus condition, the probability for a prosocial decision is not significantly influenced by state empathy.

The financial incentive increased the speed of information accumulation but not the initial decision preference.

To specify which component of the prosocial decision process was enhanced by the financial incentive, relative to prosocial decisions in the empathy-alone condition, we used hierarchical drift-diffusion modeling (HDDM) (Vandekerckhove et al., 2011; Wiecki et al., 2013), a version of the classical drift-diffusion model that exploits between-subject and within-subject variability using Bayesian parameter estimation methods. We estimated the three aforementioned DDM parameters (*v*, *z*, *a*) for every condition and participant. Comparing the observed data with 500 datasets simulated by the HDDM (Wiecki et al., 2013) showed that the HDDM fit the data with 95% credibility (see quantile comparison **Table S2.4.1**).

We compared the speed of information accumulation (drift rate; *v*-parameters), the initial prosocial decision preferences (starting point; *z*-parameters), and the amount of integrated information (*a*-parameters) between the empathy-bonus and the empathy-alone condition.

The comparison of the posteriors (Wiecki et al., 2013) revealed high probability for a larger *v*-parameter in the empathy-bonus condition compared to the empathy-alone condition, *v*-empathy-bonus (*M* = 2.03, *s.e.* = 0.21), *v*-empathy-alone (*M* = 1.25, *s.e.* = 0.19), ($p_{(v\text{-}empathy\text{-}bonus > v\text{-}empathy\text{-}alone)}$ = .99; **Figure 2.4.5A**). In contrast, the probability for a differences between the other decision parameters was relatively low, *z*-empathy-bonus (*M* = 0.47, *s.e.* = 0.01), *z*-empathy-alone (*M* = 0.47, *s.e.* = 0.01; $p_{(z\text{-}empathy\text{-}bonus > z\text{-}empathy\text{-}alone)}$ = .54), *a*-empathy-bonus (*M* = 1.94, *s.e.* = 0.08), *a*-empathy-alone (*M* = 1.85, *s.e.* = 0.09; $p_{(a\text{-}empathy\text{-}bonus > a\text{-}empathy\text{-}alone)}$ = .79). This indicates that financial incentives enhanced the efficiency of the prosocial decision process, while leaving initial prosocial preferences unchanged.

Inspired by previous studies (Chen & Krajbich, 2018; Hutcherson et al., 2015), in an additional analysis, we conducted a model that took the trial-by-trials difference in points for self vs other into account. To do so, we added the point difference (point for self vs points for other) as additional covariate effecting the drift rate (Chen & Krajbich, 2018). The results replicated the observed findings (high probability for a larger *v*-parameter in the empathy-bonus condition compared to the empathy-alone condition: *v*-empathy-bonus (*M* = 5.54, *s.e.* = 0.20), *v*-empathy-alone (*M* = 4.80, *s.e.* = 0.18), $p_{(v\text{-}empathy\text{-}bonus > v\text{-}empathy\text{-}alone)}$ = .99), no differences between the other decision parameters *z*-parameter: *z*-empathy-bonus (*M* = 0.50, *s.e.* = 0.01), *z*-empathy-alone (*M* = 0.49, *s.e.* = 0.01), $p_{(z\text{-}empathy\text{-}bonus > z\text{-}empathy\text{-}alone)}$ = .62; *a*-parameter: *a*-empathy-bonus (*M* = 1.83, *s.e.* = 0.08), *a*-empathy-alone (*M* = 1.75, *s.e.* = 0.08), $p_{(a\text{-}empathy\text{-}bonus > a\text{-}empathy\text{-}alone)}$ = .76).

The incentive-related facilitation of prosocial decisions and individual differences in empathy are associated with changes in anterior insula activation.

On the neural level, the main contrasts between the prosocial decision-related activation in the empathy-bonus vs the empathy-alone conditions revealed significant activation in the right lingual gyrus (BA 18, MNI peak coordinates, *x* = 12, *y* = -94, *z* = -13, *k* = 37, *T* = 5.50, *z* = 4.54).

A second-level regression with the neural activation during prosocial decisions in the empathy-bonus condition and the respective *v*-parameters revealed a significant activation in the left anterior insula (BA 45, MNI peak coordinates, *x* = -27, *y* = 38, *z* = 5, *k* = 107, *T* = 6.32, *z* = 4.97; **Figure 2.4.5B**), indicating that the speed of information accumulation in the empathy-bonus condition is related to a region that has also been associated with individual differences in empathy the processing of empathy (Hein et al., 2010; Lamm et al., 2011; Marsh, 2018). Moreover, the second-level regression revealed a significant activation in the right lingual gyrus (BA 19, MNI peak coordinates, *x* = 24, *y* = -67, *z* = -1, *k* = 401, *T* = 6.03, *z* = 4.82). To provide insights into larger networks, additionally, we also conducted explorative analyses using 5% FWE cluster-based inference (**Table 2.4.1**) and no correction for multiple comparisons (**Table S2.4.5**).

To test whether the neural drift rate signal in AI (**Figure 2.4.5B**) is also affected by empathy, and whether there are differential effects between the empathy-bonus and the empathy-alone condition, we conducted a linear mixed model with the beta estimates of AI activation during prosocial decisions in the empathy-bonus and the empathy-alone condition as a dependent variable.

Table 2.4.1 Neural results of the second-level regression between prosocial decision-related activity in the Empathy-bonus condition and the speed of information accumulation (v-parameter) in the Empathy-bonus condition with P < .001 uncorrected and k > 50. The asterisk indicates activations that are significant at 5% whole-brain FWE voxel-based inference.

Region	**Hemisphere**	**x y z**	**Cluster size**	***t*-value**	**P(FWE$_{\text{cluster-based}}$)**
Anterior Insula	Left	-27 38 5	107	6.32	.004*
	Left	-30 14 -13	62	4.87	.040
Lingual gyrus	Right	24 -67 -1	401	6.03	.000*
Inferior lingual gyrus	Left	-51 -58 -19	189	5.17	.000
Pallidum	Left	-18 -7 -1	70	4.64	.026

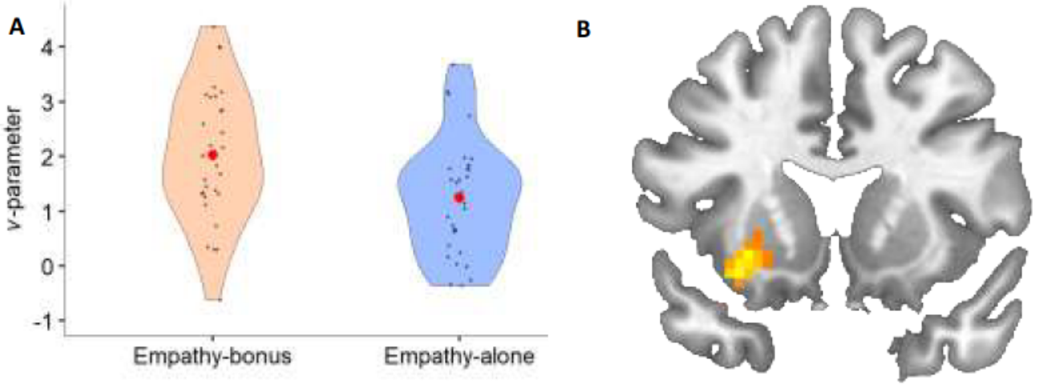


Figure 2.4.5. Distributions of the participants' drift rates (v-parameters) and the related neural response in the anterior insular cortex. **A** Distributions of the participants' drift rates (v-parameters) in the empathy-bonus (orange) and the empathy-alone condition (blue). Red circles represent means, and black dots represent the individual data points. Overall, drift rates are significantly higher in the empathy-bonus condition (orange) compared to the empathy-alone condition (blue). **B** The neural response in the anterior insula (AI) correlated with the individual v-parameters in the empathy-bonus condition (visualized using 5% FWE cluster-based inference with P < .001 cluster-forming threshold; k = 50). The higher the speed of information accumulation in the empathy-bonus condition, the stronger the neural response in AI.

The individual *v*-parameters and empathy ratings were added as predictors, condition (empathy-bonus / empathy-alone) was added as a categorical variable. The results revealed significant main effects of condition (lmm $\chi^2_{(1)}$ = 13.13, *P* < .01, *B* = 0.69, *s.e.* = 0.19), empathy ratings (lmm $\chi^2_{(1)}$ = 4.79, *P* = .03, *B* = 0.34, *s.e.* = 0.16) and the *v*-parameter (lmm $\chi^2_{(1)}$ = 25.60, *P* < .01, *B* = 0.68, *s.e.* = 0.13). Moreover, there were significant interactions between empathy ratings x *v*-parameter (lmm $\chi^2_{(1)}$ = 5.94, *P* = .01, *B* = -0.40, *s.e.* = 0.17), and condition x *v*-parameter (lmm $\chi^2_{(1)}$ = 4.35, *P* = .04, *B* = -0.41, *s.e.* = 0.20), but not between condition x empathy ratings (lmm $\chi^2_{(1)}$ = 0.08, *P* = .78, *B* = 0.07, *s.e.* = 0.21). Finally, the analysis showed a significant condition x *v*-parameter x empathy rating interaction (lmm $\chi^2_{(1)}$ = 11.52, *P* < .01, *B* = 0.72, *s.e.* = 0.21, R^2_m = .50).

To test the effect of cognitive empathy, we added the perspective taking subscale of the IRI (Davis, 1980) as additional predictor. Specifically, we conducted a linear mixed model with the beta estimates of the AI during prosocial decisions in the empathy-bonus and the empathy-alone condition as dependent variables, the individual v-parameters, empathy ratings and the perspective taking subscale of the IRI (Davis, 1980) as predictors and condition (empathy-bonus / empathy-alone) as categorical variable. The results showed no effect of the perspective taking subscale, indicating that cognitive empathy had no significant effect on neuronal activity in the anterior insula (see supplementary **Table S2.4.3** for full results).

The same analysis with beta estimates from the right lingual gyrus, i.e., the other region that was significantly related to individual differences in the *v*-parameter (see second-level regression results above), revealed no significant results, except a significant main effect the *v*-parameter (lmm $\chi^2_{(1)}$ = 12.31, *P* < .01, *B* = 0.60, *s.e.* = 0.17) and a significant condition x *v*-parameter interaction (lmm $\chi^2_{(1)}$ = 9.54, *P* < .01, *B* = -0.80, *s.e.* = 0.26), indicating that the three-way interaction between condition, empathy ratings, and the *v*-parameters was specific for left AI (for full results see **Table S2.4.4**).

To unpack the significant condition x *v*-parameter x empathy rating interaction in left AI, we tested the relationship between the *v*-parameter and the empathy ratings separately in the empathy-alone and the empathy-bonus condition. We found a significant negative empathy ratings x *v*-parameter interaction in the empathy-bonus condition (*B* = -0.37, *s.e.* = 0.14, *P* = .01), with significant main effects of *v* (*B* = 0.70, *s.e.* = 0.11, *P* < .01) and empathy ratings (*B* = 0.31, *s.e.* = 0.13, *P* = .03, R^2 = .65; **Figure 2.4.6A**). The results for the empathy-alone condition revealed a marginal significant positive empathy x *v*-parameter interaction (*B* = 0.31, *s.e.* = 0.16, *P* = .06) with a significant main effect of the empathy ratings (*B* = 0.42, *s.e.* = 0.18 *P* = .03) and no main effect of the *v*-parameter (*B* = 0.24, *s.e.* = 0.18, *P* = .19; R^2 = .32, **Figure 2.4.6B**).

To further unpack these two-way interactions, we tested the relationship between the *v*-parameter and anterior insula (AI) beta estimates, as well as the relationship between empathy ratings and AI beta estimates separately in the empathy-bonus and the empathy-alone condition. Given that empathy facilitates prosocial decisions (Batson et al., 1995; Decety et al., 2016) and correlates with neural responses in AI cortex, we assumed a positive relationship between the empathy ratings and the drift rate and between the empathy ratings and AI activation, and used one-sided tests to test these assumptions (Pfaffenberger & Patterson, 1977; Ruxton & Neuhäuser, 2010). In the empathy-alone condition, the results revealed significant positive relationships between empathy ratings and AI beta estimates (*B* = 0.43, *s.e.* = 0.18, *P* = .01, **Figure 2.4.6D**), empathy ratings and drift rate (*B* = 0.30, *s.e.* = 0.18, *P* = .05), and *v*-parameter and AI beta estimates (*B* = 0.38, *s.e.* = 0.19, *P* = .03, **Figure**

2.4.6F). In the empathy-bonus condition we observed a significant positive relationship between *v*-parameter and AI beta estimates (*B* = 0.73, *s.e.* = 0.12, *P* < .01, **Figure 2.4.6E**), while the relationships between empathy ratings and AI beta estimates (*B* = 0.23, *s.e.* = 0.16, *P* = .08, **Figure 2.4.6C**) and between empathy ratings and drift rate were not significant (*B* = 0.21, *s.e.* = 0.17, *P* = .11).

These subsequent analyses revealed a significant positive relationship between empathy ratings and neural responses in AI and between empathy ratings and drift rate in the empathy-alone condition. In the presence of a financial incentive in the empathy-bonus condition, these effects were no longer significant. Interestingly, the interaction between the empathy ratings and the drift rate reduced AI activation in the empathy-bonus condition while increasing it in the empathy-alone condition. This indicates that in the empathy-bonus condition, the empathy ratings (indicating the strength of the empathy motive before the bonus was offered) suppress the positive effect of the *v*-parameter on the neural response in AI.

To test the robustness of the differential effects in the empathy-bonus and the empathy-alone conditions, we extracted the beta-estimates of prosocial decision-related activation in the empathy-bonus and the empathy-alone condition from an independent region of interest in the AI (defined based on the peak coordinates reported in a recent meta-analysis on empathy of pain studies (Jauniaux et al., 2019). We conducted a linear mixed model with these beta-estimates as dependent variable, and condition (empathy-bonus / empathy – alone), empathy ratings, and *v*-parameters as predictors. The results replicated the significant condition x *v*-parameter x empathy rating interaction reported above (lmm $\chi^2_{(1)}$ = 5.97, *P* = .01, *B* = 0.62, *s.e.* = 0.25, R^2_m = .20), reflecting a negative relationship in the empathy-bonus condition and a positive relationship in the empathy-alone condition.

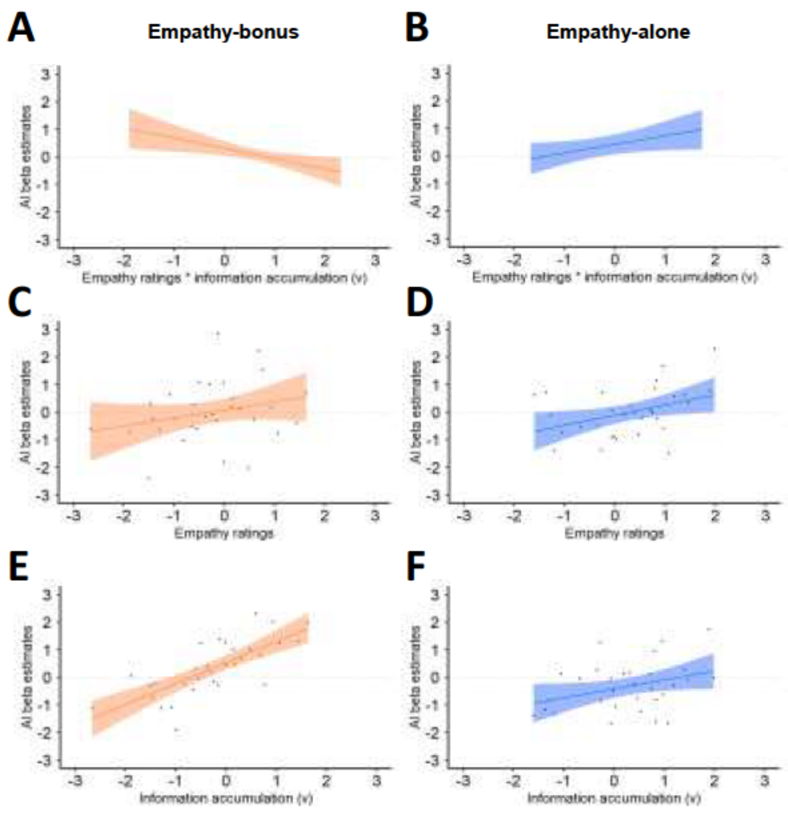


Figure 2.4.6. Relationships between anterior insula (AI) beta estimates and empathy ratings and between AI beta estimates and speed of information processing (v-parameter) in the empathy-bonus and empathy-alone conditions. The beta estimates reflect the average of AI activation from the empathy-bonus and the empathy-alone condition, extracted from the same AI clusters that correlated with the v-parameter in the empathy-bonus condition (shown in **Figure 2.4.5**). **A** Effect of the empathy ratings x v-parameters interaction on AI responses in the empathy-bonus condition. **B** Effect of the empathy ratings x v-parameters interaction on AI responses in the empathy-alone condition. **C** The relationship between the individual strength of the AI responses and the individual empathy ratings in the empathy-bonus condition was not significant. **D** Significant positive relationship between the individual strength of the AI responses and the individual empathy ratings in the empathy-alone condition. **E** Significant positive relationship between the individual strength of the AI responses and the speed of information processing (v-parameter) in the empathy-bonus condition. **F** Significant positive relationship between the individual strength of the AI responses and the speed of information processing (v-parameter) in the empathy-alone condition.

Discussion

Our study investigated how financial incentives affect empathy-related prosocial decisions. The results show that on average financial incentives increase the frequency of prosocial decisions (**Figure 2.4.4**), especially the lower an individual scored on empathy. The finding that the financial bonus enhanced the frequency of prosocial decisions is in line with previous studies showing an incentive-related increase in prosocial behaviors (Balliet et al., 2011; Stoop et al., 2018). Extending this previous evidence, our results reveal that this effect is modulated by individual differences in empathy, i.e., the effect is the stronger, the lower a person’s state empathy. Besides providing insights into the interplay between financial incentives and empathy, our results specified how financial incentives affect the prosocial decision process. The results of drift-diffusion modelling showed that the financial incentive enhanced the efficiency (i.e., speed of information accumulation captured by the v-parameter) of prosocial decisions in the empathy-bonus compared to the empathy-alone condition (**Figure 2.4.5A**). In contrast, the incentive had no significant effect on participants' initial prosocial preferences, i.e., the preference of making a selfish or prosocial decision with which they entered the decision process (captured by the z-parameter).

Outside the domain of prosocial decisions, there is evidence that the efficiency of decisions (captured by the v-parameter) is affected by individual differences in emotions (Aylward et al., 2019; Lerche et al., 2018; Roberts & Hutcherson, 2019; Thompson & Steinbeis, 2021). For example, according to the results of Thompson and Steinbeis (2021), individuals with greater state anxiety show an increased v-parameter on fearful face trials. Extending these findings, our results reveal that the speed of information accumulation is shaped by the motivation that drives participants' prosocial decisions, i.e., higher if a prosocial decision is rewarded than if it is only based on empathy.

On the neural level, the incentive-related facilitation of the prosocial decision process was strongest related to the participants' neural response in the left anterior insula (AI; **Figure 2.4.5B**), in line with previous evidence that associated the individual strength of AI responses and individual differences in drift rate (Gluth et al., 2012; Pedersen et al., 2015). Importantly, the neural response in the same AI region was also related to individual differences in empathy ratings, supporting a link between anterior insula activity and empathy (Hein, Engelmann, et al., 2016; Hein, Morishima, et al., 2016; Lamm et al., 2011; Marsh, 2018; Masten, Eisenberger, Pfeifer, & Dapretto, 2011) as well as the propensity for prosocial decisions (Hein, Engelmann, et al., 2016; Hein, Morishima, et al., 2016; Lamm et al., 2011; Marsh, 2018; Masten, Eisenberger, et al., 2011).

Adding a novel aspect, our findings reveal how financial incentives alter the effect of empathy on the computation of prosocial decisions in the anterior insular cortex. After offering a bonus in the empathy-bonus condition, the relationship between empathy ratings and drift rate and empathy ratings and AI estimates was no longer significant, indicating that in the presence of an incentive, empathy was no longer a significant driver of prosocial decisions. Interestingly, the interaction between the empathy ratings and the drift rate significantly reduced AI activation in the empathy-bonus condition (**Figure 2.4.6A**) while increasing it in the empathy-alone condition (**Figure 2.4.6B**). This indicates that in the empathy-bonus condition, the strength of the empathy motive (captured by the individual strength of the empathy ratings before the bonus was offered) suppressed the positive relationship between information accumulation during prosocial decisions and the neural response in AI. Together, these findings indicate that the anterior insula integrates self-regarding (gaining the financial incentive) and other-regarding (empathy with the other person) motives that both elicit prosocial decisions and thus forms a plausible neural basis for the impact of financial incentives on empathic motivation. They support the assumptions of influential motivation theories (Batson, 1994; Batson et al., 2011; Engel & Zhurakhovska, 2016; Hughes & Zaki, 2015; Saulin et al., 2022) which assume that most complex decisions are driven by an interaction between different motives, here an egoistic motive incited by a bonus for prosocial decisions, and the empathy motive. Adding to this theoretical framework, our results show that the extent of the motive interaction depend on individual state empathy and is captured by changes in neural responses in AI cortex.

Besides the AI, we hypothesized that the incentive-related increase in the v-parameter in the empathy-bonus compared to the empathy-alone condition may also increase the neural activation in the ACC/MCC, medial prefrontal, temporo-parietal and striatal regions, i.e., regions that have been associated with emotional and cognitive empathy and prosocial decision-making in general. Exploratory analyses on a lower threshold ($p_{uncorrected} < 0.001$) indeed showed that the efficiency of prosocial decisions (captured by individual differences in the v-parameter) is also related to changes in activation in medial prefrontal and striatal regions (**Table S2.4.5**). However, none of these regions showed the interaction between the *v*-parameter and empathy ratings observed in the AI.

The AI has mainly been linked to emotional aspects of empathy (Dvash & Shamay-Tsoory, 2014; Fan et al., 2011; Preckel et al., 2018; Schurz et al., 2021; Stietz et al., 2019). However, given evidence that many prosocial decisions are driven by both, cognitive and emotional empathy (Kanske et al., 2015; Preckel et al., 2018; Stietz et al., 2019; Zaki & Ochsner, 2012), it is still possible that cognitive empathy processes also play a role. Supporting this view, on a lower threshold, the efficiency of the prosocial decision process was also captured by medial prefrontal brain regions that have been associated with cognitive empathy (Dvash & Shamay-Tsoory, 2014; Preckel et al., 2018; Schurz et al., 2021; for results see **Table S2.4.5**). However, individual differences in cognitive empathy (measured by the perspective taking subscale by the IRI, Davis, 1980) did not modulate participants' empathy ratings and the interaction effects observed in AI (**Table S2.4.3**). These results indicate that cognitive empathy may have influenced the efficiency of the prosocial decision process but did not significantly alter the interplay between self-regarding (gaining the financial incentive) and other-regarding (empathy with the other person) motives that was observed in AI cortex.

In our study, empathy was conceptualized as a motive that can drive prosocial decisions. And indeed, the empathy ratings of our participants that correlated with empathic concern (but not with personal distress and perspective taking) facilitated the prosocial decision process in the empathy-alone condition, in line with previous findings (Batson et al., 1995; Decety et al., 2016). That said, the effect that financial incentives counteracted the facilitating effect of empathy on prosocial decisions the stronger the higher participants' state empathy, might indicate that participants with higher state empathy are less motivated to empathize in the presence of an incentive, an assumption that supports the notion that empathy itself is a motivated state (Zaki, 2014).

In the present study, the financial incentive for prosocial decisions was offered in private, and self-image concerns were reduced as far as possible, at least with regard to public reputation. However, some participants nevertheless showed an incentive-related decline in prosocial decisions (**Figure 2.4.4A**). It is conceivable that participants scoring higher on state empathy feel insulted by the bonus because "being paid to be nice" undermined their intrinsic empathic motivation that otherwise (i.e., in the empathy-alone condition) drives their prosocial decisions. Thus, although on average, our findings show that the incentive increased the frequency of prosocial decisions compared to an empathy-alone condition, it is still possible that it undermines prosocial behavior in highly empathic participants. To test this assumption, future studies should test the effect of financial incentives on empathy-based decisions in extreme groups, i.e., groups of extremely high or low empathic individuals. Moreover, it would be interesting to use a trial-by-trial bonus manipulation that allows for modeling the effect directly as part of the DDM.

In our study, we motivated prosocial behavior by empathy, known to be one of the strongest drivers of prosocial behavior (Batson, 1994; Decety et al., 2016) and, in the other condition, additionally offered a bonus for prosocial behavior. As expected, these experimental manipulations resulted in a high number of prosocial decisions. This raises the question of whether our results may be affected by ceiling effects. Addressing ceiling effects in DDM modeling, previous work has shown that the estimation of DDM parameters is robust even if participants achieve near-ceiling accuracy (over 90% correct answers) (Ratcliff & McKoon, 2008). Particularly when accuracy is at ceiling for one but not for all conditions of an experiment, the other conditions provide the error responses required for the model to estimate the variability in drift rate and starting point over the entire experiment (Ratcliff, 2014). In light of this evidence, it is unlikely that the estimation of the drift rate is strongly affected by ceiling effects. Moreover, the empathy ratings were collected during the empathy induction prior to the choice task and thus are also not affected by ceiling. Given that our main results are based on the interaction between empathy ratings and the v-parameter, it is unlikely that these findings reflect ceiling effects.

Given evidence for different allocation patterns toward a partner from the same as compared to the opposite sex (Eckel & Grossman, 1998; Saad & Gill, 2001), and for gender differences in empathy (Christov-Moore et al., 2014) and prosocial behavior (Chowdhury et al., 2017), we recruited participants in their early twenties from the same gender (female) that were paired with a partner from the same gender (confederates). Testing participants from the same gender and age group allowed us to control for unspecific gender and age effects. The current sample size of 31 participants is large enough to detect effects of incentive on prosocial decisions, comparing two dependent means, and, because of the complex setup (involving confederates) it was difficult to test a large sample. The imaging results obtained with the current sample were thresholded using 5% family wise error (FWE) correction, i.e., the method that is recommended for a reliable correction of multiple comparisons in our field (Han & Glenn, 2018). Multilevel models were used to minimize multiple comparisons (Gelman, Hill, & Yajima, 2012). However, we acknowledge that our results are based on a rather small female sample with specific demographic characteristics (e.g., a certain age group with high education) which limits their generalizability. Future studies are required to replicate our results in male participants, and larger, more diverse samples, i.e., individuals from different age groups and educational backgrounds.

In summary, our current results indicate that financial incentives offered in private facilitate prosocial decisions more the lower participants scored on state empathy.

Authors' Contributions

Iotzov, Vassil: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration; *Saulin, Anne*: Methodology, Software, Validation, Resources, Writing - Review & Editing; *Kaiser, Jochen*: Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Project administration; *Han, Shihui*: Writing - Review & Editing; *Hein, Grit*: Conceptualization, Methodology, Software, Validation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

We thank Isabelle Ehrlich for help with data collection and Marthe Gründahl and Martin Weiß for feedback on the manuscript. This work was supported by the German Research Foundation (HE 4566/2-1; HE 4566/5-1).

3 General discussion

3.1 Summary

The key research question addressed in the studies conducted in the context of this dissertation was how sustainably empathy induces social closeness and prosocial behavior in of itself as well as in comparison to and combined with other social motives.

In study 1, we examined the formation and persistence of empathy-related social closeness by adopting a reinforcement learning (RL) approach. Specifically, we used an acquisition-extinction paradigm in which empathy was reinforced by participants observing an ostensible other participant (confederate of the experimenter) receive painful stimulation. In a first block, empathy was frequently reinforced (acquisition phase) and in a second block only rarely reinforced (extinction phase). Results showed that empathy-related social closeness increased during acquisition and persisted during extinction. Using the Rescorla-Wagner reinforcement learning model, we could show that empathy-related social closeness formation and sustainability were characterized by a large recalibration of the learning signal (the so-called prediction error). From an RL perspective, this indicates that those trials serving as non-reinforcer trials were able to elicit positive prediction errors, which in turn lead to an increase as opposed to a decrease in social closeness after non-reinforced trial. In an independent sample, we replicated these behavioral and computational modelling results. On a neural level, the extent of recalibration modulated how sensitive neural activation in the superior temporal sulcus (STS), the temporo-parietal junction (TPJ), and the inferior frontal gyrus (IFG) extending into anterior insula (AI) was to the observation of another's painful stimulation compared to another's non-painful stimulation. In the acquisition phase, stronger activation in IFG/AI in response to another's non-pain was associated with increased social closeness. In the extinction phase, however, stronger activation in response to another's pain was associated with increased social closeness, especially for highly empathic individuals. In contrast to empathy-related social closeness, reciprocity-related social closeness was not sustainable but starkly decreased in the extinction phase. In line with this pattern, computational modelling showed that a simple learning rule can well describe the reciprocity-related behavior for a large portion of the participants.

In study 2, we investigated the sustainability of empathy-driven prosocial behavior. That is, we tested how the activation strength of empathy influences specific components of the empathy-driven prosocial decision process and their neural correlates as assessed using fMRI. To this end, participants performed two conditions, each comprising three blocks of the social decision task with an alleged other participant (confederates of the experimenter). In the treatment condition, the first block served as baseline measurement and was followed by strong activation of empathy by frequent observation of the partner's painful stimulation, which should lead to an increase in empathy motive strength. After the second social decision block (initial response block), empathy was only rarely activated, which in turn should lead to a decrease of empathy motive strength and thus a lower frequency of prosocial decisions in the final social decision block (sustained response block). In a parallel control condition, participants performed the same number of blocks but with activation frequency at chance level in the two motive activation phases. Results showed that the frequency of prosocial decisions driven by empathy was not significantly modulated by block number or by condition across two independent samples. Drift-diffusion modelling (DDM) however, revealed that the empathy-driven social decision process was sensitive to the different activation strengths. That is, the initial bias towards making a prosocial decision as opposed to an egoistic decision was increased after initial strong activation and remained on this level after subsequent rare activation. Hence, before participants started considering the different point options the faced in each trial of the social decision task, they already preferred the prosocial option more strongly after frequent activation as compared to the baseline block. On a neural level, we observed increasing neural activation with increasing block number in regions associated with social cognition, including the AI, the TPJ, the striatum, and IFG. Together, these results support empathy-related sustainability also with respect to prosocial behavior. In an additional experiment, we tested whether reciprocity-driven behavior was equally sustainable. However, we observed that while the frequency of reciprocity-driven prosocial decisions and the initial bias towards making prosocial decisions significantly increased after frequent activation, both indicators for prosocial behavior decreased again after rare activation. Hence, in contrast to empathy, reciprocity did not lead to sustainable prosocial decision behavior.

In study 3, we investigated the relative influence of empathy and reciprocity on the prosocial decision process by comparing prosocial decisions driven by the combination of empathy and reciprocity with prosocial decisions driven by each motive separately and a baseline without any motive induction. Before the decision task, the motives were induced towards different interaction partners (confederates of the experimenter), corresponding to the four experimental conditions: the empathy motive condition, the reciprocity motive condition, the multi-motive condition (i.e., empathy + reciprocity motive), and the baseline condition (i.e., no motive actively induced). Results showed that compared to the baseline condition, participants more frequently chose the prosocial decision option when the motives were actively induced. Decisively, participants chose the prosocial option more frequently in the multi-motive condition compared to the reciprocity condition. Thus, adding the empathy motive to the reciprocity motive increased the frequency of prosocial decisions, but not vice versa. Using DDM analyses, we observed that the combination of empathy and reciprocity increased the initial bias towards making a prosocial decision compared to the activation of reciprocity only. Neural second-level regression analyses revealed that neural activation in bilateral dorsal striatum increased the more, the more an individual's initial bias was increased in the multi-motive condition compared to the reciprocity condition. Interestingly, these findings could not be explained by mere dominance of the empathy motive compared to the reciprocity motive. That is, empathy did not simply overrule reciprocity when the motives were combined. Instead, differences between the reciprocity-driven and the combination-driven prosocial decision process was specific to the prosocial decision process driven by a complex motivational state, i.e., that is empathy as well as reciprocity, in contrast to a simple motivational state, i.e., only empathy or only reciprocity. Nonetheless, the results suggest that empathy is the stronger motive which other motives can benefit from with respect to increasing prosocial decision behavior, but not vice versa.

In study 4, we investigated the combination of empathy with the motive of outcome maximization. Results showed that when participants were offered an additional payout for making prosocial decisions, i.e., when the motive of outcome maximization was added to the empathy-driven social decision process, people made prosocial decisions more frequently. Additionally, the prosocial decision process was more efficient as indicated by an increased speed of evidence accumulation (DDM drift-rate parameter). The higher a participant's speed of evidence accumulation was when prosocial decisions were based on outcome maximization and empathy, the larger the neural activation in AI, a region that was also associated with state empathy. This effect was particularly strong for low empathic participants.

The four studies conducted as part of this dissertation demonstrated that empathy, the sharing of another's affective state, (i) lead to sustainable social closeness and prosocial decision behavior, (ii) was beneficial in combination with other social motives, and (iii) could not be easily undermined by the motive of outcome maximization. In the following sections, these findings will be discussed in light of existing works and directions for future research inspired by these findings will be put forward.

3.2 Implications

Observable empathy sustainability and stability

We investigated the question of empathy-related social closeness sustainability within the framework of reinforcement learning (study 1). We had hypothesized that the more sustainably empathy induces social closeness, the less decrease ratings of social closeness should show in the extinction block, i.e., the block with only rare reinforcement of the empathy motive. Analogously, we had hypothesized that the more sustainable the empathy-related prosocial decision behavior, the weaker the decrease in the frequency of prosocial decisions after only weak empathy activation (study 2). For both measures we observed sustainable empathy-related social behavior in two independent samples, respectively. That is, we observed no decrease in social closeness during extinction and no decrease in observable prosocial decision-making after weak empathy activation. However, for reciprocity, we observed a pattern in line with unsustainable social closeness and prosocial behavior, i.e., a decrease in social closeness during extinction and a decrease in the frequency of prosocial decisions after weak reciprocity activation. Thus, empathy led to more sustainable social behavior than reciprocity.

These findings may be understood in terms of the behavioral predictions that arise from the different goals elicited by empathy and reciprocity, respectively. While prosocial behavior based on empathy is linked to increasing the well-being of another person (Batson, 2010), prosocial behavior based on reciprocity is linked to repaying a favor (Gouldner, 1960; McCabe et al., 2003). Hence, when a previously received favor is fully repaid, prosocial behavior based on reciprocity on the one hand should deteriorate, especially when the other person stops paying favors (Fehr & Gächter, 2000; Gouldner, 1960; Nowak, 2006). Prosocial behavior based on empathy on the other hand may not deteriorate once the other person experiences more frequent non-painful stimulation, since this more positive experience does not weigh out the negative prior experience. Instead, observing non-pain after previous frequent observed pain may even further boost social closeness and prosocial behavior based on positive empathy (Andreychik, 2019; Morelli et al., 2014; Telle & Pfister, 2016). Thus, both types of trials observed can potentially promote social closeness and prosocial behavior. For reciprocity, however, observing that the other person had decided not to help can have the opposite effect and activate negative reciprocity (Chernyak et al., 2019; Kaltwasser et al., 2016). Instead of promoting social closeness and prosocial behavior, observing non-helping behavior could thus impede these social behaviors, which would contribute to the lack of reciprocity-related social behavior sustainability observed in study 2.

These first results shed light on the sustainability of empathy-related social behavior alone and in comparison with reciprocity-related social behavior. In the other two studies, we approached the question of empathy sustainability from a different angle by combining empathy with other social motives. We explicitly tested how the combination of empathy with reciprocity and the combination of empathy with the motive of outcome maximization shaped the prosocial decision process. We hypothesized that if empathy was a sustainable motive with respect to motive combinations, empathy should boost prosocial behavior based on other motives and should not be undermined by additional motives. The results showed that when prosocial decision were based on both motives, participants made prosocial decisions more frequently compared to the situation in which participants decided only based on reciprocity, but not compared to the situation in which participants decided only based on empathy (study 3). Adding empathy to reciprocity hence boosted prosocial behavior but not vice versa. Moreover, empathy-based prosocial behavior was not undermined but in fact boosted by the addition of outcome maximization (study 4).

This latter findings is surprising as it has frequently been suggested that offering monetary incentives, which is the operationalization of the motive of outcome maximization, undermines intrinsic prosocial motivation (e.g., Frey & Jegen, 2001; Promberger & Marteau, 2013). Titmuss prominently claimed that paying people for donating blood will decrease their likelihood to actually do it by undermining the initial intrinsic prosocial motive to help others (Titmuss, 1970). However, a later meta-analysis did not support an undermining effect of monetary incentives on blood donation, but rather suggested that paying or not paying people for donating blood does not influence their likelihood to do so (Niza et al., 2013). In line with other studies (meta-analysis: Balliet et al., 2011; review: Besley & Ghatak, 2018), the findings observed here also do not support a general undermining effect of an egoistic motive on empathy-based prosocial behavior. In fact, participants in our study made more prosocial decisions when they were additionally offered a monetary bonus for behaving prosocially. Hence, the motive of outcome maximization did not undermine observable empathy-based prosocial behavior adding to the notion of empathy being a sustainable and stable driver for social closeness and prosocial behavior.

Taken together, on the level of directly observable behavior as indicated by ratings of social closeness and social decision-making, empathy sustainably increased social closeness and prosocial decisions. In the following section, the computational mechanisms underlying empathy sustainability and stability as assessed in this dissertation are discussed in detail.

Computational mechanisms underlying empathy sustainability

The sustainability and stability of empathy-based social closeness and prosocial decision-making can be directly observed, indicated by increased and more stable social behavior, but it can also be characterized in terms of the computational mechanisms underlying the respective behavior. In this dissertation, I aimed at uncovering these mechanisms using reinforcement learning models (Rescorla & Wagner, 1972) to capture the temporal evolution of empathy-related social closeness (study 1) and the drift-diffusion model (Ratcliff & McKoon, 2008; Vandekerckhove et al., 2011; Wiecki et al., 2013) to better understand the empathy-related social decision process (studies 2-4).

Results of study 1 revealed that the empathy-based development of social closeness over time can be described in terms of a reinforcement learning process that allows for individual recalibration of the learning signal (cf. Bavard, Lebreton, Khamassi, Coricelli, & Palminteri, 2018). That is, observing another's pain was not generally associated with a learning signal corresponding to a value of 1 which always led to an increase in social closeness, and observing another's non-pain was not generally associated with a learning signal of 0 which always lead to a decrease in social closeness (as assumed in the simple RW learning model). Rather, these values were individually adjusted resulting in a value of smaller than 1 for reinforced trials, i.e., trial of observed pain, and values of larger than 0 for non-reinforced trials, i.e., trials of observed non-pain. This implies that individuals will increase their social closeness not only on reinforced trial, but also based on non-reinforced trials as these trials can yield a positive learning signal. Interestingly, follow-up analyses showed that the extent of recalibration differed between the block of frequent empathy reinforcement (acquisition) and the subsequent block of rare empathy reinforcement (extinction). Specifically, in the acquisition block, participants on average did not strongly recalibrate the feedback signal. Thus, observing the other person in pain entailed an increase of social closeness. In the extinction block, however, participants on average strongly recalibrated the feedback signal. Thus, observing that the other person received non-painful stimulation increased social closeness to a comparable extent to which observing the other person receive painful stimulation increased social closeness (see supplementary materials of study 1 for details). This computational account demonstrates that participants can switch the basis on which they are learning to feel closer to the other person. In detail, they do so by putting more value on seeing that the other person receives no pain once the context changes from frequent to rare incidences of painful stimulation. Previous works have shown that positive as well as negative empathy can lead to connectedness and prosocial behavior towards the other person (Andreychik, 2019; Andreychik & Migliaccio, 2015; Depow, Francis, & Inzlicht, 2021; Morelli, Lieberman, et al., 2015; Shiota, Papies, Preston, & Sauter, 2021; Telle & Pfister, 2016). The results from this dissertation are hence in line with these studies and additionally demonstrate that individuals can switch from one type of empathy to the other within one experimental session. Generally, in the framework of reinforcement learning, a context-dependent adaptation of the learning process is reported more and more frequently (Fontanesi, Palminteri, & Lebreton, 2019; Hunter & Daw, 2021; Palminteri et al., 2015; Pischedda et al., 2020; Stojić, Schulz, P Analytis, & Speekenbrink, 2020). Hunter & Daw (2021) for example highlight that the uncertainty of reward in a given environment shapes the learning process. Other works demonstrated that contextual information such as offering information about the outcome associated with the option *not* chosen by the participant, termed counter-factual information, influences the learning process (e.g., Pischedda et al., 2020). This dissertation could show that this principle of context-sensitive learning extends to the learning of motive-driven social closeness. Specifically, we showed that this principle holds with respect to who is learning (see section *Inter-individual differences*) as well as which motive the learning process was based on. That is, results showed that in contrast to empathy-related social closeness, the temporal evolution of reciprocity-related social closeness was similarly likely to be explained by the very basic RW learning rule, i.e., without assuming individual recalibration, as by the model variant which included individual recalibration. As in basic reinforcement learning, social closeness deteriorated once reciprocity was only rarely reinforced (Bouton, 2004; Shiban et al., 2015). The computational account of reciprocity-based social closeness suggests an inflexibility of the underlying mechanisms of reciprocity-based social closeness in contrast to empathy-based social closeness. Here, this inflexibility led to a decrease in social closeness. Although, this inflexibility entailed unsustainable social closeness in the present paradigm, unambiguity with respect to a social norm may actually facilitate its application in daily life. That is, if the outcome value of observed helping behavior corresponds to the value of 1 and the non-helping behavior to a value of 0 for a large portion of the population observed, there is less room for uncertainty. This in turn simplifies the use of reciprocity as a social norm that corresponds to a simple heuristic (Rand et al., 2014), facilitating prosocial behavior on a societal level (Bartlett & DeSteno, 2006; Gouldner, 1960; Nowak, 2006; Orhun, 2018; Penner, Dovidio, Piliavin, & Schroeder, 2005).

In this dissertation, empathy did not only induce sustainable social closeness but activation of empathy also led a sustained prosocial decision bias (study 2) and boosted the reciprocity-based prosocial decision process (study 3). DDM analysis revealed that strong activation of empathy increased participants' initial bias towards making a prosocial decision compared to prior baseline behavior. This result is in line with previous works which observed that changes in the motivation to behave prosocially were reflected in changes of the initial bias towards the prosocial (in contrast to the egoistic) decision option (Chen & Krajbich, 2018; Gallotti & Grujić, 2019; Yu et al., 2021). For empathy-based prosocial behavior, this bias remained high even after weak activation of empathy. Although the other person only rarely received painful stimulation, the motivation to act prosocially towards that person remained high. For reciprocity, however, this bias decreased after weak motive activation, indicating decreasing motivation to act prosocially towards that person. These results showed that when activated separately, empathy induces a more sustainable prosocial decision bias than reciprocity.

Moreover, combining empathy with reciprocity increased the prosocial decision bias compared to when only reciprocity was activated (study 3). Thus, empathy sustainability in terms of relative motive effectiveness regarding prosocial decision-making was again reflected in the changes of this initial prosocial decision bias. Previous works have suggested that the empathy motive may not be necessary once a social norm prescribing a certain behavior is active that can elicit prosocial behavior (Lay, Zagefka, González, Álvarez, & Valdenegro, 2020). Another study even showed that empathy may be actively avoided due to its high cognitive cost (Cameron et al., 2019). This view suggests a “taking over” of the cognitively less costly motive (here, the reciprocity motive) once both motives are active. In contrast, the results obtained in this dissertation show that empathy can boost the reciprocity-based social decision process whereas reciprocity cannot boost the empathy-based social decision process. These results are in keeping with other studies, that have found stronger effects of empathy compared to reciprocity on hypothetical helping behavior (Allsop, Fifield, & Seiter, 2002) or a sustaining effect of empathic reactions on the upholding of a (reciprocal) relationship (Rumble, Van Lange, & Parks, 2010). Based on studies in primates, Yamamoto & Takimoto (2012) further suggested that reciprocity as a fairness norm may stabilize prosocial behavior based on empathy but may not act as a promoter of prosocial behavior itself. Together, these studies highlight the interactional nature of empathy and reciprocity in maintaining of positive social relationships and support the idea of empathy as an active promoter of prosocial behavior.

In daily life, empathy cannot only interact with the reciprocity motive, but also with other motives, such as the motive of outcome maximization (Batson et al., 2004; Cory, 2006). We hence tested how the addition of financial incentives for prosocial behavior altered the empathy-based social decision process (study 3). DDM results revealed that the efficiency of the choice process itself, rather than the initial bias was increased when participants’ social decision behavior was based on outcome maximization was well as empathy compared to empathy only. Hence, knowing that you are additionally paid a bonus makes your prosocial decision process more efficient as compared to when your decision is only based on empathy. In keeping with empathy as a sustainable social motive and in contrast to a potential undermining effect of financial incentives (Promberger & Marteau, 2013; Takeuchi et al., 2015; Titmuss, 1970), the empathy-based prosocial decision process benefited from the addition of this motive. However, in contrast to studies 2 and 3, this beneficial effect was reflected in the efficiency of the choice process and not the initial prosocial bias. Given that different components of the decision process are affected in studies 2 and 3 compared to study 4, different mechanism appear to underlie the influence of empathy combined with reciprocity compared to empathy combined with a financial incentive on the social decision process (Voss et al., 2004). Variability in the speed of evidence accumulation, has primarily been linked to task difficulty and conflict with higher task difficulty and conflict being associated with a lower speed of evidence accumulation (Dambacher & Hübner, 2015; Schuch & Pütz, 2021; Servant et al., 2014; Voss et al., 2004). During the decisions participants face in the dictator game, they have to weigh the self-interest of maximizing their own outcome against the other-regarding preference to maximize the other's outcome (Hu et al., 2017). These two interests are in a conflict (Fehr & Camerer, 2007), a conflict that may even be exacerbated when one motive is externally enhanced. The findings in study 4 of this dissertation hence suggest that participants have a stronger conflict between self-interest and other-regarding preferences after activation of the empathy motive only as compared to the situation in which the empathy motive is activated and a financial incentive is offered. In this latter scenario, self- interest and other-regarding preferences should be aligned as they both favor the prosocial decision option, hence leading to a decrease in conflict. This in turn entails increased efficiency of the prosocial decision process as indicated by increases in the speed of evidence accumulation. In sum, combining the motive of outcome maximization with empathy appeared to boost prosocial decision-making by decreasing the conflict between self-interest and other-regarding preferences, whereas combining the motive of reciprocity with empathy appeared to boost prosocial decision-making by increasing the general motivation to act prosocially.

In this section, different mechanisms underlying the formation and sustainability of empathy-based social closeness and prosocial behavior in contrast to and in combination with other motives were discussed. The following section will focus on the neural activation related to these different mechanisms as assessed using functional magnetic resonance imaging (fMRI).

Neural underpinnings of empathy sustainability

In this section, the average effects of experimental manipulations on different aspects of empathy sustainability will be discussed. The subsequent section "Inter-individual differences" will focus on the central findings with regard to inter-individual differences in neuro-computational effects in the studies of this dissertation.

Results of study 1 showed that during the emotional appraisal of the other person's painful vs. non-painful stimulation, neural activation in the AI and the TPJ was generally increased for trials in which the other person received painful stimulation compared to trials in which the other person received non-painful stimulation. This finding is in line with previous works that have reliably established a link between empathic responses to another’s pain and neural activation in the AI (Hein et al., 2010; Lamm et al., 2011; Y. Li et al., 2020; Marsh, 2018; Patricia L. Lockwood, 2016; Saarela et al., 2007; Singer & Lamm, 2009). That is, neural activation is increased when participants observe another person in a painful situation in contrast to a non-painful situation (e.g., Lamm et al., 2011). Additionally, the larger an individual’s empathic reaction to another’s pain, the larger neural activation in the AI (e.g., Li et al., 2020). Neural activation in the TPJ has been similarly often linked to social cognition, particularly to processes of perspective taking and theory of mind (W. Li, Mai, & Liu, 2014; Moriguchi et al., 2006; Saxe & Kanwisher, 2003; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Steinbeis, 2016; A. Tusche et al., 2016). Specifically, tasks that afforded putting yourself into someone else’s shoes were associated with higher activation of the TPJ as compared to corresponding control conditions (Schurz et al., 2014). Previous works have emphasized the specificity of insular activation linked to affective empathy, i.e., sharing another’s affective state, and neural activation in the TPJ linked to cognitive empathy, i.e., mentalizing/theory of mind/perspective-taking, (Böckler et al., 2014). More recently however, it has been shown that both regions are often co-activated and that the respective activation strength corresponds to relative task affordances of emotional and cognitive empathy (Schurz et al., 2021). Within the paradigm of study 1 in this dissertation, it is plausible to assume that participants increasingly employed processes of affective empathy as well perspective-taking while appraising their emotional reaction to the other person’s painful in contrast to non-painful stimulation. These findings are therefore an indicator for successful activation of the empathy motive which was associated with well-established neural markers of empathic reactions.

During the subsequent updating of social closeness, previously observed painful stimulation was associated with increased neural activation in the dorso-medial prefrontal cortex (dmPFC) and the dorsal striatum as compared to previously observed non-painful stimulation. Interestingly, activation in the dmPFC was not only associated with sensitivity to another’s pain vs. non- pain in study 1 but was also a marker of stronger sustained as opposed to initial empathic reactions in study 2. In more detail, neural activation in the dmPFC during the empathy-based social decision process increased more strongly after weak as compared to after previous strong activation of the empathy motive. In previous studies, the dmPFC has frequently been implied in processes of mentalizing in concert with the TPJ (W. Li et al., 2014; Schurz et al., 2014; Van Overwalle, 2009). More generally, the dmPFC has been linked to social cognition (Lieberman, Straccia, Meyer, Du, & Tan, 2019) as well as social learning (Olsson et al., 2020). The results obtained in this dissertation thus extend these previous findings by demonstrating the dmPFC's implication in empathy-related updating of social closeness as well as sustained neural response in the context of empathy-based social decision-making.

The second region modulated by trial type during social closeness updating, i.e., the partner's painful vs. non-painful stimulation, was the dorsal striatum. This region, too, was implied in the empathy-based and reciprocity-based social decision process as investigated in study 3 of this dissertation (see section "Inter-individual differences" for discussion). The dorsal striatum has previously been associated with the encoding of decision preferences and the implementation of goal-directed behavior (Balleine et al., 2007; Liljeholm & O 'Doherty, 2012; O'Doherty et al., 2004; Palmiter, 2008; Robinson et al., 2006). Its implication in the updating of social closeness in study 1 as well as changes in prosocial decision bias in study 3 suggests that changes in empathy-based social closeness may indeed reflect dynamic changes in the underlying empathy motive strength.

Results of study 2 showed that neural activation during the social decision process in the ventral striatum as well as the inferior frontal gyrus (IFG) and middle cingulate cortex (MCC) was sensitive to empathy activation. In these regions, neural activation generally increased after each motive activation phase, independent of reinforcer frequency in the respective phase, i.e., independent of whether empathy was frequently or rarely reinforced in the previous activation phase. The ventral as opposed to the dorsal striatum is predominantly linked to the encoding and processing of choice values in general (Kable & Glimcher, 2007; Liljeholm & O 'Doherty, 2012; O'Doherty et al., 2004; Strait et al., 2015) as well as social value computation after motive activation in particular (Hein, Morishima, et al., 2016). Specifically, Hein and colleagues (2016) observed that neural activation in the ventral striatum during the social decision process was increased when interacting with a partner towards whom the empathy motive has been explicitly activated as compared to when interacting with a person towards whom no motive has been explicitly activated. The findings in study 2 replicated these previous observations while additionally showing that the ventral striatum was sensitive not only to one initial, but to repeated but weaker instances of empathy motive activation. Neural activation in the inferior frontal gyrus and middle cingulate cortex has also been linked to processes of social cognition in general and emotional empathic responses in particular (Gu & Han, 2007; Harari et al., 2010; Saarela et al., 2007; Walter, 2012). It is thus plausible to conclude that neural activation in those regions during social decision-making may also be sensitive to the degree of social motive activation experiences prior to these decisions.

Taken together, neural activation during trial-by-trial empathy for pain as well as during empathy-based social decision-making was modulated in regions that have frequently been implied in processes of affective (i.e., AI, MCC, IFG) as well as cognitive facets of empathy (i.e., TPJ, dmPFC, temporal poles). During the social decision process as well as during the evaluation of social closeness based on the empathy motive, value encoding regions such as the dorsal and ventral striatum were modulated by motive activation frequencies.

Beyond these average effects of the experimental manipulations the studies in this dissertation revealed results pointing towards meaningful inter-individual differences regarding computational mechanisms of empathy sustainability as well as the related neural activation. In the next section, these effects of inter-individual differences will be discussed.

3.3 Inter-individual differences

Previous studies in the field of social neuroscience have demonstrated that the combination of computational modelling and fMRI offers valuable insights into the mechanisms underlying social learning and social decision behavior on the level of average manipulation effects (Chang & Sanfey, 2013; Hutcherson et al., 2015; R. M. Jones et al., 2011; Lockwood et al., 2016; Lockwood & Klein-Flügge, 2021; Anita Tusche & Bas, 2021), but also on the level of inter-individual differences such as individual learning rates (e.g., Hein, Engelmann, et al., 2016). In this dissertation, we adopted this approach and tested whether determinants of the computational models capturing participants' individual behavior modulated concurrent neural activation differentially in the experimental manipulations.

In study 1, model comparison had revealed that a reinforcement learning model accounting for individual recalibration of the feedback signal value best described participants' behavior. While appraising the emotional reaction, the larger an individual's recalibration, the more sensitive the neural activation to the other's pain vs. non-pain in the treatment condition in regions comprising the STS, the IFG, and the AI. While rating social closeness, large individual recalibration correlated with stronger neural activation in the precuneus, posterior cingulate gyrus, and TPJ after observed pain vs. non-pain, in the treatment condition as compared to the control condition.

Whereas the STS, IFG, AI, and TPJ have frequently been linked to empathy or mentalizing as discussed in the previous section (e.g., Böckler et al., 2014; Hein & Singer, 2008; Lamm et al., 2007; A. Tusche et al., 2016), the precuneus and posterior cingulate gyrus are more specifically related to social learning (Lambert, Declerck, Emonds, & Boone, 2017; Petrini, Piwek, Crabbe, Pollick, & Garrod, 2014; Stanley, 2016). Stanley (2016) for example observed that neural activation in the precuneus and posterior cingulate cortex was more closely related to prediction errors when learning about another person's generosity, i.e., social learning, as compared to learning about the likelihood to win in an analogous lottery task, i.e., non-social learning. In a different study, neural activation in the precuneus was sensitive to whether a biological motion display of two people exhibited a typical, i.e., expected, or atypical, i.e., unexpected, motion pattern (Petrini et al., 2014). Precuneus activation was larger for the unexpected social stimuli in contrast to the socially expected stimuli, again signalling a social prediction error. In study 1 of this dissertation, the extent of recalibration directly influenced participants' prediction error. Hence, based on these previous studies, observing that the extent of recalibration modulated the neural sensitivity to the empathy-reinforcing event (i.e., observed pain) in precuneus suggests that participants increased their empathy-based social closeness towards the other person in a reinforcement learning like process.

Exploratory follow-up analyses showed that individual trait empathic concern and trait perspective-taking modulated how neural activation was associated with reported social closeness in those regions linked to recalibration during the emotional reaction to the other's pain vs. non-pain. The higher an individual's empathic concern score (empathic concern subscale of the IRI, Davis (1980)), the more strongly increased neural responses to another's pain in the IFG/STS were linked to decreased social closeness during acquisition, and were linked to decreased social closeness during extinction. In the STS/TPJ, however, the lower an individual's perspective-taking score (perspective-taking subscale of the IRI, Davis, (1980)), the more decreased neural responses to another's pain were linked to increased social closeness during acquisition as well as during extinction. Hence, the link between neural sensitivity to the other's pain vs. non-pain in IFG/AI and reported social closeness is more strongly reversed from acquisition to extinction, the more empathic an individual. This finding in IFG is particularly interesting as a previous study has found that neural activation in the IFG was connected to empathic reappraisal (Naor et al., 2020). In the study by Naor and colleagues (2020), participants observed pictures of painful experiences (e.g., accidentally cutting yourself with scissors) and were instructed to empathize. Subsequently, they were either asked to simply watch the picture or they were instructed to reappraise their initial empathic feeling. When participants were instructed to reappraise, neural activation in the IFG was higher than when they were instructed to only watch the scene. Additionally, IFG was more strongly connected to regions linked to empathy for pain (ACC and AI) during reappraisal of painful in contrast to neutral scenes. These findings indicate that the IFG may play an important role in adapting one's empathic emotional reaction in response to observed pain. The results of study 1 in this dissertation further strengthen this hypothesis by showing that neural activation in the IFG is (i) modulated by individual recalibration of the empathy-related feedback value and (ii) differentially linked to social closeness depending on the setting of frequent (acquisition) or rare (extinction) empathy reinforcement.

During the prosocial decision process after phases of strong empathy activation vs. weak empathy activation neural activation was predominantly shaped by individual differences in trait empathic concern and individual general prosocial decision bias. Specifically, activation changes in TPJ and dmPFC from baseline to after strong empathy activation were the larger, the higher participants self-reported trait empathy (empathic concern and to a lesser extent perspective-taking), corresponding to an initial empathic response. In contrast, activation changes in the TPJ and the dmPFC from after strong to after weak empathy activation were associated with a stronger prosocial decision bias. This effect represents a sustained empathic response. This differential modulation is interesting as it suggests that different processes may be at play after the initial as opposed to during the sustained influence of empathy on the social decision process. At the initial stage, the participants' affective reaction to the other's pain and the derived preferences of that other person potentially linked to activation in TPJ and dmPFC (R. M. Carter, Bowling, Reeck, & Huettel, 2012; Morishima et al., 2012) may predominantly underlie the social decision process. At the sustained stage, however, more general tendencies to act prosocially towards someone who was suffering may predominantly underlie the social decision process.

In line with the block-wise increase of striatal activation in study 2, the individual relative increase in prosocial decision bias after combined activation of empathy and reciprocity as compared to reciprocity only was associated with an increase in neural activation in bilateral dorsal striatum (study 3). This latter effect on the level of individual differences may reflect an increased motivation for prosocial behavior (Palmiter, 2008), but based on the combination of empathy and reciprocity rather than to increasing levels of empathic motivation. Together, these findings indicate that increased prosocial motivation may shape the social decision process by biasing the encoding of decision preferences and the derived goal-directed behavior. In accordance with results in the realm of motivation more generally (Gluth et al., 2012; Palmiter, 2008), this effect of increased social motivation is potentially implemented in the dorsal striatum (Balleine et al., 2007; Liljeholm & O 'Doherty, 2012; O'Doherty et al., 2004; Palmiter, 2008; Robinson et al., 2006).

In contrast to this account of empathy sustainability via increased motivation to act prosocially, results of study 4 suggest that the combination of empathy with the motive of outcome maximization increased participants' efficiency for making prosocial decisions. The efficiency of the social decision process was associated with individual neural activation in the AI. Specifically, the more efficient an individual's decision process as indicated by a larger drift-rate in the DDM, the higher the insular activation. Follow-up mixed-models analyses taking into account participants' state empathy, individual insular activation during the decision process, decision efficiency, and condition (empathy + financial incentive vs. empathy only) revealed that individuals with low state empathy particularly benefited from the additional offer of a financial incentive. Compared to individuals with high state empathy, these individuals showed larger increases in decision efficiency when financial incentives were added and individual decision efficiency during the social decision process based on empathy as well as outcome maximization was more strongly linked to insular activation. Additionally, individuals with high state empathy yielded increased insular activation. These results suggest that the increased neural activation in the AI, associated with a combined activation of the outcome maximization and empathy motive, was related to increased empathic motivation for high empathic individuals, but to increased efficiency for choosing prosocially in low empathic individuals.

Taken together, individual differences in empathy sustainability as investigated in this dissertation were linked to differences in behavioral markers as well as neural activation. In the context of empathy-based social closeness (study 1), the extent of individual recalibration, an indicator of empathy-based social closeness sustainability, modulated the individual neural sensitivity to observed pain vs. non-pain during the emotional reaction in regions linked to emotional empathic responses (IFG, AI), mentalizing (STS, TPJ), and social learning (precuneus/PCC). In the context of different levels (study 2) and kinds (study 3) of empathy-related prosocial motivation, individual differences in empathy sustainability were indicated by variations in the initial prosocial decision bias. In study 2, the individual general prosocial decision bias modulated the sustained responses to empathy activation in regions previously linked to mentalizing (dmPFC, TPJ). Individual increases in initial prosocial bias based on empathy and reciprocity (study 3) were related to individual increases of neural activation in a region previously linked to the encoding of choice preferences and subsequent goal-directed behavior (dorsal striatum). Combining empathy with the motive of outcome maximization showed that people low in state empathy in particular benefit from the additional offer of a financial bonus, which was linked to individual insular activation during the social decision-making process, a region strongly linked to empathic reactions.

These findings shed light on the ways in which empathy drives and maintains social closeness and prosocial behavior and how these mechanisms are shaped by individual differences. In the following section, potential real-world applications are highlighted that may benefit from the results of this dissertation.

3.4 The real world

Since the empathy-altruism hypothesis (Batson et al., 1991) has been coined, there have been debates on whether empathy can actually incite prosocial behavior in a reliable fashion (Bloom, 2017; Graziano, Habashi, Sheese, & Tobin, 2007; Wilhelm & Bekkers, 2010). By demonstrating persistence of empathy-based social closeness and prosocial decision behavior in different experimental settings of social interactions, the studies presented in this dissertation strongly support the idea of empathy as a reliable driver of prosocial behavior (Batson et al., 1991; Decety et al., 2016). The results obtained may hence tentatively inform applications beyond our laboratory by providing mechanisms that could inspire interventions aiming at sustainably promoting social closeness and prosocial behavior.

If for example, the goal of a mayor of a town was to devise an intervention to promote prosocial behavior and social closeness towards refugees, she may rely on emphasizing the painful experiences refugees made on their way and are still making since they have arrived. The results of this dissertation suggest that she does not need to keep stressing these painful experiences for months. Instead, it may be more helpful to strongly emphasize these experiences in the beginning of the planned intervention, but to only rarely emphasize them towards the end. Such an approach could also decrease effects of empathy habituation (Preis, Kröner-Herwig, Schmidt-Samoa, Dechent, & Barke, 2015). Especially the results from study 3 suggest that such an approach could even further boost prosocial behavior of people who already show prosocial behavior based on other social motives such as reciprocity. Inspired by the results of study 4, offering monetary compensation for helping the refugees may additionally promote prosocial behavior. However, in the study conducted, the financial incentive was given in private in order to exclude potential undermining effects due to reputation considerations (Ferguson, Cameron, & Inzlicht, 2020; Hilbe et al., 2018). Thus, a comparable incentive scheme in a public context should be closely monitored.

Stressing the advantage of motive combinations of empathy with other motives, the combinations of empathy and reciprocity or empathy and financial incentives could also inspire single-person based interventions. For example, on the scale of small teams in a company, an intervention in which empathy towards the other members of the team and the general norm to return helping behavior is activated could sustainably promote social closeness and prosocial behavior within the team and hence improve the working environment. Such an approach may additionally benefit from the mutual strengthening of these two motives over time, as already observed by other studies (Cameron, Conway, & Scheffer, 2022; Mestre, Carlo, Samper, Malonda, & Mestre, 2019; Simpson & Willer, 2008; Von Biebersteinid, Esslid, & Friedrichid, 2021)

In yet another context, the results of study 1 could inform developments in the realm of artificial intelligence when robots learn to (re)act empathically towards humans (Bagheri et al., 2021). This could be accomplished by computationally implementing the learning model derived for human participants in study 1 into the software of the artificial intelligence actor. Based on the model, robots may learn to react in an empathic fashion as well as to sustainably act prosocially towards a human who does not momentarily experience pain but has suffered pain in the past.

Taken together, the results of this dissertation may serve as a starting point to develop and test interventions for sustainably promoting social closeness and prosocial behavior in diverse target groups, ranging from a large-scale audience to small team-based and single-person application contexts.

3.5 Limitations

Despite providing evidence for empathy as a sustainable driver of social closeness and prosocial behavior, the studies conducted in this dissertation face limitations that must be considered and may be addressed in future studies.

First, the participant samples were comprised of females. Previous works have demonstrated that gender significantly influences behavioral and neural empathic responses (Christov-Moore et al., 2014) as well as social decision behavior (Böckler, Tusche, & Singer, 2016; Chowdhury et al., 2017; Eckel & Grossman, 1998; Saad & Gill, 2001). Thus, the findings obtained may not directly translate to male-only contexts as well as gender-mixed contexts. That said, future studies should include participants from all genders.

Second, particularly in studies 1 and 2, sustainability of social closeness and prosocial behavior based on reciprocity was only assessed on a behavioral level. As such, these studies do not show whether the neural regions associated with empathy sustainability are specific to this motive or may also subserve comparable mechanisms based on other social motives such as reciprocity. In order to determine how specific the present neural findings are to empathy-based behavior, future studies need to include other motives in an fMRI study.

Third, the present studies focussed on empathy-based behavior towards strangers, i.e., towards people whom participants did not know before. There is ample evidence that empathy (Beeney et al., 2011) and prosocial behavior (Maner & Gailliot, 2007; Morelli, Knutson, & Zaki, 2018; Padilla-Walker & Christensen, 2011) are influenced by the relationship between the participant and the interaction partner, i.e., whether the other person is for example a friend or a stranger. The observed behavior and underlying (neural) mechanisms reported in this dissertation may thus not apply one-to-one to social behavior towards friends, family, or other people whom you already have a relationship with. Future studies should vary the relational status between participant and interaction partner to test how this influences the sustainability of empathy-based social behavior.

Fourth, in the paradigms used, empathy and reciprocity were activated using an analogous procedure which postulates that trials in which an interaction partner received non-painful stimulation in the empathy context corresponds to trials in which an interaction partner decided not to help in the reciprocity context. However rather than not activating positive reciprocity in this latter type of trials, these trials may in fact activate negative reciprocity instead. Such an account may explain the lack of sustainable social closeness and prosocial behavior based on reciprocity in studies 1 and 2. Follow-up analyses showed that negative trait reciprocity was marginally linked to changes in social closeness in study 1 of this dissertation, but did not differentially influence social closeness during extinction. Future studies may more explicitly account for this potential alternative explanation or develop a different solution that does not activate negative reciprocity.

Lastly, for analysis of the presented fMRI data, univariate neuroimaging analyses were conducted. Previous findings have implied that multivariate approaches are more sensitive (e.g., Haxby, 2012; Huang et al., 2021; Norman, Polyn, Detre, & Haxby, 2006). Multivariate approaches such as multivariate pattern analysis which allows for the detection of more general patterns potentially existing in the data obtained could also bring to light additional neural dynamics linked to the sustainability of empathy-based social behavior as observed in the present studies.

3.6 Future directions

The previous section addressed some limitations of the studies conducted as part of this dissertation, and potential next steps to overcome these limitations in future studies were suggested. However, beyond merely overcoming limitations of the studies presented, future work could more clearly elucidate the mechanisms underlying empathy sustainability in terms of empathy by itself as well as in combination with other motives.

Moving in one possible direction, future work may more closely investigate the neural and behavioral mechanism underlying empathy sustainability by expanding the paradigm applied in study 1. One possible approach could be to focus on the extinction block and parametrically vary the frequency of observed pain (which was set to 20% in our study). This approach may yield an indicator for how sustainably empathy or other motives can incite social closeness and prosocial behavior.

Moreover, in the present studies, sustainability of empathy-based prosocial behavior was quantified using a pre - post design. Analogously to the trial-by-trial ratings of social closeness, one may use trial-by-trial ratings of social decision-making during acquisition and extinction. This would additionally enable the application of RL-DDMs, for which DDM is combined with RL modelling by making the RW updating rule ($V_t = V_{t-1} + a \times \delta_t$) the choice rule for the drift-diffusion decision process (Fontanesi, Gluth, Spektor, & Rieskamp, 2019; Peters & D'Esposito, 2020). Thus, in this model, the parameters characterizing the decision process are updated trial by trial based on the RW learning rule. Model outputs could then be added as trial-by-trial variables in the analysis of the fMRI data.

Another possible direction could follow-up on the combination of the different motives and test more specifically how sustainably empathy in combination multiple other motives incites social closeness and/or prosocial behavior. This approach could shed more light on how sustainable the empathy motive is in combination with more than one motive relative to these other motives alone, as well as relative to combinations of these other motives.

Yet another possibility could be to specify the role of individual differences in empathy sustainability. The studies of this dissertation as well as previous works (e.g., Banissy, Kanai, Walsh, & Rees, 2012; Edele et al., 2013; FeldmanHall et al., 2015; Hein & Singer, 2008)) have shown that (neural) empathic reactions and empathy-based social behavior are shaped by individual differences. Additionally, empathy or the lack thereof has been implicated in psychiatric disorders such as autism spectrum disorder (Fletcher-Watson & Bird, 2020; Patricia L. Lockwood, 2016), depression (Guhn et al., 2020; O'Connor, Berry, Lewis, Mulherin, & Crisostomo, 2007), or psychopathy (Rijnders, Terburg, Bos, Kempes, & van Honk, 2021; van Dongen, 2020). Future studies may include more comprehensive personality measures (e.g., based on the big five ; Caprara, Barbaranelli, Borgogni, & Perugini (1993), obtain indicators of affectivity (e.g., using the positive and negative affect scale (PANAS); Crawford & Henry (2004)), and assess individual scores for autism spectrum disorder, depression, and psychopathy (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Goldstein & Naglieri, 2009; Williams, Nathanson, Paulhus, & others, 2003). Including this additional information could shed light on how much the sustainability of empathy-based behavior is actually affected by personality traits as well as how it may change in psychiatric disorders.

3.7 Conclusion

From the works in this dissertation, we have learnt that empathy for pain can lead to sustainable social closeness towards strangers even when the rate of observed pain decreases over time (study 1). On a mechanistic level, this sustainability can be explained by participants individually adjusting how much the increase in social closeness can also be driven by observed non-pain as opposed to observed pain. The more participants adjusted this value of observed pain vs. non-pain, the more sensitive the neural activation in regions of social cognition and social learning (STS, TPJ, IFG, AI) were to the other's pain vs. non-pain. In terms of prosocial decision-making, the sustainability of the empathy-based initial bias towards making a prosocial decision was paralleled by neural activation in the striatum as well as in dmPFC and TPJ, which are both part of the mentalizing network (study 2). The effect of empathy boosting reciprocity-based decision-making (study 3) by increasing the initial prosocial decision bias was associated with increases in neural activation in dorsal striatum, a region closely linked to changes in motivational state. Results of the final study demonstrated that empathy is resilient to potential undermining effects of additional financial incentives as demonstrated by a more efficient prosocial decision process when driven by empathy and the motive of outcome maximization, a process linked to stronger neural activation in the AI.

Together, the results of this dissertation demonstrate that empathy can incite sustainable prosocial behavior and social closeness drawing on mechanisms that imply neural regions of social cognition, reward learning, and motivation.

4 References

Adams, M. M., & Miller, J. G. (2022). The flexible nature of everyday reciprocity: reciprocity, helping, and relationship closeness. *Motivation and Emotion*, 1–15.

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences*, *3*(12), 469–479. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10562726

Allsop, K., Fifield, K., & Seiter, J. S. (2002). Empathy and generalized reciprocity in compliance with requests for help. *Psychological Reports*, *91*(1), 241–242.

Andreychik, M. R. (2019). Feeling your joy helps me to bear feeling your pain: Examining associations between empathy for others' positive versus negative emotions and burnout. *Personality and Individual Differences*, *137*, 147–156. https://doi.org/10.1016/j.paid.2018.08.028

Andreychik, M. R., & Migliaccio, N. (2015). Empathizing With Others' Pain Versus Empathizing With Others' Joy: Examining the Separability of Positive and Negative Empathy and Their Relation to Different Types of Social Behaviors and Social Emotions. *Basic and Applied Social Psychology*, *37*, 274–291. https://doi.org/10.1080/01973533.2015.1071256

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555. https://doi.org/10.1257/aer.99.1.544

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*. https://doi.org/10.1126/science.7466396

Aylward, J., Hales, C., Robinson, E., & Robinson, O. J. (2019). Translating a rodent measure of negative bias into humans: The impact of induced anxiety and unmedicated mood and anxiety disorders. *Psychological Medicine*, *50*(2), 237–246. https://doi.org/10.1017/S0033291718004117

Báez-Mendoza, R., & Schultz, W. (2013). The role of the striatum in social behavior. *Frontiers in Neuroscience*, *7*(233). https://doi.org/10.3389/fnins.2013.00233

Bagheri, E., Roesler, O., Cao, H.-L., & Vanderborght, B. (2021). A reinforcement learning based cognitive empathy framework for social robots. *International Journal of Social Robotics*, *13*(5), 1079–1093.

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The Role of the Dorsal Striatum in Reward and Decision-Making. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1554-07.2007

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Banissy, M. J., Kanai, R., Walsh, V., & Rees, G. (2012). Inter-individual differences in empathy are reflected in human brain structure. *NeuroImage*, *62*(3), 2034–2039. https://doi.org/10.1016/J.NEUROIMAGE.2012.05.081

Bartlett, L., & DeSteno, D. (2006). Gratitude and prosocial behavior. *Psychological Science*, *17*(4), 319–325. https://doi.org/10.1111/j.1467-9280.2006.01705.x

Barton, K. (2019). MuMIn: Multi-modal inference. Model selection and model averaging based on information criteria (AICc and alike) R package version 1.43.15. *Http://Cran.r-*

Project.Org/Web/Packages/MuMIn/Index.Html.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21767–21772. https://doi.org/10.1073/pnas.0908104107

Basten, Ulrike, Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21767–21772. https://doi.org/10.1073/pnas.0908104107

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(48). https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Eigen, C. (2014). Package "lme4." *Comprehensive R Archive Network (CRAN)*.

Batson, D. (1994). Why Act for the Public Good? Four Answers. *Personality and Social Psychology Bulletin*, *20*(5), 603–610. https://doi.org/10.1177/0146167294205016

Batson, D. (2010). Empathy-induced altruistic motivation. In *Prosocial motives, emotions, and behavior: The better angels of our nature.* https://doi.org/http://dx.doi.org/10.1037/12061-001

Batson, D., Ahmad, N., & Stocks, E. L. (2011). Four forms of prosocial motivation egoism, altruism, collectivism, and principlism. In D. Dunning (Ed.), *Social Motivation* (pp. 103–126). New York: Psychology Press. https://doi.org/10.4324/9780203833995

Batson, D., Ahmad, N., & Tsang, J.-A. (2004). Four Motives for Community Involvement. *Journal of Social Issues*, *58*(3), 429–445. https://doi.org/10.1111/1540-4560.00269

Batson, D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., & Todd, R. M. (1991). Empathic Joy and the Empathy-Altruism Hypothesis. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.61.3.413

Batson, D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*(2), 290–302. https://doi.org/10.1037/0022-3514.40.2.290

Batson, D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., & Griffitt, C. (1988). Five Studies Testing Two New Egoistic Alternatives to the Empathy-Altruism Hypothesis. *Journal of Personality and Social Psychology*, *55*(1), 52–77. https://doi.org/10.1037/0022-3514.55.1.52

Batson, D., & Shaw, L. L. (1991). Evidence for Altruism: Toward a Pluralism of Prosocial Motives. *Psychological Inquiry*, *2*(2), 107–122. https://doi.org/10.1207/s15327965pli0202_1

Batson, D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information Function of Empathic Emotion. *Journal of Personality and Social Psychology*, *68*(2), 300–313. Retrieved from 10.1037/0022-3514.68.2.300

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The Neural Circuitry of a Broken Promise. *Neuron*. https://doi.org/10.1016/j.neuron.2009.11.017

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06781-2

Beck, A. T., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). Beck depression inventory (BDI). *Arch Gen Psychiatry*, *4*(6), 561–571.

Beeney, J. E., Franklin, R. G., Levy, K. N., & Adams, R. B. (2011). I feel your pain: emotional closeness modulates neural responses to empathically experienced rejection. *Social Neuroscience*, *6*(4), 369–376. https://doi.org/10.1080/17470919.2011.557245

Bellucci, G., Camilleri, J. A., Eickhoff, S. B., & Krueger, F. (2020). Neural signatures of prosocial behaviors. *Neuroscience and Biobehavioral Reviews*, *118*, 186–195. https://doi.org/10.1016/j.neubiorev.2020.07.006

Ben-Ner, A., & Kramer, A. (2011). Personality and altruism in the dictator game: Relationship to giving to kin, collaborators, competitors, and neutrals. *Personality and Individual Differences*, *51*(3), 216–221. https://doi.org/10.1016/j.paid.2010.04.024

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Besley, T., & Ghatak, M. (2018). Prosocial Motivation and Incentives. *Annual Review of Economics*, *10*(10), 411–438. https://doi.org/10.1146/annurev-economics-063016-103739

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: Social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 61–73. https://doi.org/10.1002/wcs.1266

Blockley, N. P., Griffeth, V. E. M., Simon, A. B., & Buxton, R. B. (2013). A review of calibrated blood oxygenation level-dependent (BOLD) methods for the measurement of task-induced changes in brain oxygen metabolism. *NMR in Biomedicine*. https://doi.org/10.1002/nbm.2847

Bloom, P. (2017). Empathy and Its Discontents. *Trends in Cognitive Sciences*, *21*(1), 24–31. https://doi.org/10.1016/J.TICS.2016.11.004

Böckler, A., Kanske, P., Trautwein, F.-M., & Singer, T. (2014). The EmpaToM: A novel fMRI-task separating affective and cognitive routes to social cognition. In *20th Annual Meeting of the Organization for Human Brain Mapping (OHBM)*.

Böckler, A., Tusche, A., & Singer, T. (2016). The Structure of Human Prosociality: Differentiating Altruistically Motivated, Norm Motivated, Strategically Motivated, and Self-Reported Prosocial Behavior. *Social Psychological and Personality Science*, *7*(6), 530–541. https://doi.org/10.1177/1948550616639650

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bohnet, I., & Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review*, *89*(1), 335–339. https://doi.org/10.1257/aer.89.1.335

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Bottemanne, L., & Dreher, J. C. (2019). Vicarious rewards modulate the drift rate of evidence accumulation from the drift diffusion model. *Frontiers in Behavioral Neuroscience*. https://doi.org/10.3389/fnbeh.2019.00142

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective and Behavioral Neuroscience*. https://doi.org/10.3758/CABN.7.4.356

Bouton, M. E. (2004). Context and Behavioral Processes in Extinction. *Learning & Memory*, *11*(5), 485–494. https://doi.org/10.1101/LM.78804

Bowles, S. (2008). Policies designed for self-interested citizens may undermine "The moral sentiments": Evidence from economic experiments. *Science*, *320*, 1605–1609. https://doi.org/10.1126/science.1152110

Brett, M, Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using an SPM toolbox - abstract presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. *NeuroImage*. https://doi.org/http://dx.doi.org/10.1016/S1053-8119(02)90010-8

Brett, Matthew, Anton, J., Valabregue, R., & Poline, J. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*.

Bruneau, E. G., Pluta, A., & Saxe, R. (2012). Distinct roles of the "Shared Pain" and "Theory of Mind" networks in processing others' emotional suffering. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2011.11.008

Buckner, R. L. (1998). Event-related fMRI and the hemodynamic response. *Human Brain Mapping*. https://doi.org/10.1002/(sici)1097-0193(1998)6:5/6<373::aid-hbm8>3.3.co;2-g

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. *Journal of SocioEconomics* (Vol. 32). https://doi.org/10.1016/j.socec.2003.10.009

Cameron, C. D., Conway, P., & Scheffer, J. A. (2022). Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology*, *44*, 188–195. https://doi.org/10.1016/J.COPSYC.2021.09.011

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000595

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and Individual Differences*, *15*(3), 281–288. https://doi.org/10.1016/0191-8869(93)90218-R

Carter, C. S., Lesh, T. A., & Barch, D. M. (2016). Thresholds, Power, and Sample Sizes in Clinical Neuroimaging. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 99–100. https://doi.org/10.1016/j.bpsc.2016.01.005

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, *337*(6090), 109–111.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*(3), 277–284. https://doi.org/10.1093/scan/nsr094

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.1162/003355302760193904

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: promises and pitfalls. *Social Neuroscience*. https://doi.org/10.1080/17470919.2018.1518834

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*(1), 1–10.

https://doi.org/10.1038/s41467-018-05994-9

Chernyak, N., Leimgruber, K. L., Dunham, Y. C., Hu, J., & Blake, P. R. (2019). Paying Back People Who Harmed Us but Not People Who Helped Us: Direct Negative Reciprocity Precedes Direct Positive Reciprocity in Early Development. *Psychological Science*. https://doi.org/10.1177/0956797619854975

Chowdhury, S. M., Jeon, J. Y., & Saha, B. (2017). Gender Differences in the Giving and Taking Variants of the Dictator Game. *Southern Economic Journal*, *84*(2), 474–483. https://doi.org/10.1002/soej.12223

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience and Biobehavioral Reviews*, *46*(4), 604–627. https://doi.org/10.1016/j.neubiorev.2014.09.001

Churamani, N., Barros, P., Strahl, E., & Wermter, S. (2018). Learning Empathy-Driven Emotion Expressions using Affective Modulations. In *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2018.8489158

Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A. L. (1987). Empathy-Based Helping: Is It Selflessly or Selfishly Motivated? *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.52.4.749

Cory, G. A. (2006). A behavioral model of the dual motive approach to behavioral economics and social exchange. *Journal of Socio-Economics*, *35*(4), 592–612. https://doi.org/10.1016/j.socec.2005.12.017

Cox, C. L., Uddin, L. Q., Di martino, A., Castellanos, F. X., Milham, M. P., & Kelly, C. (2012). The balance between feeling and knowing: Affective and cognitive empathy are reflected in the brain's intrinsic functional dynamics. *Social Cognitive and Affective Neuroscience*, *7*(6), 727–737. https://doi.org/10.1093/scan/nsr051

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*(3), 245–265. https://doi.org/10.1348/0144665031752934

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/10.1016/j.neuroimage.2018.09.009

Dambacher, M., & Hübner, R. (2015). Time pressure affects the efficiency of perceptual processing in decisions under conflict. *Psychological Research*, *79*(1), 83–94.

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003441

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*. https://doi.org/10.1.1.462.7754

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

Davis, M. H. (2006). Empathy. In *Handbook of the sociology of emotions* (pp. 443–466). Springer.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*. https://doi.org/10.1016/j.neuron.2011.02.027

Dawes, C. T., Loewen, P. J., Schreiber, D., Simmons, A. N., Flagan, T., McElreath, R., … Paulus,

M. P. (2012). Neural basis of egalitarian behavior. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1118653109

De Hollander, G., Forstmann, B. U., & Brown, S. D. (2016). Different Ways of Linking Behavioral and Neural Data via Computational Cognitive Models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *33*(4), 1400–1410. https://doi.org/10.1016/j.bpsc.2015.11.004

de Lange, F. P., Rahnev, D. A., Donner, T. H., & Lau, H. (2013). Prestimulus oscillatory activity over motor cortex reflects perceptual expectations. *Journal of Neuroscience*, *33*(4), 1400–1410. https://doi.org/10.1523/JNEUROSCI.1094-12.2013

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., Jackson, P. L., Sommerville, J. A., Caminade, T., & Meltzoff, A. N. (2004). Investigation. *NeuroImage*, *23*(2), 744–751. https://doi.org/10.1016/j.neuroimage.2004.05.025.The

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668. https://doi.org/10.1037//0033-2909.125.6.627

Depow, G. J., Francis, Z., & Inzlicht, M. (2021). The Experience of Empathy in Everyday Life. *Psychological Science*, *32*(8), 1198–1213. https://doi.org/10.1177/0956797621995202

Dittrich, M. (2015). Gender differences in trust and reciprocity: evidence from a large-scale experiment with heterogeneous subjects. *Applied Economics*. https://doi.org/10.1080/00036846.2015.1019036

Domenech, P., Redouté, J., Koechlin, E., & Dreher, J. C. (2018). The Neuro-Computational Architecture of Value-Based Selection in the Human Brain. *Cerebral Cortex*, *28*(2), 585–601. https://doi.org/10.1093/cercor/bhw396

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Dunsmoor, J. E., Kroes, M. C. W., Moscatelli, C. M., Evans, M. D., Davachi, L., & Phelps, E. A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour*, *2*(4), 291–299. https://doi.org/10.1038/s41562-018-0317-4

Dvash, J., & Shamay-Tsoory, S. G. (2014). Theory of mind and empathy as multidimensional constructs: Neurological foundations. *Topics in Language Disorders*, *34*(4), 282–295. https://doi.org/10.1097/TLD.0000000000000040

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., & Leibo, J. Z. (2020). Learning reciprocity in complex sequential social dilemmas. *Trends in Cognitive Sciences*, *24*(10), 802–813.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*, 181–191. https://doi.org/10.1006/game.1996.0081

Eckel, C. C., & Grossman, P. J. (1998). Are women less selfish than men?: Evidence from dictator experiments. *Economic Journal*, *108*(448), 726–735. https://doi.org/10.1111/1468-0297.00311

Edele, A., Dziobek, I., & Keller, M. (2013). Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. *Learning and Individual Differences*, *24*, 96–102. https://doi.org/10.1016/j.lindif.2012.12.020

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Ellis, P. D. (2012). Power analysis and the detection of effects. In *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results* (pp. 47–72). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511761676.004

Engel, C., & Zhurakhovska, L. (2016). When is the risk of cooperation worth taking? The prisoner's dilemma as a game of multiple motives. *Applied Economics Letters*, *23*(16), 1157–1161. https://doi.org/10.1080/13504851.2016.1139672

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*. https://doi.org/10.1016/j.geb.2008.12.006

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

Eres, R., Decety, J., Louis, W. R., & Molenberghs, P. (2015). Individual differences in local gray matter density are associated with differences in affective and cognitive empathy. *NeuroImage*, *117*, 305–310. https://doi.org/10.1016/j.neuroimage.2015.05.038

Exley, C. (2018). Incentives for prosocial behavior: The role of reputations. *Management Science*, *64*(5), 2460–2471.

Falbén, J. K., Golubickis, M., Tamulaitis, S., Caughey, S., Tsamadi, D., Persson, L. M., … Macrae, C. N. (2020). Self-relevance enhances evidence gathering during decision-making. *Acta Psychologica*. https://doi.org/10.1016/j.actpsy.2020.103122

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*. https://doi.org/10.1016/j.geb.2005.03.001

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

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2007.09.002

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*. https://doi.org/10.1007/s12110-002-1012-7

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

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

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., Montez, D. F., Phelps, E. A., Davachi, L., & Murty, V. P. (2021). Hippocampus guides adaptive learning during dynamic social interactions. *Journal of Neuroscience*, *41*(6), 1340–1348.

Ferguson, A. M., Cameron, C. D., & Inzlicht, M. (2020). Motivational effects on empathic choices. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2020.104010

Flagan, T., Mumford, J. A., & Beer, J. S. (2017). How do you see me? The neural basis of motivated meta-perception. *Journal of Cognitive Neuroscience*, *29*(11), 1908–1917. https://doi.org/10.1162/jocn_a_01169

Fletcher-Watson, S., & Bird, G. (2020). Autism and empathy: What are the real links? *Autism*, *24*(1), 3–6. https://doi.org/10.1177/1362361319883506

Foerde, K., Steinglass, J. E., Shohamy, D., & Walsh, B. T. (2015). Neural mechanisms supporting maladaptive food choices in anorexia nervosa. *Nature Neuroscience*. https://doi.org/10.1038/nn.4136

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin and Review*, *26*, 1099–1126. https://doi.org/10.3758/s13423-018-1554-2

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cognitive, Affective, \& Behavioral Neuroscience*, *19*(3), 490–502.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E. J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0805903105

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Forstmann, B. U., & Wagenmakers, E. J. (2015). *An introduction to model-based cognitive neuroscience*. *An Introduction to Model-Based Cognitive Neuroscience*. New York: Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-2236-9

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347–369.

Fox, J., & Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition. *Thousand Oaks CA: Sage.*

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., & Baud-Bovy, G. (2018). Package “car.” *R Documentation*.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *NeuroImage*. https://doi.org/10.1006/nimg.1997.0306

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

Frolichs, K., Rosenblau, G., & Korn, C. (2021). How do humans learn about other people?

Incorporating social knowledge structures into reinforcement learning.

Functional Imaging Laboratory, W. T. C. f. N. (2019). *SPM12 manual*. Institute of Neurology, UCL.

Gächter, S., & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics*. https://doi.org/10.1111/1467-9442.00269

Gallotti, R., & Grujić, J. (2019). A quantitative description of the transition between intuitive altruism and rational deliberation in iterated Prisoner's Dilemma experiments. *Scientific Reports*, *9*, 1–11. https://doi.org/10.1038/s41598-019-52359-3

Garbers, Y., & Konradt, U. (2014). The effect of financial incentives on performance: A quantitative review of individual and team-based financial incentives. *Journal of Occupational and Organizational Psychology*, *87*(1), 102–137. https://doi.org/10.1111/joop.12039

Garrett, N., & Daw, N. D. (2020). Biased belief updating and suboptimal choice in foraging decisions. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16964-5

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. Retrieved from doi.org/10.1214/ss/1177011136

Germar, M., Albrecht, T., Voss, A., & Mojzisch, A. (2016). Social conformity is due to biased stimulus processing: Electrophysiological and diffusion analyses. *Social Cognitive and Affective Neuroscience*, *11*(9), 1449–1459. https://doi.org/10.1093/scan/nsw050

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*(31), 10686–10698. https://doi.org/10.1523/JNEUROSCI.0727-12.2012

Goldstein, S., & Naglieri, J. A. (2009). *Autism spectrum rating scales (ASRS)*. Multi-Health System North Tonawanda, NY.

Gouldner, A. W. (1960). The Norm of Reciprocity : A Preliminary Statement. *American Sociological Review*, *25*(2), 161–178.

Graziano, W. G., Habashi, M. M., Sheese, B. E., & Tobin, R. M. (2007). Agreeableness, empathy, and helping: A person × situation perspective. *Journal of Personality and Social Psychology*, *93*(4), 583–599. https://doi.org/10.1037/0022-3514.93.4.583

Grynberg, D., & Konrath, S. (2020). The closer you feel, the more you care: Positive associations between closeness, pain intensity rating, empathic concern and personal distress to someone in pain. *Acta Psychologica*, *210*, 1031–1075. https://doi.org/10.1016/J.ACTPSY.2020.103175

Gu, X., & Han, S. (2007). Attention and reality constraints on the neural processes of empathy for pain. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2007.02.025

Guhn, A., Merkel, L., Hübner, L., Dziobek, I., Sterzer, P., & Köhler, S. (2020). Understanding versus feeling the emotions of others: How persistent and recurrent depression affect empathy. *Journal of Psychiatric Research*, *130*, 120–127. https://doi.org/10.1016/j.jpsychires.2020.06.023

Haacke, E. M., Lai, S., Reichenbach, J. R., Kuppusamy, K., Hoogenraad, F. G. C., Takeichi, H., & Lin, W. (1997). In vivo measurement of blood oxygen saturation using magnetic resonance imaging: A direct validation of the blood oxygen level-dependent concept in functional brain imaging. *Human Brain Mapping*. https://doi.org/10.1002/(SICI)1097-0193(1997)5:5<341::AID-HBM2>3.0.CO;2-3

Han, H., & Glenn, A. L. (2018). Evaluating methods of correcting for multiple comparisons implemented in SPM12 in social neuroscience fMRI studies: an example from moral psychology. *Social Neuroscience*, *13*(3), 257–267. https://doi.org/10.1080/17470919.2017.1324521

Harari, H., Shamay-Tsoory, S. G., Ravid, M., & Levkovitz, Y. (2010). Double dissociation between cognitive and affective empathy in borderline personality disorder. *Psychiatry Research*. https://doi.org/10.1016/j.psychres.2009.03.002

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

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, *324*(5927), 646–648. https://doi.org/10.1126/science.1168450

Hare, T. A., Hakimi, S., & Rangel, A. (2014). Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. *Frontiers in Neuroscience*, *8*(8 MAR), 1–15. https://doi.org/10.3389/fnins.2014.00050

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1109322108

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Hein, G., Engelmann, J. B., Vollberg, M. C., & Tobler, P. N. (2016). How learning shapes the empathic brain. *Proceedings of the National Academy of Sciences*, *113*(1), 80–85. https://doi.org/10.1073/pnas.1514539112

Hein, G., Morishima, Y., Leiberg, S., Sul, S., & Fehr, E. (2016). The brain's functional network architecture reveals human motives. *Science*, *351*(6277), 1074–1078. https://doi.org/DOI: 10.1126/science.aac7992

Hein, G., Silani, G., Preuschoff, K., Batson, 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

Hein, G., & Singer, T. (2008). I feel how you feel but not always: the empathic brain and its modulation. *Current Opinion in Neurobiology*, *18*(2), 153–158. https://doi.org/10.1016/J.CONB.2008.07.012

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., & Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, *115*(48), 12241–12246. https://doi.org/10.1073/pnas.1810565115

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *American Economic Review*, *86*(9), 653–660. https://doi.org/10.1017/cbo9780511528347.009

Holmås, T. H., Kjerstad, E., Lurås, H., & Straume, O. R. (2010). Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay.

Journal of Economic Behavior and Organization, *75*(2), 261–267. https://doi.org/10.1016/j.jebo.2010.03.024

Hu, J., Li, Y., Yin, Y., Blue, P. R., Yu, H., & Zhou, X. (2017). How do self-interest and other-need interact in the brain to determine altruistic behavior? *NeuroImage*, *157*(June), 598–611. https://doi.org/10.1016/j.neuroimage.2017.06.040

Huang, Q., Li, D., Zhou, C., Xu, Q., Li, P., & Warren, C. M. (2021). Multivariate pattern analysis of electroencephalography data reveals information predictive of charitable giving. *NeuroImage*, *242*, 111475. https://doi.org/10.1016/j.neuroimage.2021.118475

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64. https://doi.org/10.1016/j.tics.2014.12.006

Hunter, L. E., & Daw, N. D. (2021). Context-sensitive valuation and learning. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/j.cobeha.2021.05.001

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

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of Social and Monetary Rewards in the Human Striatum. *Neuron*. https://doi.org/10.1161/STROKEAHA.107.501643

Jagers, S. C., Linde, S., Martinsson, J., & Matti, S. (2017). Testing the Importance of Individuals' Motives for Explaining Environmentally Significant Behavior. *Social Science Quarterly*, *98*(2), 644–658. https://doi.org/10.1111/ssqu.12321

Janczyk, M., & Lerche, V. (2019). A diffusion model analysis of the response-effect compatibility effect. *Journal of Experimental Psychology: General*, *148*(2), 237–251.

Jauniaux, J., Khatibi, A., Rainville, P., & Jackson, P. L. (2019). A meta-analysis of neuroimaging studies on pain empathy: Investigating the role of visual information and observers' perspective. *Social Cognitive and Affective Neuroscience*, *14*(8), 789–813. https://doi.org/10.1093/scan/nsz055

Jennings, J. H., Sparta, D. R., Stamatakis, A. M., Ung, R. L., Pleil, K. E., Kash, T. L., & Stuber, G. D. (2013). Distinct extended amygdala circuits for divergent motivational states. *Nature*, *496*(7444), 224. https://doi.org/10.1038/nature12041

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*(4), 283–286. https://doi.org/10.1111/j.1467-9280.2006.01699.x

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, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct? *Emotion*, *16*(8), 1107–1116. https://doi.org/10.1037/emo0000228

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*. https://doi.org/10.1038/nn2007

Kaltwasser, L., Hildebrandt, A., Wilhelm, O., & Sommer, W. (2016). Behavioral and neuronal determinants of negative reciprocity in the ultimatum game. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsw069

Kanske, P., Böckler, A., Trautwein, F. M., Lesemann, F. H. P., & Singer, T. (2016). Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. *Social Cognitive and Affective Neuroscience*, *11*(9), 1383–1392. https://doi.org/10.1093/SCAN/NSW052

Kanske, P., Böckler, A., Trautwein, F. M., & Singer, T. (2015). Dissecting the social brain:

Introducing the EmpaToM to reveal distinct neural networks and brain-behavior relations for empathy and Theory of Mind. *NeuroImage*, *122*, 6–19. https://doi.org/10.1016/j.neuroimage.2015.07.082

Katsimpokis, D., Hawkins, G. E., & van Maanen, L. (2020). Not all Speed-Accuracy Trade-Off Manipulations Have the Same Psychological Effect. *Computational Brain and Behavior*. https://doi.org/10.1007/s42113-020-00074-y

Kennedy, P. J., & Shapiro, M. L. (2009). Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0903259106

Kim, B., & Im, H.-I. (2018). The role of the dorsal striatum in choice impulsivity. *Annals of the New York Academy of Sciences*.

Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms16033

Klimecki, O. M., Mayer, S. V., Jusyte, A., Scheeff, J., & Schönenberg, M. (2016). Empathy promotes altruistic behavior in economic interactions. *Scientific Reports*, *6*(1), 1–5. https://doi.org/10.1038/srep31961

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*(5800), 829–832. https://doi.org/10.1126/science.1129156

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience*, *25*(19), 4806–4812. https://doi.org/10.1523/JNEUROSCI.0642-05.2005

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A Common Mechanism Underlying Food Choice and Social Decisions. *PLoS Computational Biology*, *11*(10). https://doi.org/10.1371/journal.pcbi.1004371

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2018). A theory of goal systems. In *The motivated mind* (pp. 215–258). Routledge. https://doi.org/10.4324/9781315175867

Lambert, B., Declerck, C. H., Emonds, G., & Boone, C. (2017). Trust as commodity: social value orientation affects the neural substrates of learning to cooperate. *Social Cognitive and Affective Neuroscience*, *12*(4), 609–617. https://doi.org/10.1093/SCAN/NSW170

Lamm, C., Batson, D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn.2007.19.1.42

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

Lay, S., Zagefka, H., González, R., Álvarez, B., & Valdenegro, D. (2020). Don’t forget the group! The importance of social norms and empathy for shaping donation behaviour. *International Journal of Psychology*, *55*(4), 518–531. https://doi.org/10.1002/ijop.12626

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-017-0067

Lenth, R., Singman, H., Love, J., Buerkner, P., & Herve, M. (2019). Emmeans package: Estimated Marginal means, aka Least-Squares Means. *R Package Version 1.15-15*.

Leong, Y. C., Hughes, B. L., Wang, Y., & Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. *Nature Human Behaviour*, *3*(9), 962–973. https://doi.org/10.1038/s41562-019-0637-z

Lerche, V., Neubauer, A. B., & Voss, A. (2018). Effects of implicit fear of failure on cognitive processing: A diffusion model analysis. *Motivation and Emotion*, *42*(3), 386–402. https://doi.org/10.1007/s11031-018-9691-5

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01324

Lewin, K. (1951). *Field theory in social sciences*. New York: Harper & Row.

Lewin, K., Cartwright, D., & Price, D. (1951). *Field theory in social science: Selected theoretical papers.* (D. Cartwright, Ed.), *American Sociological Review*. https://doi.org/10.1037/0021-9010.69.1.85

Li, W., Mai, X., & Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us. *Frontiers in Human Neuroscience*, *8*, 74. https://doi.org/10.3389/fnhum.2014.00074

Li, Y., Li, W., Zhang, T., Zhang, J., Jin, Z., & Li, L. (2021). Probing the role of the right inferior frontal gyrus during Pain-Related empathy processing: Evidence from fMRI and TMS. *Human Brain Mapping*, *42*(5), 1518–1531.

Li, Y., Zhang, T., Li, W., Zhang, J., Jin, Z., & Li, L. (2020). Linking brain structure and activation in anterior insula cortex to explain the trait empathy for pain. *Human Brain Mapping*, *41*(4), 1030–1042.

Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, *58*, 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

Lieberman, M. D., Straccia, M. A., Meyer, M. L., Du, M., & Tan, K. M. (2019). Social, self,(situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence. *Neuroscience \& Biobehavioral Reviews*, *99*, 311–328.

Liljeholm, M., & O ’Doherty, J. P. (2012). Contributions of the striatum to learning, motivation, and performance: an associative account Anatomical and functional delineations of the striatum. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2012.07.007

Linares, D., & López-Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *R Journal*. https://doi.org/10.32614/rj-2016-008

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., & Klein-Flügge, M. C. (2021). Computational modelling of social cognition and behaviour - A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsaa040

Logothetis, N. K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. In *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2002.1114

Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models.

Journal of Open Source Software, *3*(26), 772. https://doi.org/10.21105/joss.00772

Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (2017). A Flexible and Efficient Hierarchical Bayesian Approach to the Exploration of Individual Differences in Cognitive-model-based Neuroscience. In *Computational Models of Brain and Behavior*. https://doi.org/10.1002/9781119159193.ch34

Majdandžić, J., Amashaufer, S., Hummer, A., Windischberger, C., & Lamm, C. (2016). The selfless mind: How prefrontal involvement in mentalizing with similar and dissimilar others shapes empathy and prosocial behavior. *Cognition*, *157*, 24–38. https://doi.org/10.1016/j.cognition.2016.08.003

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*. https://doi.org/10.1016/S1053-8119(03)00169-1

Maner, J. K., & Gailliot, M. T. (2007). Altruism and egoism: Prosocial motivations for helping depend on relationship context. *European Journal of Social Psychology*, *37*(2), 347–358. https://doi.org/10.1002/ejsp.364

Marsh, A. A. (2018). The neuroscience of empathy. *Current Opinion in Behavioral Sciences*, *19*, 110–115. https://doi.org/10.1016/j.cobeha.2017.12.016

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika*, *81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x

Masten, C. L., Eisenberger, N. I., Pfeifer, J. H., & Dapretto, M. (2011). Witnessing peer rejection during early adolescence: Neural correlates of empathy for experiences of social exclusion. *Social Neuroscience*, *5*, 496–507. https://doi.org/10.1080/17470919.2010.490673.Witnessing

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

Matlab Inc. (2015). Matlab. Natick, Massachusetts: The MathWorks Inc.

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/PBR.16.5.798

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, *52*(2), 267–275. https://doi.org/10.1016/S0167-2681(03)00003-9

McClelland, D. C. (1985). How motives, skills, and values determine what people do. *American Psychologist*. https://doi.org/10.1037//0003-066x.40.7.812

McClelland, D. C. (2014). How Motives Interact with Values and Skills to Determine What People Do. In *Human Motivation*. https://doi.org/10.1017/cbo9781139878289.015

Mestre, M. V., Carlo, G., Samper, P., Malonda, E., & Mestre, A. L. (2019). Bidirectional relations among empathy-related traits, prosocial moral reasoning, and prosocial behaviors. *Social Development*, *28*(3), 514–528.

Mill, J. S. (1836). On the Definition of Political Economy, and on the Method of Investigation Proper to it. *Don and Westminster Review*, *4*(October), 120–164.

Morelli, S. A., Knutson, B., & Zaki, J. (2018). Neural sensitivity to personal and vicarious reward differentially relate to prosociality and well-being. *Social Cognitive and Affective Neuroscience*, *13*(8), 831–839. https://doi.org/10.1093/scan/nsy056

Morelli, S. A., Lieberman, M. D., & Zaki, J. (2015). The emerging study of positive empathy.

Social and Personality Psychology Compass, *9*(2), 57–68. https://doi.org/10.1111/spc3.12157

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., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., … Komaki, G. (2006). Impaired self-awareness and theory of mind: An fMRI study of mentalizing in alexithymia. *NeuroImage*, *32*(3), 1472–1482. https://doi.org/10.1016/j.neuroimage.2006.04.186

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

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. https://doi.org/10.1523/jneurosci.4156-11.2012

Murayama, K., Matsumoto, M., Izuma, K., & Matsumoto, K. (2010). Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(49), 20911–20916. https://doi.org/10.1073/pnas.1013305107

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS Biology*, *18*(12). https://doi.org/10.1371/journal.pbio.3001028

Naor, N., Rohr, C., Schaare, L. H., Limbachia, C., Shamay-Tsoory, S., & Okon-Singer, H. (2020). The neural networks underlying reappraisal of empathy for pain. *Social Cognitive and Affective Neuroscience*, *15*(7), 733–744. https://doi.org/10.1093/SCAN/NSAA094

Neyer, F. J., Wrzus, C., Wagner, J., & Lang, F. R. (2011). Principles of relationship differentiation. *European Psychologist*. https://doi.org/10.1027/1016-9040/a000055

Niza, C., Tung, B., & Marteau, T. M. (2013). Incentivizing blood donation: Systematic review and meta-analysis to test titmuss' hypotheses. *Health Psychology*, *32*(9), 941–949. https://doi.org/10.1037/a0032740

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology*, *76*(B), 117–130. https://doi.org/10.1016/j.jmp.2016.03.003

O'Connor, L. E., Berry, J. W., Lewis, T., Mulherin, K., & Crisostomo, P. S. (2007). Empathy and depression: The moral system on overdrive. In *Empathy in Mental Illness* (p. 75). https://doi.org/10.1017/CBO9780511543753.005

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004).

Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*. https://doi.org/10.1126/science.1094285

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.87.24.9868

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

Olsson, A., & Spring, V. (2018). The Vicarious Brain: Integrating Empathy and Emotional Learning. *Neuronal Correlates of Empathy: From Rodent to Human*, 7–23. https://doi.org/10.1016/B978-0-12-805397-3.00002-4

Orhun, A. Y. (2018). Perceived motives and reciprocity. *Games and Economic Behavior*. https://doi.org/10.1016/j.geb.2018.01.002

Padilla-Walker, L. M., & Christensen, K. J. (2011). Empathy and self-regulation as mediators between parenting and adolescents' prosocial behavior toward strangers, friends, and family. *Journal of Research on Adolescence*, *21*(3), 545–551. https://doi.org/10.1111/j.1532-7795.2010.00695.x

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9096

Palmiter, R. D. (2008). Dopamine signaling in the dorsal striatum is essential for motivated behaviors: Lessons from dopamine-deficient mice. *Annals of the New York Academy of Sciences*, (1129), 35. https://doi.org/10.1196/annals.1417.003

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*, 15964. https://doi.org/10.1038/ncomms15964

Patricia L. Lockwood. (2016). The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behavioural Brain Research*, *311*, 255–266.

Pedersen, M. L., Endestad, T., & Biele, G. (2015). Evidence accumulation and choice maintenance are dissociated in human perceptual decision making. *PLoS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0140361

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behaviour: Multilevel perspectives. *Annual Review of Psychology*, *56*(1), 365–392. https://doi.org/10.1146/annurev.psych.56.091103.070141

Perry, A., & Shamay-Tsoory, S. (2013). Understanding emotional and cognitive empathy: A neuropsychological perspective. *Understanding Other Minds: Perspectives from Developmental Social Neurosciene, Oxford University Press*.

Perugini, M., Gallucci, M., Presaghi, F., & Ercolani, A. P. (2003). The Personal Norm of Reciprocity. *European Journal of Personality*. https://doi.org/10.1002/per.474

Peters, J., & D'Esposito, M. (2020). The drift diffusion model as the choice rule in inter-temporal and risky choice: A case study in medial orbitofrontal cortex lesion patients and controls. *PLoS Computational Biology*, *16*(4). https://doi.org/10.1371/journal.pcbi.1007615

Petrini, K., Piwek, L., Crabbe, F., Pollick, F. E., & Garrod, S. (2014). Look at those two!: The precuneus role in unattended third-person perspective of social interactions. *Human Brain Mapping*, *35*(10), 5190. https://doi.org/10.1002/HBM.22543

Pfaffenberger, R. C., & Patterson, J. H. (1977). *Statistical Methods for Business and Economics.* Georgetown, Ontario: Irwin-Dorsey Ltd.

Pischedda, D., Palminteri, S., & Coricelli, G. (2020). The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *The Journal of Neuroscience*. https://doi.org/10.1101/2020.01.08.898841

Powers, K. E., Chavez, R. S., & Heatherton, T. F. (2015). Individual differences in response of dorsomedial prefrontal cortex predict daily social behavior. *Social Cognitive and Affective Neuroscience*, *11*(1), 121–126. https://doi.org/10.1093/scan/nsv096

Preckel, K., Kanske, P., & Singer, T. (2018). On the interaction of social affect and cognition: empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences*, *19*, 1–6. https://doi.org/10.1016/j.cobeha.2017.07.010

Preis, M. A., Kröner-Herwig, B., Schmidt-Samoa, C., Dechent, P., & Barke, A. (2015). Neural correlates of empathy with pain show habituation effects. An fMRI study. *PLoS ONE*, *10*(8), 1–19. https://doi.org/10.1371/journal.pone.0137056

Preston, S. D. (2007). A perception-action model for empathy. In *Empathy in Mental Illness*. https://doi.org/10.1017/CBO9780511543753.024

Promberger, M., & Marteau, T. M. (2013). When do financial incentives reduce intrinsic motivation? Comparing behaviors studied in psychological and economic literatures. *Health Psychology*. https://doi.org/10.1037/a0032727

R Core Team. (2019). R: A Language and Environment for Statistical Computing.

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

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677. https://doi.org/10.1038/ncomms4677

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 870–888. https://doi.org/10.1037/a0034954

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Reeve, J., & Lee, W. (2012). Neuroscience and Human Motivation. In *The Oxford Handbook of Human Motivation*. https://doi.org/10.1093/oxfordhb/9780195399820.013.0021

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W.F. Prokasy (Eds.), *Clasical conditioning II: current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Rijnders, R. J. P., Terburg, D., Bos, P. A., Kempes, M. M., & van Honk, J. (2021). Unzipping empathy in psychopathy: Empathy and facial affect processing in psychopaths. *Neuroscience and Biobehavioral Reviews*, *131*, 1116–1126. https://doi.org/10.1016/j.neubiorev.2021.10.020

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*. https://doi.org/10.1146/annurev.psych.121208.131647

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*, 1694–1703. https://doi.org/10.1016/j.neuroimage.2004.04.015

Roberts, I. D., & Hutcherson, C. A. (2019). Affect and Decision Making: Insights and Predictions from Computational Models. *Trends in Cognitive Sciences*, *23*(7), 602–614. https://doi.org/10.1016/j.tics.2019.04.005

Robinson, S., Sotak, B. N., During, M. J., & Palmiter, R. D. (2006). Local dopamine production in the dorsal striatum restores goal-directed behavior in dopamine-deficient mice. *Behavioral Neuroscience*, *120*(1), 196. https://doi.org/10.1037/0735-7044.120.1.000

Rode, J., Gómez-Baggethun, E., & Krause, T. (2015). Motivation crowding by economic incentives in conservation policy: A review of the empirical evidence. *Ecological Economics*, *117*, 270–282. https://doi.org/10.1016/j.ecolecon.2014.11.019

Roiser, J. P., Linden, D. E., Gorno-Tempinin, M. L., Moran, R. J., Dickerson, B. C., & Grafton, S. T. (2016). Minimum statistical standards for submissions to Neuroimage: Clinical. *NeuroImage: Clinical*, *12*, 1045–1047. https://doi.org/10.1016/j.nicl.2016.08.002

Rumble, A. C., Van Lange, P. A. M., & Parks, C. D. (2010). The benefits of empathy: When empathy may sustain cooperation in social dilemmas. *European Journal of Social Psychology*, *40*(5), 856–866.

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, *1*(2), 114–117. https://doi.org/10.1111/j.2041-210x.2010.00014.x

Saad, G., & Gill, T. (2001). The effects of a recipient's gender in a modified dictator game. *Applied Economics Letters*, *8*(7), 463–466. https://doi.org/10.1080/13504850010005260

Saarela, M. V., Hlushchuk, Y., Williams, A. C. D. C., Schürmann, M., Kalso, E., & Hari, R. (2007). The compassionate brain: Humans detect intensity of pain from another's face. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhj141

Salamone, J. D., & Correa, M. (2012). The Mysterious Motivational Functions of Mesolimbic Dopamine. *Neuron*, *76*(3), 470–485. https://doi.org/10.1016/j.neuron.2012.10.021

Salamone, J. D., Pardo, M., Yohn, S. E., López-Cruz, L., Sanmiguel, N., & Correa, M. (2016). Mesolimbic dopamine and the regulation of motivated behavior. In *Behavioral neuroscience of motivation* (pp. 231–257). Springer, Cham. https://doi.org/10.1007/7854_2015_383

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G. (2022). The neural computation of human prosocial choices in complex motivational states. *NeuroImage*, *247*, 118827. https://doi.org/10.1016/J.NEUROIMAGE.2021.118827

Saxe, R. R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schier, U. K., Ockenfels, A., & Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology*, *56*, 107–115. https://doi.org/10.1016/j.joep.2016.06.004

Schmidt, L., Tusche, A., Manoharan, N., Hutcherson, C., Hare, T., & Plassmann, H. (2018). Neuroanatomy of the vmPFC and dlPFC predicts individual differences in cognitive regulation during dietary self-control across regulation strategies. *Journal of Neuroscience*, *38*(25), 5799–5806. https://doi.org/10.1523/JNEUROSCI.3402-17.2018

Schuch, S., & Pütz, S. (2021). Mood state and conflict adaptation: an update and a diffusion model analysis. *Psychological Research*, *85*(1), 322–344.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience \& Biobehavioral Reviews*, *42*, 9–34.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., … Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: A review using probabilistic atlases from different imaging modalities. *Human Brain Mapping*, *38*(9), 4788–4805. https://doi.org/10.1002/HBM.23675

Servant, M., Montagnini, A., & Burle, B. (2014). Conflict tasks and the diffusion framework: Insight in model constraints based on psychological laws. *Cognitive Psychology*, *72*, 162–195.

Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist*, *17*(1), 18–24.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617–627. https://doi.org/10.1093/brain/awn279

Shamay-Tsoory, S. G., & Hertz, U. (2022). Adaptive Empathy: A Model for Learning Empathic Responses in Response to Feedback. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916211031926

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*. https://doi.org/10.1016/j.neuron.2010.07.020

Shiban, Y., Wittmann, J., Weißinger, M., & Mühlberger, A. (2015). Gradual extinction reduces reinstatement. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00254

Shiota, M. N., Papies, E. K., Preston, S. D., & Sauter, D. A. (2021). Positive affect and behavior change. *Current Opinion in Behavioral Sciences*, *39*, 222–228.

Shohamy, D. (2011). Learning and motivation in the human striatum. *Current Opinion in Neurobiology*, *21*(3), 408–414. https://doi.org/10.1016/j.conb.2011.05.009

Simpson, B., & Willer, R. (2008). Altruism and indirect reciprocity: The interaction of person and situation in prosocial behavior. *Social Psychology Quarterly*. https://doi.org/10.1177/019027250807100106

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., & Hein, G. (2012). Human empathy through the lens of psychology and social neuroscience. In F. B. M. de Waal & F. P. Francesco (Eds.), *The primate mind: Built to connect with other minds.*

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology*. https://doi.org/10.1016/j.cub.2014.06.054

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/j.1749-6632.2009.04418.x

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*. https://doi.org/10.1126/science.1093535

Small, D. M., Jones-Gotman, M., & Dagher, A. (2003). Feeding-induced dopamine release in dorsal striatum correlates with meal pleasantness ratings in healthy human volunteers. *NeuroImage*. https://doi.org/10.1016/S1053-8119(03)00253-2

Son, J. Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific Reports*. https://doi.org/10.1038/s41598-019-48050-2

Stanley, D. A. (2016). Getting to know you: general and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience*, *11*(4), 525–536. https://doi.org/10.1093/SCAN/NSV145

Steinbeis, N. (2016). The role of self-other distinction in understanding others' mental and emotional states: Neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150074. https://doi.org/10.1098/rstb.2015.0074

Stietz, J., Jauk, E., Krach, S., & Kanske, P. (2019). Dissociating empathy from perspective-taking: Evidence from intra- And inter-individual differences research. *Frontiers in Psychiatry*, *10*, 126. https://doi.org/10.3389/fpsyt.2019.00126

Stojić, H., Schulz, E., P Analytis, P., & Speekenbrink, M. (2020). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, *149*(10), 1878.

Stoop, J., van Soest, D., & Vyrastekova, J. (2018). Rewards and cooperation in social dilemma games. *Journal of Environmental Economics and Management*, *88*(C), 300–310. https://doi.org/10.1016/j.jeem.2017.12.007

Strait, C. E., Sleezer, B. J., & Hayden, B. Y. (2015). Signatures of value comparison in ventral striatum neurons. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1002173

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*, *112*, 1619–1624. https://doi.org/10.1073/pnas.1414715112

Tabibnia, G., & Lieberman, M. D. (2007). Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1118*, 90–101. https://doi.org/10.1196/ANNALS.1412.001

Takeuchi, R., Bolino, M. C., & Lin, C. C. (2015). Too many motives? The interactive effects of multiple motives on organizational citizenship behavior. *Journal of Applied Psychology*, *100*(4), 1239. https://doi.org/10.1037/apl0000001

Telle, N. T., & Pfister, H. R. (2016). Positive empathy and prosocial behavior: A neglected link. *Emotion Review*, *8*(2), 154–163. https://doi.org/10.1177/1754073915586817

Terlecki, M. A., & Buckner, J. D. (2015). Social anxiety and heavy situational drinking: Coping and conformity motives as multiple mediators. *Addictive Behaviors*, *40*, 77–83. https://doi.org/10.1016/j.addbeh.2014.09.008

Thompson, A., & Steinbeis, N. (2021). Computational modelling of attentional bias towards threat in paediatric anxiety. *Developmental Science*, *24*(3), e13055. https://doi.org/10.1111/desc.13055

Timmers, I., Park, A. L., Fischer, M. D., Kronman, C. A., Heathcote, L. C., Hernandez, J. M., & Simons, L. E. (2018). Is empathy for pain unique in its neural correlates? A meta-analysis of neuroimaging studies of empathy. *Frontiers in Behavioral Neuroscience*, *12*, 289. https://doi.org/10.3389/fnbeh.2018.00289

Titmuss, R. M. (1970). *The Gift Relationship: From Human Blood to Social Policy. Original ed. with new chapters ed. by Ann Oakley and John Ashton ed. New York, NY: New Press*. London: Allen & Unwin.

Toelch, U., Panizza, F., & Heekeren, H. R. (2018). Norm compliance affects perceptual decisions through modulation of a starting point bias. *Royal Society Open Science*, *5*(3), 171268. https://doi.org/10.1098/rsos.171268

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

Tusche, Anita, & Bas, L. M. (2021). Neurocomputational models of altruistic decision-making and social motives: Advances, pitfalls, and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(6), e1571.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

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*. https://doi.org/10.1093/scan/nsp009

Van Dijk, E., & De Dreu, C. K. W. (2021). Experimental Games and Social Decision Making. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-081420-110718

van Dongen, J. D. M. (2020). The Empathic Brain of Psychopaths: From Social Science to Neuroscience in Empathy. *Frontiers in Psychology*, *11*, 695. https://doi.org/10.3389/fpsyg.2020.00695

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*. https://doi.org/10.1002/hbm.20547

Van Rossum, G. (2007). Python programming language. *Paper Presented at the USENIX Annual Technical Conference.*

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*. https://doi.org/10.3758/BRM.40.1.61

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical Diffusion Models for Two-Choice Response Times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., … Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2005.07.022

Von Biebersteinid, F., Esslid, A., & Friedrichid, K. (2021). Empathy: A clue for prosocialty and driver of indirect reciprocity. https://doi.org/10.1371/journal.pone.0255071

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/1232672

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory and Cognition*, *32*(7), 1206–1220. https://doi.org/10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*. https://doi.org/10.3758/BF03192967

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Wallace, D. L., Aarts, E., Dang, L. C., Greer, S. M., Jagust, W. J., & D'Esposito, M. (2014). Dorsal striatal dopamine, food preference and health perception in humans. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0096319

Walter, H. (2012). Social Cognitive Neuroscience of Empathy: Concepts, Circuits, and Genes. *Emotion Review*, *4*(1), 9–17. https://doi.org/10.1177/1754073911421379

Wei, L. T., & Yazdanifard, R. (2014). The impact of Positive Reinforcement on Employees' Performance in Organizations. *American Journal of Industrial and Business Management*, *4*(1), 4. https://doi.org/10.4236/ajibm.2014.41002

White, C. N., Curl, R. A., & Sloane, J. F. (2016). Using Decision Models to Enhance Investigations of Individual Differences in Cognitive Neuroscience. *Frontiers in Psychology*, *7*, 81. https://doi.org/10.3389/fpsyg.2016.00081

White, C. N., Liebman, E., & Stone, P. (2018). Decision mechanisms underlying mood-congruent emotional classification. *Cognition and Emotion*, *32*(2), 249–258. https://doi.org/10.1080/02699931.2017.1296820

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*, 1686. https://doi.org/10.21105/joss.01686

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Wilhelm, M. O., & Bekkers, R. (2010). Helping behavior, dispositional empathic concern, and the principle of care. *Social Psychology Quarterly*, *73*(1), 11–32. https://doi.org/10.1177/0190272510361435

Williams, K. M., Nathanson, C., Paulhus, D. L., & others. (2003). Structure and validity of the self-report psychopathy scale-III in normal populations. In *111th annual convention of the American Psychological Association* (pp. 1–12).

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058

Wood, R. M., Rilling, J. K., Sanfey, A. G., Bhagwagar, Z., & Rogers, R. D. (2006). Effects of tryptophan depletion on the performance of an iterated Prisoner's Dilemma game in healthy adults. *Neuropsychopharmacology*. https://doi.org/10.1038/sj.npp.1300932

Yamamoto, S., & Takimoto, A. (2012). Empathy and fairness: Psychological mechanisms for eliciting and maintaining prosociality and cooperation in primates. *Social Justice*

Research, *25*(3), 233–255.

Yeung, A. W. K. (2018). An updated survey on statistical thresholding and sample size of fMRI studies. *Frontiers in Human Neuroscience*, *12*(1), 1–7. https://doi.org/10.3389/fnhum.2018.00016

Yu, H., Siegel, J. Z., Clithero, J. A., & Crockett, M. J. (2021). How peer influence shapes value computation in moral decision-making. *Cognition*. https://doi.org/10.1016/j.cognition.2021.104641

Zaki, J. (2014). Empathy: a motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/A0037679

Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*(5), 675–680. https://doi.org/10.1038/nn.3085

Zhang, J., & Rowe, J. B. (2014). Dissociable mechanisms of speed-accuracy tradeoff during visual perceptual learning are revealed by a hierarchical drift-diffusion model. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2014.00069

Zhou, Y., Lindström, B., Soutschek, A., Kang, P., Tobler, P. N., & Hein, G. (2022). Learning from ingroup experiences changes intergroup impressions. *Journal of Neuroscience*, *42*(36), 6931–6945.

5 Appendix

Manuscripts and publications in peer reviewed journals

Saulin, A., Ting, C.C., Engelmann, J.B., & Hein, G. (manuscript) Empathy induces sustained social closeness.

Saulin, A. & Hein, G. (manuscript) Empathy incites a sustainable prosocial decision bias.

*Weiß, M., *Saulin, A., Iotzov, V., Hewig, J., & Hein, G. (registered report: in principle acceptance) Can monetary incentives overturn fairness-based decisions? *Royal Society Open Science**equal contribution

Iotzov, V., Saulin, A., Kaiser, J., Han, S., & Hein, G. (2022) Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females. *Social Neuroscience*

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G. (2022) The neural computation of human prosocial choices in complex motivational states. *Neuroimage.*

Contribution at conferences

Saulin, A., Ting, C.C., Engelmann, J.B., & Hein, G.. 24 June 2022; 6th Science Conference of the Center of Mental Health, Center of Mental Health, Würzburg, Germany; Poster “Learned empathy results in sustainable social closeness”

Saulin, A., Ting, C.C., Engelmann, J.B., & Hein, G.. Talk “Learning prosocial motives: Modelling empathy and reciprocity driven closeness”. 16 – 18 June 2022; 47th Psychology and the Brain conference, Deutsche Gesellschaft für Psychologie, Freiburg, Germany;

Saulin, A., Ting, C.C., Delius, K., Engelmann, J.B., & Hein, G.. Talk “Learning prosocial motives”. 10 – 11 June 2021; 17th NeuroPsychoEconomics Conference, Association for NeuroPsychoEconomics, Amsterdan (virtual), The Netherlands

Saulin, A., Ting, C.C., Delius, K., Engelmann, J.B., & Hein, G.. Talk “Learning prosocial motives”. 14 – 16 Mar 2021 ;63rd Conference of experimental psychologists, Deutsche Gesellschaft für Psychologie, Ulm (virtual), Germany

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G.. Poster "How multiple motives affect the computation of social decisions in the human brain". 20 – 22 June 2019; 45th Psychology and the Brain conference, Deutsche Gesellschaft für Psychologie, Dresden, Germany

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G.. Poster "How different motives interact in the human brain". 10 – 13 June 2019; 25th Annual Meeting of the Organization for Human Brain Mapping (OHBM), OHBM, Rome, Italy

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G.. Talk "How multiple motives affect the computation of social decisions in the human brain". 20 - 23 2019; 13th Göttingen Meeting of the German Neuroscience society, German Neuroscience Society, Göttingen, Germany

Affidavit

I hereby confirm that my thesis entitled “Sustainability of empathy as driver for prosocial behavior and social closeness: insights from computational modelling and functional magnetic resonance imaging” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis. Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Anne Saulin Würzburg
____
Doctoral Researcher’s Name Date Place Signature

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation „Nachhaltigkeit von Empathie als Motiv für prosoziales Verhalten und soziale Nähe: Erkenntnisse auf Grundlage von computational modelling und funktioneller Magnetresonanztomographie“ eigenständig, d.h. insbesondere selbstständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.
Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Anne Saulin Würzburg
____
Doctoral Researcher’s Name Date Place Signature

Individual author contributions

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions and of legal second publication rights

Publication (complete reference):

Saulin, A., Ting, C.-C.., Engelmann, J.B., & Hein, G. (in prep). Empathy induces sustained social closeness.

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design	AS GH	JBE			
Methods Development	AS CC	GH	JBE		
Data Collection	AS				
Data Analysis and Interpretation	AS CC	GH	JBE		
Manuscript Writing					
Writing of Introduction	AS	GH			
Writing of Materials & Methods	AS	CC	GH		
Writing of Discussion	AS	GH			
Writing of First Draft	AS				

Explanations (if applicable):

Publication (complete reference):

Saulin, A.& Hein, G. (in prep). Empathy incites a sustainable prosocial decision bias.

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design	AS GH				
Methods Development	AS	GH			
Data Collection	AS				
Data Analysis and Interpretation	AS	GH			
Manuscript Writing					
Writing of Introduction	AS	GH			
Writing of Materials & Methods	AS				
Writing of Discussion	AS	GH			
Writing of First Draft	AS				

Explanations (if applicable):

Publication (complete reference):

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G. (2022). The neural computation of human prosocial choices in complex motivational states. *NeuroImage*, *247*, 118827.

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design	GH AS	JK	ML		
Methods Development	GH AS	JK	ML		
Data Collection	AS UH	ML			
Data Analysis and Interpretation	AS	GH	UH	JK	
Manuscript Writing					
Writing of Introduction	AS GH				
Writing of Materials &	AS	UH	GH		
Methods	AS	UH	GH		
Writing of Discussion	AS GH				
Writing of First Draft	AS	GH			

Explanations (if applicable):

Publication (complete reference):

Iotzov, V., Saulin, A., Kaiser, J., Han, S., & Hein, G. (2022) Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females. *Social Neuroscience*

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design	GH VI	JK			
Methods Development	VI	GH	AS	JK	SH
Data Collection	VI				
Data Analysis and Interpretation	VI	GH	AS	JK	SH
Manuscript Writing					
Writing of Introduction	VI GH	AS			
Writing of Materials &	VI	AS			
Methods					
Writing of Discussion	VI GH	AS			
Writing of First Draft	VI GH				

Explanations (if applicable):

The doctoral researcher confirms that she/he has obtained permission from both the publishers and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Anne Saulin 28.09.2022 Würzburg

Doctoral Researcher's Name Date Place Signature

Prof. Dr. Grit Hein 21.09.2022 Würzburg

Primary Supervisor's Name Date Place Signature

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions to figures/tables/chapters included in the manuscripts

Publication (complete reference):

Saulin, A., Ting, C.-C.., Engelmann, J.B., & Hein, G. Empathy induces sustained social closeness

Figure	**Author Initials,** Responsibility decreasing from left to right				
Fig 1	AS	GH			
Fig 2	AS	CCT	GH		
Fig 3	AS	CCT	GH		
Table 1	AS	GH	CCT	JBE	
Fig 4	AS	CCT	GH		
Table 2	AS	GH	JBE		
Fig 5	AS	GH	JBE		

(If required please use more than one sheet)

Explanations (if applicable):

Publication (complete reference):

Saulin, A.& Hein, G. Empathy incites a sustainable prosocial decisions bias

Figure	**Author Initials,** Responsibility decreasing from left to right				
Fig 1	AS	GH			
Table 1	AS	GH			
Fig 2	AS	GH			
Table 2	AS	GH			
Table 3	AS	GH			
Fig 3	AS	GH			
Fig 4	AS	GH			

Explanations (if applicable):

Publication (complete reference):

Saulin, A., Horn, U., Lotze, M., Kaiser, J., & Hein, G. (2022). The neural computation of human prosocial choices in complex motivational states. *NeuroImage*, *247*, 118827.

Figure	**Author Initials,** Responsibility decreasing from left to right				
Fig 1	AS	GH			
Fig 2	AS	GH			
Fig 3	AS	GH			
Fig 4	AS	GH	UH		

Explanations (if applicable):

Publication (complete reference):

Iotzov, V., Saulin, A., Kaiser, J., Han, S., & Hein, G. (2022) Financial incentives facilitate the neural computation of prosocial decisions stronger in lower empathic adult females *Social neuroscience*

Figure	**Author Initials,** Responsibility decreasing from left to right				
Fig 1	VI	AS	GH		
Fig 2	VI	AS	GH		
Fig 3	VI	GH	AS		
Table 1	VI	GH			
Fig 4	VI	AS	GH		
Fig 5	VI	AS	GH		
Fig 6	VI	AS	GH		

Explanations (if applicable):

I also confirm my primary supervisor's acceptance.

Anne Saulin	28.09.2022	Würzburg	
Doctoral Researcher's Name	Date	Place	Signature