The (Dis)Honest and (Un)Fair Brain

Investigating the Neural Underpinnings of Moral Decisions

The (Dis)Honest and (Un)Fair Brain

Investigating the Neural Underpinnings of Moral Decisions

Het (on)eerlijke brein

Onderzoek naar het neurale substraat van morele keuzes

Thesis

to obtain the degree of Doctor from the

Erasmus University Rotterdam

by command of the

rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Friday 19 November 2021 at 13.00 hrs

by

Sebastian Peter Heinz Speer

born in Bad Friedrichshall (Germany)

Erasmus University Rotterdam

Doctoral Committee

Promotor: Prof. dr. A. Smidts

Other members: Prof. dr. E.A.M. Crone

Prof. dr. A.G. Sanfey

Dr. N. Abe

Co-promotor: Dr. M.A.S. Boksem

Erasmus Research Institute of Management – ERIM

The joint research institute of the Rotterdam School of Management (RSM)

and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam

Internet: www.erim.eur.nl

ERIM Electronic Series Portal: repub.eur.nl/

ERIM PhD Series in Research in Management, 537

ERIM reference number: EPS-2021- 537-MKT

ISBN 978-90-5892-616-6

Cover design by: Jannika Speer

Design: PanArt, www.panart.nl

Print: OBT bv, www.obt.eu

This publication (cover and interior) is printed on FSC® paper Magno Satin MC.

FSC
www.fsc.org
MIX
Papier van
verantwoorde herkomst
FSC® C022586

Chapter 1
Introduction

"*Always do what is right. It will gratify half of mankind and astound the other*." – Mark Twain

Our daily life abounds with situations that confront us with a conflict between selfish urges and virtuous alternatives that benefit the welfare of others. Do we donate a recently received windfall to charity or do we spend it indulgingly on a luxurious dinner? Do we incriminate a colleague to obtain an advantageous position in a hiring process? Do we deliberately fail to report income we receive to avoid paying taxes? Do we decide to cheat on an exam, or do we risk failing? Do we illegally stream a movie we would otherwise have to pay for? In all these examples we are exposed to the temptation to violate moral standards to serve our selfish interests. As Mark Twain alluded to in the quote above, not everyone solves these conflicts in the same manner. While highly virtuous individuals such as Martin Luther King or Mother Theresa may devote their lives to improve the condition of the oppressed or the less fortunate, more selfish individuals such as Jordan 'Wolf of Wallstreet' Belfort, who greedily engaged in fraudulent stock market manipulations that led to investor losses of more than 200 million US Dollars, tend to focus on maximizing their own gains and in some cases even exploit others. These large differences in how individuals weigh their own benefits against moral standards and social norms that may benefit the welfare of others, are pivotally important in understanding moral decision-making that is the fundament of cooperation on a societal level. The aim of this dissertation is to combine neuroimaging methods with behavioral experiments to investigate the neurocognitive underpinnings of individual differences in moral decision-making.

1.1 Cheating/honesty vs fairness/unfairness

To provide guidance for the reader and delineate the scope of this dissertation, it is important to note that morality is generally conceived of as a concept that consists of five universal moral foundations: Harm/Care, Fairness/Honesty, Ingroup/Loyalty, Authority/Respect and Purity/Sanctity (Graham et al., 2011). This dissertation will focus exclusively on the foundation of Fairness/Honesty: that is, impartial and just treatment or behavior without favoritism, discrimination or deviation from the truth. Within this domain we must further distinguish between fairness which involves a directly identifiable victim, and honesty when no such immediate victim can be identified.

Most of the research on Fairness/Honesty so far, especially in social neuroscience, has been concerned with unfair or dishonest behavior in which a directly identifiable victim is present. In the first paragraph, the first three examples: 'not investing in a charity', 'incriminating a colleague', and 'taking credit for someone else's idea', are all examples of selfish (unfair) behavior where a victim of this selfishness can be easily identified. Experimental paradigms measuring unfairness or dishonesty with an identifiable victim are typically zero-sum games, where higher payoff for the participant results in lower payoff for the opponent, such as in sender-receiver, ultimatum or dictator games (Güth et al., 1982, Gneezy, 2005).

In contrast, in the latter two examples: 'cheating on an exam' and 'engaging in software or movie piracy', there is no clearly identifiable victim. Often the counterpart or victim is an organization, such as a university in case of plagiarism, a clothing company in the case of wardrobing or big tech companies in the case of software piracy, and thus remains less identifiable and more ambiguous. Previous studies that have investigated dishonest behavior without identifiable victim used experimental paradigms such as the coin-flip task (Bucciol & Piovesan, 2011) or die-roll task (Fischbacher & Heusi, 2013), in which it is not specified what happens to the money that is not paid out: the more someone cheats, the more money is paid out (positive sum games). Consequently, there is no identifiable individual victim, and the negative consequences of cheating are less clear.

Therefore, when engaging in unfair or deceptive behavior towards an identifiable victim, feelings of guilt towards the other may arise as compared to when cheating on an exam or illegally downloading a music file, where the moral conflict may be more internal and focused on the tension between financial reward and one's moral standards.

In this dissertation, one chapter (Chapter 6) will be devoted to study the neural mechanisms underlying (un)fair decisions with an identifiable victim, whereas the remaining chapters will focus on dishonest behavior without identifiable victim.

In accordance with this distinction (identifiable versus no identifiable victim), a meta-analysis on neuroimaging research on moral decision-making in the Fairness/Honesty domain, conducted by Lisofsky and colleagues (2014), reported that experimental deception paradigms that involved an identifiable victim, unfair decisions were associated with increased activation in the right temporal parietal junction and the bilateral temporal pole, which have been associated consistently with theory of mind processes (Bahnemann et al., 2010) as compared to less interactive studies without identifiable victim. In the domain of fairness, there is converging evidence that perspective taking, or also termed theory of mind (ToM), is a crucial element in promoting prosocial sharing (e.g., Eisenberg and Miller, 1987; Batson et al., 1991; Penner et al., 2005; Pavey et al., 2012; Edele et al., 2013; Artinger et al., 2014). Perspective taking here refers to the process of taking the viewpoint of another person, to better understand their motivations, emotions and thoughts. Previous fMRI studies research has identified the ToM network as the basic system that enables social understanding (Saxe and Wexler, 2005; Schurz et al., 2014; Young et al., 2010b; for an overview on the neural correlates see Section 1.6). Based on these findings, Lisofsky et al. (2014) argue that, particularly in studies involving social interaction and an identifiable victim, perspective-taking processes are important.

In contrast, contexts in which there is no identifiable victim present and dishonest behavior is targeted at an anonymous organization, less perspective taking may be involved. Consequently, brain regions involved in moral decisions may be activated differently. While research on fairness and honesty with identifiable victims may be a valuable source of insights, as the two types certainly share some neural substrates such as regions related to cognitive control and affective processing (for details see Lisofsky et al., 2014), cheating behavior without identifiable victim also demands more in-depth exploration. This is because many of the costliest forms of dishonest behavior, such as tax deception or software and music piracy do not have a directly identifiable victim.

To be concise, we will from now on refer to the context with identifiable victim as fairness/selfishness and to the context without identifiable victim as honesty/cheating.

1.2 Will vs Grace

When exposed to the opportunity to cheat or be selfish, clearly, (monetary) rewards play a crucial role. In the context of cheating: the higher the reward, the more attractive the decision to cheat (Becker, 1968; Allingham & Sando, 1972). Accumulating research has demonstrated that greedy people find a variety of moral transgressions more acceptable and engage in such transgressions more often as compared to less greedy people (Seuntjens et al., 2019). Accordingly, neural responses in anticipation of reward have been shown to be predictive of (dis)honesty in a subsequent unrelated task (Abe & Greene, 2014). Similarly, in the context of selfishness, standard economic theory prescribes that a purely rational actor is driven by his own self-interest and should consequently maximize his or her own reward (Edgeworth, 1881; Kahneman et al., 1986; Forsythe et al., 1994), rather than sharing with others. Collectively, these findings emphasize that higher rewards and stronger sensitivity to reward should increase the likelihood of dishonest and selfish behavior.

In the last decades, research from psychology, economics, and neuroscience has generated evidence, however, that people care about more than only maximizing their own monetary payoff, which is evident in the abundance of prosocial behaviors such as honesty, altruism and reciprocity. Humans have adopted social norms and use these as a guideline and benchmark against which they compare their own behavior (Campbell, 1964, Henrich et al., 2001).

As already mentioned above, with regard to fairness, it is commonly agreed upon that perspective taking is a strong determinant of prosocial sharing (e.g. Eisenberg and Miller, 1987; Batson et al., 1991; Penner et al., 2005; Pavey et al., 2012; Edele et al., 2013; Artinger et al., 2014). The reasoning here is that taking the perspective of the other will lead to considering the consequences of one’s actions for the other person and the resulting thoughts and feelings. As a result of perspective taking, people may then refrain being selfish to avoid the negative consequences for the other. For example, in the context of the Ultimatum Game, when proposers were instructed to imagine being in the position of a responder, they offered significantly higher amounts of money (Hoffman et al., 2000). It has also been found that perspective taking enhances a preference for fairness in children (Takagishi et al., 2010). In this study the authors found that preschoolers who had acquired theory of mind proposed significantly higher offers than children who had not yet acquired theory of mind. In addition, neuroimaging evidence suggests that neural responses related to ToM promote prosocial behavior, that is participants who showed higher activity in brain regions associated with theory of mind and empathy in response to viewing someone else in pain, were more likely to engage in prosocial behavior towards the victim later (Masten, Morelli & Eisenberger, 2011).

In the context of honesty, the way we view ourselves, our self-concept (Bem, 1972; Aronson, 1969, Baumeister, 2010), may allow us to refrain from cheating (Mazar et al., 2008). Humans highly appreciate honesty in others and have deep faith in their own morals (Dhar & Wertenbroch, 2012). Thus, violating one’s own moral standards will demand a negative update of one’s self-concept which is highly unpleasant (Berthoz et al., 2006). In line with this self-concept maintenance theory, behavioral research shows that people indeed behave dishonesty enough to profit from an opportunity to cheat but yet honesty enough to still maintain their believes in their own integrity (Mazar et al., 2008). Stated differently, a small degree of dishonesty provides a taste of profit but prevents ruining a positive self-image.

It becomes apparent that there is a conflict between the selfish urge to obtain monetary reward and maintaining a positive self-concept and considering the welfare of others. Several lines of research put forward that cognitive control is needed to resolve this tension (Abe & Greene, 2014, Gino et al., 2011; Greene & Paxton, 2009; Marechal et al., 2017; Mead et al., 2009). Self-control is the psychological capacity that allows an individual to enact behaviors that are consistent with his/her long-term goals (e.g., of being an ethical person) and refrain from engaging in behaviors that are driven by short-term, selfish interests such as dishonest behavior. However, while it is clear that cognitive control is instrumental in reconciling this conflict, the precise role of cognitive control in moral decisions remains controversial (Köbis et al., 2019). Two competing theories have been proposed: the 'Will' and the 'Grace' hypothesis (Greene & Paxton, 2009).

The *Will hypothesis* proposes that humans have a natural inclination to be selfish and dishonest and that to overcome their immoral impulses, deliberate cognitive control is required. 'Will' thus refers to the willpower needed to remain honest or fair. According to this hypothesis, honesty and fairness are the product of the arduous resistance to temptation (see also Bereby-Meyer & Shalvi, 2015), similar to the cognitive control processes that allow individuals to delay gratification (McClure et al., 2004). This proposition aligns with behavioral research reporting that participants who are cognitively depleted by challenging exercises, sleep deprived, or under time pressure are more inclined to cheat or act selfishly (Gino et al., 2011, Mead et al., 2009, Barnes et al., 2011, Shalvi et al., 2012). In concert, this evidence suggests that humans serve their self-interest per default and require effortful cognitive control to refrain from the temptation of selfish reward. Further, neuroimaging research has shown that activity in regions associated with cognitive control, such as the dorsolateral prefrontal cortex, are associated with overriding automatic selfish impulses (Knoch et al., 2006) and are activated when participants refrain from lying (Greene & Paxton, 2009, Abe

& Greene, 2014). When activity in the dlPFC is disrupted with transcranial magnetic stimulation (TMS; Knoch et al., 2006) or impaired by lesions, selfish impulses are not inhibited anymore and consequently prevail. Thus, these neuroimaging studies support the Will-hypothesis.

In contrast, the *Grace hypothesis* advances that humans are honest and fair per default and need cognitive control to benefit from an opportunity to cheat and be selfish on occasion. Here, 'Grace' thus refers to the natural inclination to be courteous and good mannered.

This hypothesis emphasizing moral grace is supported by research showing that people react faster when asked to tell the truth as compared to lying (for meta-analyses, see refs. Suchotzki et al., 2017, Verschure et al., 2018) and are more honest, cooperative and prosocial under time pressure (Carparo, 2017, Carparo et al., 2019). Cheating requiring cognitive capacity is also supported by findings that people cheat less when taxed by a cognitively demanding memory task as compared to a less taxing task (Van't Veer et al., 2014). Collectively, this research suggests that fairness and honesty are the default and cognitive control is necessary to override this default and act selfishly or cheat.

This Grace hypothesis also dovetails with the social intuitionist model, put forward by Haidt (2001), emphasizing the role of emotions and social intuition in moral judgement and decision-making rather than deliberate cognitive control. The model deemphasizes deliberate reasoning done by individuals and emphasizes instead the importance of social and cultural influences. Intuitionist approaches in moral psychology state that moral intuitions (including moral emotions) precede and directly cause moral judgments. Moral intuition is defined as the sudden appearance in consciousness of a moral judgment, including an affective valence (e.g., good-bad, like-dislike), without any conscious awareness of deliberation such as weighing evidence, or inferring a conclusion. Stated differently, moral judgments appear in consciousness automatically and effortlessly as the result of moral intuitions. These intuitions are shaped by previous experience and learning. This is supported by evidence showing that much of social cognition operates automatically and implicitly (Bargh & Chartrand, 1999; Greenwald & Banaji, 1995). In line with the Grace hypothesis, Haidt's social intuitionist model proposes that intuitive processes are the default, handling everyday moral judgments in a rapid, effortless, and holistic way. It is only when intuitions are conflicting, or when the social situation demands thorough examination of all aspects of a scenario, that the deliberate reasoning processes are engaged.

Neuroimaging evidence in support of the Grace hypothesis is provided by early fMRI studies using moral judgements on statements with moral connotations (Moll et al., 2002) or moral dilemmas (Greene et al., 2001, Greene & Haidt, 2002). These studies show that regions associated with self-referential thinking and perspective taking, such as the (ventro-) medial prefrontal cortex, temporal pole, superior temporal sulcus show stronger activation to personal moral dilemmas as opposed to impersonal moral or neutral statements (for discussion of the neural correlates of self-referential thinking see Section 1.6). It has been suggested that engagement of this self-referential thinking network may be associated with prosocial moral sentiments, such as guilt, compassion and empathy, which may arise automatically when exposed to an opportunity to be unfair or (dis)honest, thus supporting the Grace hypothesis.

1.3 The importance of studying individual differences in moral decision making

This dissertation aims at investigating how cognitive control resolves the conflict between an external financial reward and one's self concept and how this decision process unfolds in the brain. One possible explanation for the contradictory findings favoring either the *Will* or the *Grace hypothesis* is that people vary to the extent with which they are inclined to be honest and fair, or dishonest and selfish. As eluded to above, our social environment abounds with examples of great individual differences in prosocial and honest behavior ranging from small acts of kindness in daily life to self-sacrificing altruism in situations of extreme danger. Unfortunately, however, economic, psychological and neuroscientific research on moral decision-making in the fairness domain have so far mainly focused on explaining aggregate behavior, while paying little attention to the considerable individual differences with regard to moral default. Therefore, it is crucial to explore how individual differences in underlying neurocognitive processes can account for differences in moral decisions. So far, only few studies have explored these individual differences in moral default and their psychological and neural basis. On the behavioral level, it has been found that more dishonest or selfish behavior is linked to impulsivity and risk taking (Anderman et al., 2009), greed or reward responsiveness (Scheres & Sanfey, 2006; Seuntjens et al., 2019), creativity (Gino et al., 2012) and Machiavellianism (Tang & Chen, 2008). Neuroimaging research has supported the role of greed by demonstrating that neural responses to anticipated reward, reflected in activity in the nucleus accumbens (NAcc) predicts cheating behavior in a subsequent unrelated task (Abe & Greene, 2014).

Furthermore, neuroimaging research has found that individual differences in fairness in economic exchanges are linked to cognitive control processes (Spitzer, Fischbacher, Herrnberger, Grön, & Fehr, 2007; Steinbeis, Bernhardt, & Singer, 2011). The evidence from these studies suggests that stronger cognitive control capacities, reflected in higher cortical thickness or more activity in regions associated with cognitive control, such as the dorsolateral prefrontal cortex (dlPFC) is associated with more fairness and honesty. These findings suggest that individuals may process moral conflicts in a fundamentally different way and that these differences may be underpinned by different underlying neural mechanisms.

Therefore, this dissertation is devoted to investigating the neurocognitive mechanisms underlying individual differences in (un)fair and (dis)honest decisions. Specifically, this dissertation explores how cognitive control processes resolve the tension between selfish urges versus maintaining a positive self-concept and caring for the welfare of others. To this end, we will combine several neuroimaging methods ((rs)fMRI & EEG) with different behavioral paradigms to obtain comprehensive insights into the neural correlates of honesty and fairness. The details on methods used will be discussed in the remainder of the introduction.

1.4 Methods applied in the social neuroscience of fairness and honesty

1.4.1 Experimental behavioral paradigms suitable for neuroimaging research

In order to study the neural underpinnings of individual differences in fairness and honesty we need to repeatedly elicit situations with a conflict between a selfish or dishonest but rewarding option and an honest and fair but personally costly option. However, questions concerning whether to take credit for someone else's idea, donate to charity, cheat on an exam or illegally stream recently released movies, all seem intractable at face value as there are numerous factors at play. Fortunately, these conflicts can be modeled by relatively simpler experimental paradigms, that are stripped down models of more complex situations. By means of limiting the number of options and outcomes available to participants they restrict the possible motivation and associated cognitive mechanisms underlying the observed behavior.

1.4.2 Behavioral paradigms to study fairness

A common framework to investigate the conflict between fairness and selfishness is the ultimatum game (UG) (Güth et al.,1982). In the UG a proposer starts with an endowment and can propose how to divide this endowment, and a recipient can subsequently accept or reject this proposal. In case the recipient accepts the offer, the money is split as proposed, whereas both players receive nothing in case the responder rejects the offer. This paradigm imitates decisions in the real world such as a shop owner setting the price of the products or negotiations of a peace treaty. Thus, this paradigm provides a simple model enabling the investigation of social uncertainty and strategic reasoning, such as deciding whether or not to pay the bill after dining out, betraying a competitor in the hiring process, or take credit for someone else's valuable idea.

Numerous studies have found that most proposers divide the endowment approximately evenly and offer on average 40% of the initial endowment (Oosterbeek et al., 2004; Henrich et al., 2005). To understand why proposers do not keep all money for themselves, two competing mechanisms have been proposed. Proposers may divide approximately evenly because they care for the welfare of their opponent and may therefore be motivated by fairness concerns (Thaler, 1988; Oosterbeek et al., 2004). Alternatively, proposers may offer even splits to minimize the probability of (future) rejection and thus strategically maximize their monetary gains, representing selfish motives (Forsythe et al., 1994; Fehr and Schmidt, 1999).

To investigate which of these motives dominates, the UG can be paired with the dictator game (DG), in which the recipients cannot reject the offers. Consequently, there is no punishment threat in place to prevent selfish behavior. Therefore, the proposers' offers in the DG are a direct reflection of their social motivations, as they are uncontaminated by strategic considerations. Critically, proposers driven by fairness concerns are expected to offer evenly in both paradigms as the importance of welfare of the opponent is unchanged across games, whereas selfish proposers should strategically split their money evenly in the UG to minimize the probability of rejection and offer low amounts in the DG. Using the combination of these two games we can thus try to identify the neurocognitive mechanisms that dissociate decisions of selfish from more prosocial individuals. This dissertation contributes to the existing literature by pairing the combination of these two games with neuroimaging analysis and machine learning methods, and will be described in more detail in Chapter 6.

1.4.3 Paradigms to study (dis)honesty

In psychology and economics several types of behavioral paradigms to study (dis)honesty have been developed. Here, we will briefly introduce the four most commonly used paradigms. For a detailed review on the advantages and disadvantages of these paradigms, we suggest consulting the meta-analysis by Gerlach and colleagues (2019).

First, *sender-receiver* games (Gneezy, 2005), consist of two interacting participants. The *sender* is informed about the rewards for two or more choice options. The options differ regarding how lucrative the rewards are for each of the two players. However, the total amount of reward of the two participants are usually a constant sum and what one player's gains is lost to the other player. The sender can then choose whether to send a false statement, for instance "Option A will result in a higher payoff for you", or a true statement "Option B will result in a higher payoff for you". After receiving the message, the receiver picks one of the options, not knowing the actual rewards associated. However, it is within the power of the receiver to choose the reward for both players. The sender is therefore confronted with a conflict between sending a true message or a dishonest but possibly more lucrative message. The true message, if chosen by the receiver, will result in a lower reward for the sender.

Second, in *coin-flip* tasks (Bucciol & Pisovesan, 2011), players are requested to report the outcome of a self-generated coin toss, not visible to the experimenter. Reporting one specific outcome, for instance 'heads', will result in some reward, whereas resulting the other outcome, does not. The drawback of this paradigm is that the proportion of (dis)honesty can only be inferred on the aggregate level. If sufficient data from many participants are pooled, the researchers can contrast the rate of reported wins, for instance 75% to the expected baseline rate of truly randomly generated coin tosses, namely 50%. In contrast to *sender-receiver* games, *coin-flip* tasks are positive sum games, in which the victim of dishonesty is not clearly identified and in which the more people cheat, the more money is paid out.

A third, and similar to the *coin-flip* tasks, is a *die-roll* task (Fischbacher & Heusi, 2013). Here, players are requested to report the outcome of a roll of die. In contrast to the *coin-flip* task, the outcome of the *die-roll* task is continuous. Typically, the higher the outcome the more reward. As with the *coin-flip* task, the limitation of this paradigm is that the proportion of (dis)honesty can only be inferred on the group level. Just as the *coin-flip* task, the *die-roll* task is a positive sum game.

Fourth, the matrix task exposes participants to a number of matrices, which contain 12 three-digit number, such as 3.87 (Mazar et al., 2008). The goal is to identify the number pair in every matrix that sums to exactly 10.00. Yet, most participants do not manage to solve all matrices within the time limit. After the time limit is reached, participants are requested to report how many matrices they have solved correctly and are rewarded accordingly. Participants are assigned to two conditions: one in which an experimenter grades the matrices, which leaves no opportunity to cheat, and one which is self-graded, in which participants can cheat. The proportion of cheating can then be estimated on the group level by comparing the amount of cheating in the experimenter-graded group to the self-graded group. As the two preceding paradigms, this is a positive sum game.

1.4.4 Challenges in studying (dis)honesty in combination with neuroimaging methods

Developing suitable experimental paradigms to elicit dishonest behavior in the MRI scanner or when wearing an EEG cap is extremely challenging. The goal is to evoke (dis)honesty that is voluntary, spontaneous and unobtrusive. However, the fact that participants are in a scanner while their brains are scanned evokes the sense of being observed which complicates the goal of eliciting voluntary, spontaneous and unobtrusive (dis)honesty.

To evoke dishonesty in the scanner, most early studies requested participants to answer incorrectly in one condition and to tell the truth in another. A prominent example of such an experimental procedure is the Guilty Knowledge Test (GKT, Lykken, 1959), in which participants are asked to deny that they have seen certain stimuli. Several fMRI studies using different versions of this paradigm converge on the finding that a network of regions associated with cognitive control processes is consistently activated in association with this instructed form of dishonesty (Spence et al., 2001; Langleben et al., 2002; Abe et al., 2006; Gamer et al., 2007; Kozel et al., 2009; Abe, 2009; Christ et al., 2009). However, a considerable methodological concern with these studies is that they relied on instructed lies, often in hypothetical scenarios. This instructed and hypothetical nature of these paradigms differs significantly from actual dishonest behavior, which is spontaneous and voluntary as opposed to instructed and has real life consequences rather than being hypothetical. It is therefore reasonable to assume that the neural underpinnings observed in these studies may not accurately reflect the neural mechanisms underlying voluntary and spontaneous (dis)honesty. In addition, all these studies focused on dishonesty with an identifiable victim, which highlights the lack of studies on cheating behavior without identifiable victim.

To our knowledge, the only studies that allowed participants to engage in voluntary and spontaneous cheating behavior without identifiable victim were conducted by Greene and Paxton (2009) and Abe & Greene (2014). In their studies, participants engaged in the *coin-flip* task. To realize this in the scanner, participants were asked to predict the outcomes of random computerized coin-flips and were financially rewarded for accuracy and punished for inaccuracy. As explained above, in the experimental condition participants were given the chance to cheat on a series of trials as they were rewarded based on their self-reported accuracy of their predictions. To maximize the monetary reward, they could thus exaggerate the accuracy of their predictions. Their results indicated that more dishonest participants recruited the cognitive control network (anterior cingulate cortex (ACC) and dlPFC) more strongly than honest individuals. As discussed above, the main limitation of the studies by Greene & Paxton (2009) and Abe & Greene (2014) is that the paradigm does not allow assessing whether the participants cheated on a trial-by-trial basis but can only be estimated on the aggregate level. As a result, the design did not allow assessing the cheating magnitude or whether cheating evolves over time.

Another important issue to consider in experimental paradigms measuring voluntary and spontaneous (dis)honesty is suspicion about the real purpose of the task. If participants are aware of the real purpose of the study, they may behave more honestly than they would usually to make a good impression on the experimenter, which is known as the social desirability bias (Edwards, 1957). Conversely, participants may feel tricked or manipulated and may react by cheating more to avenge themselves, which is commonly referred to as reactance effect (Brehms, 1989). It becomes evident that suspicion about the real purpose of the study has several potential negative consequences. For example, in the study by Abe and Greene (2014) introduced above, participants were led to believe that the experiment is about investigating paranormal abilities to predict the future and that cheating is a known but unintended by-product of the experimental design. This may raise suspicion as it seems improbable that researcher at a university investigate paranormal abilities.

1.4.5 A paradigm to study cheating unobtrusively and trial-by-trial

Due to these methodological challenges and the lack of neuroimaging research on (dis)honesty without identifiable victim, the neurocognitive mechanisms underlying this type of (dis)honesty remains elusive. Therefore, a more thorough investigation of these processes is needed. Another contribution of this dissertation is therefore to fill this gap and to investigate how cognitive control processes resolve the conflict between reward and self-concept on a trial-by-trial basis, by designing an innovative task. This task is based on a general paradigm proposed by Gai & Puntoni (2021), in which participants were able to cheat repeatedly, deliberately, and voluntarily inside the MRI scanner without suspicion of the real purpose of the experiment. Particularly, the advantage of this task, termed the Spot-The-Difference Task, is that we can directly track on which trials the participants cheated, enabling us to study within subject variation in moral decisions and its neural underpinnings. In other words, this enables us to not only identify neural mechanisms promoting honesty in general, but also explore the neural processes that underlie an honest decision made by someone who generally cheats or the decision to cheat by someone who is generally honest. Notably, participants are led to believe that the experimenter does not know that they are cheating, which is critical in light of the social desirability bias and the reactance effect introduced above. The Spot-The-Difference paradigm is therefore a behavioral paradigm that assesses cheating behavior inconspicuously on a trial-by-trial basis enabling us to study individual differences in neurocognitive processes underlying cheating behavior.

In the Spot-The-Difference task as applied here, subjects view pairs of images and are told that there are always three differences between the image pairs. Differences can be objects that are added to or removed from an image, or objects that differ in color between images. However, images can actually contain one, two, or three differences. Participants are requested to find three differences between the images. Since reward is dependent on participants stating that they find all three differences, without having to point them out, this design encourages cheating behavior (i.e., reporting having found all three differences, even when objectively fewer than three differences are actually present in the images). Participants are instructed that the purpose of the study is to study the underlying neurocognitive processes of visual search for marketing purposes such as searching for a product in an assortment or information on a website. A more detailed explanation will follow in the subsequent empirical chapters.

1.5 The motivation for using neuroimaging methods to study fairness and (dis)honesty

While the behavioral paradigms discussed above provide simplified models of complex behavior, by limiting the amount of options and outcomes available to participants, there are still several possible psychological traits or processes that may underlie every single decision. Stated differently, there is no one-to-one correspondence between a decision and an assumed underlying psychological process. In the Spot-The-Difference task, for example, a participant may be honest because they are not attracted by the opportunity to cheat or because their self-concept interferes with the pursuit of immoral reward, or because they may be afraid of being caught. Thus, while our behavioral paradigms aim at minimizing the complexity of fair or honest behavior to a more manageable set of factors, we are still inevitably exposed to the ambiguity regarding the psychological determinants of human behavior and associated cognitive processes. The motivation for combining neuroimaging methods with said behavioral paradigms is to enhance our capacity to understand the psychological and concomitant neurophysiological substrates underlying fair or honest decisions.

However, a caveat with most neuroimaging research on dishonesty so far is that inferences about the neurocognitive processes underlying (dis)honest decisions were frequently based on reverse inference, where the cognitive operations are inferred from activation in an observed region of the brain. While reverse inference can be insightful, it should be treated with caution contingent on how selective these brain regions of interest are activated by a specific cognitive process (Poldrack, 2006). Specifically, a given region of the brain may be implicated in numerous different cognitive tasks, which complicates the conclusion that activation observed in this area can be attributed to one specific function.

One approach to circumvent and reduce these reverse inference problems, is be comparing the activation maps resulting from our neuroimaging analysis to meta-analytically derived activation maps. These maps based on meta-analysis of several hundred related neuroimaging studies give an indication of how likely a region is activated by a certain task, stimulus or condition. Observing overlap between these meta-analytically derived activation maps and our own findings would thus provide more confidence that the observed activation can indeed be linked to a given cognitive process. To further reduce reverse inference, localizer tasks can be used, where the neural signature of a neurocognitive process of interest is evoked by an external task. Typically, this independent localizer is a task that is well established and validated in evoking a cognitive process of interest selectively (e.g., the Stroop or Go-NoGo task for response inhibition). The neural patterns derived from the localizer task can then be compared to the neural patterns observed in the task of interest (e.g., Spot-The-Difference task). If overlap and/or similarity is observed, confidence increases that activation observed in the task of interest (e.g., Spot-The-Difference task) can indeed be attributed to the cognitive process of interest (e.g., cognitive control). By means of taking advantage of both these methods (meta-analytical maps and localizer tasks), this dissertation contributes by more rigorously investigating the neural mechanisms underlying dishonest decisions and reducing issues of reverse inference.

In addition, the combination of neuroimaging methods with behavioral paradigms eliciting voluntary and spontaneous selfish and dishonesty behavior, may be particular fruitful as this combination allows direct and relatively uncontaminated access to the underlying psychological processes, which could be concealed in more traditional research methods. Specifically, as mentioned above, the measurement of the psychological determinants of socially undesirable behaviors, such as dishonesty, obtained by self-report methods, may suffer from social desirability bias. Participants may not want to admit or reveal that they are dishonest, selfish, impulsive, or greedy and consequently try to conceal their undesirable behavioral tendencies. Such a bias may obscure the psychological and cognitive mechanisms that determine whether someone will cheat or not and consequently make it hard to identify cheaters. Concealing motivations and cognitive processes is not possible when measuring neural activity directly when decisions are made. Therefore, neuroimaging paired with the behavioral paradigms introduced above is a promising approach to identify neural correlates of dishonesty and fairness.

1.6 Applying fMRI and EEG to study fairness and (dis)honesty

In this dissertation I apply both fMRI and EEG to study the underlying processes such as to take advantage of both the high spatial resolution of fMRI and the high temporal resolution of EEG (details about these techniques will be explained below). This dissertation thus contributes by combining these two methods and thus providing a more comprehensive view on the neurocognitive mechanisms underlying moral decisions as these two neuroimaging techniques capture different aspects of the associated neural processes. In addition, replicating findings across different neuroimaging modalities would provide convergent evidence for our results, which further reduces issues of reverse inference.

1.6.1 Functional magnetic resonance imaging (fMRI)

In a fMRI study, a participant completes a task (such as one of the behavioral paradigms introduced above) while lying supine in an MRI scanner. Throughout the task, the MRI scanner utilizes magnetic pulses to obtain precise measures of the blood flow in the participant's brain. Specifically, fMRI measures the flow of oxygenated and deoxygenated blood throughout the brain. When neurons are active, they consume more oxygen. So more oxygenated blood will flow to an area with high activity and more deoxygenated blood will leave this area. As a consequence, the rate of oxygenated to deoxygenated blood and its associated change in magnetic properties can be used to infer neural activity at a spatial scale of approximately a few millimeters. Due to the capacity of fMRI to non-invasively measure neural activity even in deep subcortical brain areas, which are crucially involved in decision-making, it is the most frequently used method in social neuroscience and neuroeconomics. Most of the studies reported in the following chapters also utilize fMRI. One of the main limitations of fMRI is its low temporal resolution, which is due to the slow response of blood flow to changes in neural activity. Nonetheless, by means of designing appropriate event-related fMRI experiments, such as the behavioral paradigm detailed above, it is still feasible to accurately map the neural response to a given experimental manipulation. With an optimal design and the right analysis methods temporal inferences up to the 100ms resolution range can be accomplished (Ogawa et al., 2000), but typically resides in the order of 1-2 seconds. In addition, even though fMRI is providing only correlational evidence, using cross-validation techniques, and comparing neural activation observed to activity maps from meta-analysis of related studies or external localizer tasks, can provide highly valuable insights into the neural mechanisms of interest.

In *Chapter 5* we also make use of resting-state fMRI (rsfMRI). To acquire rsFMRI scans participants are asked to lie in the scanner while looking at a fixation cross, to prevent them from falling asleep, without any additional tasks. Participants are asked to let their mind wander and relax for 8 minutes. Resting state FMRI captures low-frequency fluctuations in neural activity and connectivity between different regions. This means that we are measuring interactions between regions while the brain is resting. Individual differences in whole brain resting state functional connectivity is significant across individuals, and accumulating evidence has shown that an individual's functional connectome is robust and reliable across sessions over time (Cao et al., 2014; Zuo and Xing, 2014; Finn et al., 2015). RsfMRI has been applied to successfully connect functional connectivity to individual differences in self-reported personality (Nostro et al., 2018; Cai et al., 2020) and impulsivity in intertemporal choice (Li et al., 2013). In this dissertation we employ rsfMRI as it may offer us unique unobstructed insights to the neurocognitive determinants of (dis)honesty.

1.6.2 Electroencephalogram (EEG)

To combat the low temporal resolution of fMRI, EEG can be used. While it comes at a lower spatial resolution than fMRI, it provides a more granular diagnosis of the temporal evolution of different cognitive mechanisms in the order of milliseconds. Communication in the brain occurs through billions of interconnected neurons that send electrical signals to each other. This electrical activity, reflecting communication between neurons, can be measured using EEG. To record EEG, electrodes are placed in an elastic cap that is fixed on the head of the participant. The caps commonly consist of either 32 or 64 electrodes to cover the entire head, and measure neural activity recorded at all sites. EEG thus measures the potential for electrical currents to pass between different locations on the scalp, expressed in voltage changes (Luck, 2005). The electrical activity picked up by EEG reflects the summed activity of millions of neurons at the surface of the brain (Stern, Ray, Quigley, 2001), which explains the comparatively lower spatial resolution. Changes in behavior or decision-making have been found to be associated with changes in frequency (i.e., the number of oscillations per second) and amplitude of the EEG signal. Consequently, oscillations in the EEG signal are defined in terms of different frequency bands (with specific topography) that are typically associated with psychological constructs such as attention, memory or cognitive control. Due to the higher temporal resolution of EEG, it perfectly complements fMRI and in combination, these two methods have the potential to provide comprehensive insights into the neural underpinnings of fairness and honesty. In addition, pairing different neuroimaging modalities has the potential to significantly increase the convergent validity of observation made and results reported, as it allows to capture different aspects of the neural underpinnings of the neurocognitive processes of interest.

1.7 Neural networks involved in cognitive control, reward and perspective taking

As has become evident throughout this chapter, the cognitive control, reward and self-referential thinking/ToM network seem of paramount importance in determining the outcome of (un)fair and (dis)honest decisions. To assist the reader, the most important brain regions linked to the three networks are summarized below,

Reward. Regarding reward, this dissertation particularly focuses on brain regions that are consistently associated with the anticipation of reward. This is due to the fact that the anticipation of reward has been shown to increase the probability of (dis)honesty (Abe & Greene, 2014). Accumulative neuroimaging research has reliably linked the anticipation of reward to the striatum, insula and thalamus (for a review see Oldham et al., 2018). Neuroimaging research has identified the striatum and within it the nucleus accumbens as particularly important of the anticipation of reward (Knutson et al., 2001; Ballard & Knutson, 2009; Bartra et al., 2013; Oldham et al., 2018).

Cognitive control. In the context of this dissertation, cognitive control refers to the cognitive capacity that allows an individual to enact behaviors that are consistent with his long-term goals (e.g., of being an ethical person) and inhibit often predominant behaviors that are driven by short-term selfish interests such as dishonest behavior. Converging neuroimaging research has consistently reported the cognitive control network to be composed of the dlPFC, inferior frontal gyrus (IFG) and anterior cingulate cortex (ACC) amongst others. The dlPFC has been frequently implicated in cognitive control to overcome impulsive behavior (Dalwani et al., 2011; Weygandt et al., 2015), such as delaying reward (Knutson & Ballard, 2009). Similarly, IFG has been consistently associated with the inhibition of predominant responses (Wager et al., 2005; Verbruggen and Logan, 2008; Sharp et al., 2010; Stokes et al., 2011). Lastly, the ACC has been reliably linked to conflict detection in cognitive tasks that require to inhibit predominant responses, such as the Stroop task or the Go-NoGo task (MacDonald et al., 2000; Kerns et al., 2004; Wang et al., 2005)

Perspective taking. It is important to note that the self-referential thinking network involved in honest decisions (without identifiable victim) and the ToM network employed in fair decisions (with identifiable victim) are highly similar. This makes sense conceptually as they share the common underlying mechanism of perspective taking. In the context of (dis)honesty an agent has to engage in self-referential thinking, whereby the agent has to think about how her actions in the present moment match with her values derived from actions and experiences in the past. In the context of fairness, the agent has to engage theory of mind processes, where she has to take the perspective of how another person will feel and react to the agents' actions. To summarize, we will refer to the self-referential thinking and ToM network as the *perspective taking* network. Neuroimaging research has identified the perspective taking network as consisting of the medial prefrontal cortex (MPFC), bilateral temporo-parietal junctions (TPJ) and bilateral medial temporal gyri (MTG) amongst others (Saxe and Wexler, 2005; Schurz et al., 2014; Young et al., 2010). In this network, the TPJ has been reliably linked to reorienting of attention and perspective taking (Gallagher and Frith, 2003; Krall et al., 2015). In addition, the TPJ has been associated with appreciating differences between one's own and others' perspectives in cases of conflict (Van Overwalle and Baetens, 2009; Hétu et al., 2012). Similarly, neuroimaging research has found the MTG to be preferentially activated for inferring others' beliefs (Zaitchik et al., 2010; Bruneau et al., 2012). A meta-analysis of 40 fMRI experiments also found the MPFC to be an important area in perspective taking responsible for inferring others' thoughts and feeling in order to make accurate predictions of their next moves (see review Van Overwalle and Baetens, 2009, Schurz et al., 2014). Moreover, higher connectivity between these areas during rest is associated with higher-level metacognitive operations such as self-reflection and introspection (Qin & Northoff, 2011) and increased internal awareness (Vanhaudenhuyse et al., 2011; Carhart-Harris et al., 2014).

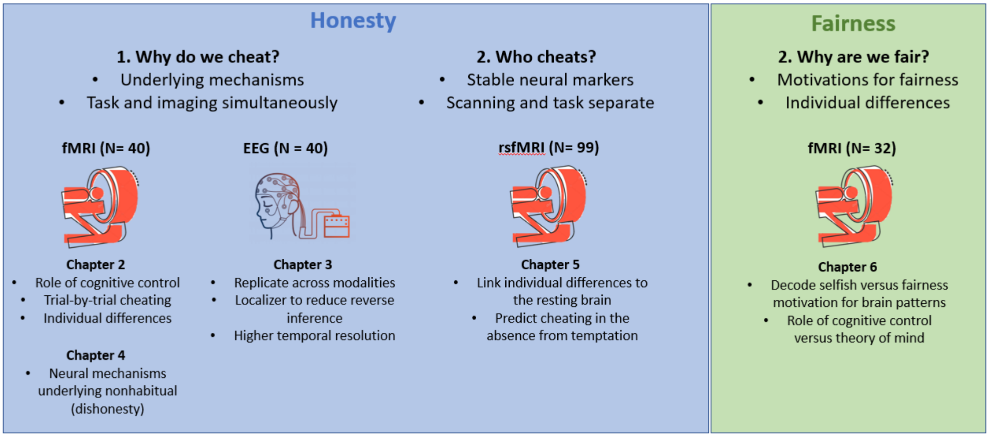
1.8 Societal relevance of studying moral decision making

Immoral behavior is an integral part of our social world as it has an impact on domains spanning from finance, marketing and politics to personal relationships. On the corporate level, companies such as Enron, Tyco and WorldCom have caused one of the biggest financial scandals in U.S. corporate history, which exemplifies the decline in business principles and ethics. A recent example of such dishonest and fraudulent behavior is the Volkswagen emissions scandal, where Volkswagen intentionally manipulated their emission data in order to be allowed to sell cars in the United States. In addition, consumers are found to behave in an increasingly immoral and dishonest manner (Vitell 2003). For instance, a survey in the U.S. on insurance fraud revealed that 25% of U.S. citizens admitted having overstated their insurance claims (Accenture, 2003). On the individual level dishonest behavior is highly prevalent, too. For instance, "wardrobing", which refers to the purchase, use and then return of used clothes costs the U.S. retail industry an estimated 16 billion dollars annually (Speights and Hilinski 2005). Another highly relevant form of immoral behavior is intellectual property theft, such as music, movie and software piracy. It has been calculated that worldwide intellectual property theft costs U.S. companies at least 250$ per year (U.S. Department of Justice 2004). However, the largest cost caused by consumer dishonesty is tax deception (e.g., omitting income; inflating deductions). According to research from the Internal Revenue Services (IRS) based on specialized audits from randomly selected individual tax returns for the time period of 2016-2020 the tax gap is estimated to be around 200 billion dollars (~1% of total tax) annually. Therefore, investigating the causes for moral misconduct and dishonesty and how this behavior can be prevented, is a highly relevant topic and of great interest to policy makers.

However, targeting dishonesty with interventions requires a thorough understanding of the underlying (neuro)cognitive processes. As introduced above, neuroimaging techniques combined with behavioral paradigms may be particularly useful in identifying factors conducive to dishonesty and selfish and fraudulent behavior, because it provides direct and unobstructed insights into the underlying neurocognitive mechanisms that give rise to these behaviors. When measuring brain activity while a person is cheating, there is no chance for the person to conceal the motives or thought processes that lead to the decision to cheat. The insights gained from neuroimaging research on (dis)honesty may thus help to mitigate such immoral behaviors and help to design interventions to prevent them in the first place.

1.9 Outline of the thesis

This thesis investigates the neurocognitive underpinnings of individual differences in fairness and dishonesty, by means of combining different neuroimaging methods with well-established as well as newly developed experimental paradigms. The empirical chapters are outlined in detail below. Chapters 2 to 5 will focus on (dis)honesty without identifiable victim, whereas Chapter 6 will focus on fairness with an identifiable victim.



Honesty

1. Why do we cheat?
- Underlying mechanisms
- Task and imaging simultaneously

2. Who cheats?
- Stable neural markers
- Scanning and task separate

fMRI (N= 40)

Chapter 2
- Role of cognitive control
- Trial-by-trial cheating
- Individual differences

Chapter 4
- Neural mechanisms underlying nonhabitual (dishonesty)

EEG (N = 40)

Chapter 3
- Replicate across modalities
- Localizer to reduce reverse inference
- Higher temporal resolution

rsfMRI (N= 99)

Chapter 5
- Link individual differences to the resting brain
- Predict cheating in the absence from temptation

Fairness

2. Why are we fair?
- Motivations for fairness
- Individual differences

fMRI (N= 32)

Chapter 6
- Decode selfish versus fairness motivation for brain patterns
- Role of cognitive control versus theory of mind

Figure 1.1. Overview of the chapters in the dissertation, including the method used, the sample sizes and the main contribution.

Chapter 2: Cognitive control increases honesty in cheaters but cheating in those who are honest

On a daily basis, we are faced with the conflict between the temptation to cheat for financial gains and maintaining a positive image of ourselves as being a 'good person'. While greed has been identified as determinant of cheating behavior, the processes that prevent us from engaging in unethical behavior, particularly in a context in which there is no directly identifiable victim, remain elusive. Moreover, although there are clearly individual differences in honesty, it remains an open question whether the same neurocognitive processes promote honesty for all of us or whether there are individual differences in how we remain honest when exposed to an opportunity to cheat. In this chapter we set out to answer these questions, by identifying the neural mechanism underlying decisions to cheat. We contribute to the existing literature by applying the Spot-The-Difference task, introduced above, which pioneers in measuring cheating inconspicuously on a trial-by-trial basis in the MRI scanner. This allowed us to not only measure the neurocognitive mechanisms underlying (dis)honest decisions, but to be the first to also more specifically identify which neural processes enable a generally honest person to cheat and a cheater to remain honest.

While replicating previous research showing that greed motivates dishonesty, we reveal that self-referential thinking processes drive honest behavior. Furthermore, we reveal that cognitive control does not serve the same purpose across individuals but facilitates honest decisions for cheaters, whereas it enables cheating for honest participants. Our findings thus contribute by reconciling the longstanding controversy between Will and Grace. The role of cognitive control is thus to override an individual's moral default. In addition, we are the first to provide neural evidence for the importance of self-referential thinking processes involved in maintaining a positive self-image and promoting honesty. These novel insights contribute to a deeper understanding of neural correlates of individual differences in moral decision making.

Using machine learning methods combined with cross-validation, we demonstrated that our models trained on neural activation and functional connectivity patterns generalize to unseen data, which highlights the generalizability and replicability of our findings and may contribute to society by making an important first step in the development of useful neuroimaging-based biomarkers of dishonesty with real-world applicability.

Chapter 3: Replicating the contingency of cognitive control on moral default

In Chapter 2 we investigated how the brain arbitrates between reward and self-concept. The study pioneered in demonstrating that brain regions associated with cognitive control helped dishonest participants to be honest, while they enabled cheating for more honest participants. Thus, we found that cognitive control does not promote (dis)honesty per se; it depends on one's moral default.

Despite the compelling evidence of this finding, it has become evident over the last decade that one study does not suffice as support for a novel hypothesis. The pivotal importance of conducting replications has been stressed as several authors have questioned the validity of numerous findings in neuroscience (e.g. Ioannidis, 2005; Button, et al., 2013; Botvinik-Nezer et al., 2020). Low replicability could have detrimental effects on the scientific community as future work is built on fragile ground. Moreover, most neuroimaging studies on dishonesty thus far have relied on reverse inference to interpret the neural mechanisms underlying (dis)honesty, where the cognitive operations are inferred from activation in an observed region of the brain. While often informative, this approach is deductively invalid and can be misleading.

To address these concerns, this chapter contributes by focusing on the replicability of our previous finding with a different neuroimaging method (EEG instead of fMRI), to capture different aspects of the underlying neural processes (higher temporal resolution) and provide convergent evidence across modalities. In addition, we employed the Stroop task as an external localizer task to identify the neural signature of cognitive control and reduce reverse inference problems.

By using a well-established localizer task for cognitive control in combination with EEG this study contributes by providing convergent validation across modalities of the differential effect of cognitive control on (dis)honesty. The results confirm the findings in Chapter 2. Our study thus contributes by more rigorously validating the neural mechanisms underlying dishonest decisions and reducing issues of reverse inference.

Chapter 4: Different neural mechanisms underlie non-habitual honesty and non-habitual cheating

As discussed in the beginning of this chapter, there is a long-standing debate with regard the neurocognitive processes underlying honesty: Is it a matter of willpower or does it flow naturally and gracefully? Proponents of the Will hypothesis argue that cognitive control is needed to suppress an individual’s temptation to cheat and allow honesty. In direct opposition, proponents of the Grace hypothesis argue that honesty is automatic without the need for active resistance to temptation. In Chapters 2 and 3, we provide reconciliation between these competing hypotheses, revealing that the cognitive control network orchestrates both, the honesty of individuals who

are generally dishonest and cheating in those who are generally honest, depending on the individual's moral default.
In a commentary on our work, Abe (2020) pointed out that while our findings provided insights into the role of cognitive control in overriding moral default, they did not reveal whether overriding honest default behavior is the same as overriding dishonest default behavior at the neural level. This is important to investigate because it may reveal whether the IFG has access to the moral valence (overriding the 'good' or 'bad' default) of a given decision. In this short chapter, we set out to answer these questions by applying multivariate pattern analysis to compare neural patterns associated with the nonhabitual decision to cheat (overriding the default for honesty) to patterns associated with nonhabitual honesty (overriding the default for cheating).

Chapter 5: Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest

As discussed in the beginning of this chapter, it is evident that some people are more successful than others in striking the balance between selfish reward and moral self-concept. However, because the underlying mechanisms of socially undesirable traits and behaviors are difficult to measure, it remains poorly understood whether there are stable markers of these individual differences in (dis)honesty.

In this chapter we investigated whether the brain's functional connectome at rest could be used to reliably distinguish cheaters from honest people even in the absence of any temptation to cheat. We combined resting state fMRI (rsfMRI) with an extended version of connectome-based predictive modelling (CPM) and the Spot-The-Difference Task. This approach allowed us to not only measure which task-independent functional patterns promote honesty but also to explore which functional connections contribute most to predicting (dis)honesty in unseen participants.

Our analyses revealed that functional connectivity between brain networks linked to self-referential thinking, cognitive control, and reward can reliably and accurately predict, out-of-sample, whether a participant is honest, or a cheater. Participants who cheated the most also scored highest on a diverse set of well-established self-report impulsivity measures, which highlights the ecological validity of our task. Importantly, when comparing neural and questionnaire measures, the neural measures were found to be substantially better at predicting honesty.

The current study expands on the previous chapters by demonstrating that the neural determinants of (dis)honesty can be detected in stable and reliable patterns of functional connectivity of the brain at rest in the absence of temptation, as opposed to context dependent task-evoked neural activations in our previous study. Further, we provide evidence for the ecological validity of our measure for dishonesty and show the superiority in predictive performance of our neural measure as compared to self-report measures.

Chapter 6: Decoding fairness motivations from multivariate brain activity patterns

Our social life abounds with situations in which there is a conflict between selfish urges and the welfare of others, and it is clear that there are substantial differences in how individuals resolve this conflict. Moreover, even when individuals act prosocially, they may do so out of selfish motivations. In this chapter, we set out to investigate whether we can decode selfish versus fairness motivations from neural patterns during decision making in economic games.

In the literature on social and economic decision making there has been a disconnect between neuroimaging research focusing on the role of cognitive control, represented in particular by neural activation in the dlPFC, and behavioral research emphasizing the importance of theory of mind processes, represented by areas such as the TPJ, in prosocial behavior. In our study we aimed at elucidating the contributions of brain networks related to cognitive control and theory of mind to individual differences in prosocial behavior. We collected fMRI data from participants while they played several rounds of the Ultimatum and Dictator Game and related their behavior during these interactions to patterns of neural activation during decision-making, using machine learning methods.

The results showed that there are indeed large individual differences in fairness and that we were able to decode selfish versus fairness motivations from the participants neural patterns while making offers in the two games. Contributing by bridging the disconnect in the literature, we thus conclude that heterogeneity in prosocial behavior may be driven by variations in the propensity to engage theory of mind and cognitive control processes during decision making. As such, our results contribute by providing a deeper understanding of the underlying neurocognitive mechanisms of human prosocial behavior.

Chapter 7: Discussion

In the discussion chapter (Chapter 7), I will discuss the three main contributions we can derive from the five empirical chapters. I will highlight the theoretical contribution of the empirical work to the literature and elaborate on the relevance of these insights for business and society. I will provide suggestions for possible intervention and provide examples of how they might be applied. In addition, limitations of my research are discussed, and I will point out promising future avenues for research moral decisions relating to fairness and honesty.

1.10 Declaration of contribution

In the following paragraphs, I declare my contribution to the empirical chapters (Chapters 2-6) of this dissertation and also acknowledge the contribution of other parties involved.

For Chapter 2, the author of the dissertation (SPHS), the promoter (AS) and the daily supervisor (MASB) collaboratively formulated the research question and designed the experiment. SPHS conducted a literature review and executed the data collection with the assistance of Mayca Thijssen. SPHS conducted the data analysis, interpreted the findings and wrote the manuscript with the input from AS and MASB. All authors have edited and approved the final manuscript. The chapter has been published in the *Proceedings of the National Academy of Sciences of the United States of America* (Speer, Smidts & Boksem, 2020).

For Chapter 3, The same contributions as in Chapter 2 apply. This chapter has been published in the *Journal of Neuroscience* (Speer, Smidts & Boksem, in press).

For Chapter 4, the same data as in Chapter 2 were used as it was written as a response to the commentary on Chapter 2. The same contributions as in Chapter 2 apply. This chapter has been published in *Frontiers in Neuroscience* (Speer, Smidts & Boksem, 2021).

For Chapter 5, SPHS, AS and MASB collaboratively formulated the research question and designed the experiment. The behavioral and neuroimaging data was collected by SPHS, but some also by Hang-Yee Chan (rsfMRI data study 3) and Nynke van der Laan and Christin Scholz at the University of Amsterdam (rsfMRI data study 2) as part of a collaboration arranged by AS. SPHS conducted the data analysis, interpreted the findings and wrote the manuscript with the input from AS and MASB. All authors have edited and approved the final manuscript. The chapter is currently under second-round review at *NeuroImage*.

For Chapter 6, MASB formulated the research question and designed the experiment and both MASB and SPHS collected the data. SPHS conducted the data analysis, interpreted the findings and wrote the manuscript with the input from MASB. All authors have edited and approved the final manuscript. The chapter is published in *Social, Cognitive & Affective Neuroscience* (Speer & Boksem, 2020).

SPHS wrote Chapter 1 and Chapter 7 and implemented feedback from AS and MASB.

Chapter 2

Cognitive control increases honesty in cheaters but cheating in those who are honest [1]

2.1 Abstract

Every day, we are faced with the conflict between the temptation to cheat for financial gains and maintaining a positive image of ourselves as being a ‘good person’. While it has been proposed that cognitive control is needed to mediate this conflict between reward and our moral self-image, the exact role of cognitive control in (dis)honesty remains elusive. Here, we identify this role, by investigating the neural mechanism underlying cheating. We developed a novel task which allows for inconspicuously measuring spontaneous cheating on a trial-by-trial basis in the MRI scanner. We found that activity in the Nucleus Accumbens promotes cheating, particularly for individuals who cheat a lot, while a network consisting of Posterior Cingulate Cortex, Temporoparietal Junction and Medial Prefrontal Cortex promotes honesty, particularly in individuals who are generally honest. Finally, activity in areas associated with Cognitive Control (Anterior Cingulate Cortex and Inferior Frontal Gyrus) helped dishonest participants to be honest, whereas it enabled cheating for honest participants. Thus, our results suggest

[1] This chapter has been published in PNAS (Speer, Smidts & Boksem., 2020)

that cognitive control is not needed to be honest or dishonest per se, but that it depends on an individual’s moral default.

2.2 Introduction

Imagine a friend sends you a link to a website where you can illegally stream recently released movies for free. Would you decide to stream the movie which you otherwise would have paid for? If so, how many movies would you stream? On a daily basis we are faced with the conflict between the temptation to violate moral standards to serve our self-interest and to uphold these moral standards, but how the brain resolves this conflict remains elusive.

When exposed to the opportunity to cheat, clearly the (financial) rewards play a crucial role: the higher the reward, the more attractive the decision to cheat (Becker 1968; Allingham and Sandmo 1972). As the renowned British novelist Jonathan Gash so eloquently stated: ‘Fraud is the daughter of greed’ (Gash, 1992). In line with this sentiment, behavioral research has demonstrated that greedy people find a variety of moral transgressions more acceptable and engage in such transgressions more often as compared to less greedy people (Seuntjens et al., 2019). Indeed, neural responses in anticipation of reward, reflected in activity in the nucleus accumbens (NAcc), predict cheating behavior in a subsequent task (Abe & Greene, 2014). Collectively, these findings emphasize that higher rewards and stronger sensitivity to reward increase the likelihood of dishonesty.

Accumulating evidence from psychology, economics and neuroscience has demonstrated, however, that people care about more than only maximizing their own monetary payoff, which is reflected in the high prevalence of prosocial behaviors such as altruism and reciprocity. People have internalized social norms and use these as an internal benchmark against which they compare their own behavior (Campbell 1964; Henrich et al. 2001). In the context of dishonesty, the way we view ourselves, our self-concept (Aronson 1969; Baumeister 1998; Bem 1972), may prevent us from cheating. People highly value honesty in others and also have strong beliefs in their own moral standards (Dhar & Wertenbroch, 2012). Violating one's own moral standards will require a negative update of one's self-concept which is highly aversive (Berthoz et al., 2006). As a consequence, people are motivated to uphold their self-concept even if it comes at the cost of foregoing potential monetary gains (Mazar, Amir, & Ariely, 2008). Hence, when given the opportunity to cheat, people are torn between the conflicting motivations to obtain desirable monetary gains versus the long-term goal of maintaining a positive self-image. Whereas the neural pattern associated with greed have been linked to cheating behavior (Abe & Green, 2014), the neural processes linked to maintaining a positive self-image are yet to be discovered. Based on the behavioral research presented here, we hypothesize that self-referential thinking processes linked to a network of regions including the medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC) and bilateral temporoparietal junctions (TPJs; Gusnard et al, 2001; Meffert et al., 2013; van Buuren et al., 2010) may underlie self-concept maintenance and promote honesty.

Several lines of research have proposed that cognitive control is needed to resolve the tension between reward and self-concept (Abe & Greene, 2014; Gino, Schweitzer, Mead, & Ariely, 2011; Greene & Paxton, 2009; Maréchal, Cohn, Ugazio, & Ruff, 2017; Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009). It has been found that this conflict is often settled with a compromise in which participants behave dishonestly enough to profit from the opportunity to cheat, but honestly enough to maintain a positive self-image (Mazar et al., 2008). While it is evident that cognitive control plays a crucial role in resolving this conflict, the precise nature of the role of cognitive control in moral decisions remains controversial (Köbis et al., 2019).Two competing theories have been proposed: the 'Will' and the 'Grace' hypothesis (Greene & Paxton, 2009).

The 'Will' hypothesis puts forward that people are per default selfish and dishonest and that in order to be honest, deliberate cognitive control needs to be exerted. Thus, honesty is a result of the effortful resistance of temptation, similar to the cognitive control processes that allow individuals to delay gratification (McClure, Laibson, Loewenstein, & Cohen, 2004). This hypothesis is supported by behavioral studies that have shown that participants who are cognitively depleted by demanding tasks, sleep deprived or under time pressure are more prone to dishonest behavior (Gino et al., 2011; Mead et al., 2009; Barnes, Schaubroeck, Huth, & Ghumman, 2011; Shalvi, Eldar, & Bereby-Meyer, 2012). Collectively, these studies suggest that people automatically serve their self-interest and require cognitive control to resist the temptation to cheat in order to maintain a positive self-image.

In contrast, the ‘Grace’ hypothesis proposes that people are intuitively honest and require cognitive control to override their dominant honest impulses to occasionally profit from an opportunity to cheat. The hypothesis that cheating rather than honesty is a complex cognitive function demanding cognitive effort, is supported by research showing that people react faster when asked to tell the truth as compared to lying (for meta-analyses, see Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; Verschuere, Köbis, Bereby-Meyer, Rand, & Shalvi, 2018) and are more honest under time pressure (Capraro, 2017; Capraro et al., 2019). Cheating requiring cognitive capacity is also supported by findings that people cheat less when taxed by a cognitively demanding memory task as compared to a less taxing task (van’t Veer, Stel, & Van Beest, 2014). In sum, these findings suggest that honesty is intuitive and cognitive control is required to override this default intuition in order to benefit from an opportunity to cheat.

In light of these evidently contradictory findings, this study aims at investigating how cognitive control resolves the conflict between external financial rewards and one’s self concept and more specifically, how this decision process unfolds in the brain. A better understanding of the function of cognitive control in the decision to cheat may help reconcile the controversy between the Will and Grace hypothesis.

In order to study how reward, self-concept and cognitive control influence cheating on a trial-by-trial basis, we developed an innovative task, based on a general paradigm proposed by Gai and Puntoni (2017), in which participants could cheat repeatedly, deliberately and voluntarily inside the MRI scanner without suspicion of the real purpose of the task. Specifically, the advantage of this task, which we call the Spot-The-Difference task, is that it allows for directly tracking on which trials the participants cheated, enabling us to study within subject variation in moral decisions and its neural underpinnings. Importantly, previous neuroimaging studies on cheating behavior have not been able to answer these questions as they used tasks such as the coin-flip task (Greene & Paxton, 2009; Abe & Greene, 2014), where cheating is inferred from the aggregate behavior at the end of the task thus eliminating the possibility to study trial-by-trial variation in behavior. Notably, participants believed that the experimenter did not know that they were cheating, which is critical as participants are found to cheat less if participants believe experimenters can observe their choices and know the truth (Gneezy, Kajackaite, & Sobel, 2018). The Spot-the-Difference paradigm is therefore the first behavioral paradigm to assess cheating behavior inconspicuously on a trial-by-trial basis enabling us to study individual differences in neurocognitive processes underlying cheating behavior while also being sensitive to within subject variation. This will enable us to not only identify neural mechanisms promoting honesty in general but also explore the neural processes that underlie an honest decision made by someone who generally cheats.

In our analysis, we first conducted an exploratory whole brain analysis to identify the brain networks underlying the decision to cheat or to be honest. We first identified the brain networks engaged when exposed to the opportunity to cheat and when making the decision to cheat or to be honest. To reduce the reverse inference problem (Poldrack, 2006), we then assessed the neural overlap between our results and meta-analytically derived maps associated with, respectively, reward, self-concept and cognitive control from Neurosynth (Yarkoni et al., 2011). Subsequently, we used the Regions Of Interest (ROIs) obtained from this conjunction analysis to conduct a trial-by-trial analysis to study the neural mechanisms underlying within-subject variation in cheating behavior and also to explore functional connectivity between the resulting networks of regions. To test the generalizability and replicability of our results, we then used cross-validation to explore whether we can use neural activation to predict unseen trials, and functional connectivity patterns to distinguish between cheaters and honest participants. Here, we use ‘cheater’ and ‘honest’ as shorthand to indicate individuals who cheated (or not) in our task, which may or may not generalize to stable traits of (dis)honesty.

We find a central role for the cognitive control network, but that its effects depend on a person’s moral default. For more honest participants, cognitive control is needed in order to cheat, whereas for cheaters cognitive control is required to be honest. We demonstrate the generalizability and replicability of our findings by means of significant out-of-sample prediction of cheated decisions and cheating individuals, based on neural activation levels and connectivity patterns.

2.3 Methods

2.3.1 Participants

The reported analyses are based on 40 participants (30 females; age 18-35 years; *M* = 23.7 *SD* = 3.2) recruited from an online community for university students, where students can sign up for experiments. An initial screening interview ensured that all participants were right-handed with normal or corrected to normal vision, spoke English fluently, were not on any psychoactive medication influencing cognitive function and had no record of neurological or psychiatric illness. The study was approved by the ERIM internal review board and was conducted according to the Declaration of Helsinki.

2.3.2 Task and stimuli

2.3.2.1 Spot-The-Difference task

In the Spot-The-Difference task, participants were presented with pairs of images and were told that there were always three differences between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were asked to find three differences between the images. Because reward (see below) was contingent on participants *reporting* that they had found all three differences, without having to point them out, this design encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were present in the images).

Participants were told that the purpose of the study was to investigate the underlying neural mechanisms of visual search for marketing purposes such as searching for a product in an assortment or information on a webpage. In order to increase credibility of this cover story a simple visual search task was added at the beginning of the experiment (see Appendix 2.1), which was also performed in the scanner while participants were undergoing localizer scans. Further, participants were instructed that the neurocognitive effect of motivation, elicited by monetary reward, on speed and accuracy of visual search was investigated. Although participants were told that there were three differences in all trials, in 25% of the trials there were only two differences and in 25% there was only one difference. All stimuli were standardized in size and were presented on a white background on a computer screen. The ratio of 50% - 50% (three differences vs less than three differences) was chosen based on the results of pilot studies that indicated this ratio to be optimal in reducing suspicion that the pairs did not always contain three differences.

Trials were further categorized into normal (50%), hard (25%) and very hard trials (25%), for which participants could receive 5cts, 20cts, and 40cts, respectively. All of the trials with three differences (the filler trials) were categorized as normal trials, whereas trials with less than three differences (the trials of interest) were randomly categorized as hard or very hard trials. Consequently, the reward was independent of the number of differences in the image pair for the trials of interest, which is important in order to be able to disentangle the effects of reward and cheating magnitude (the actual number of differences) on cheating behavior. The different levels of difficulty were added to reduce suspicion about the real purpose of task. It was assumed that if trials are labeled as hard or very hard it would be more credible to the participant that the image pair actually contained three differences, but they were just too hard to spot. In addition, levels of difficulty were introduced to eliminate possible demand effects: we wanted participants to cheat for monetary reward and not to prevent seeming incompetent, which may be associated with different underlying neural mechanisms and consequently confound the analysis.

To further reduce suspicion about the purpose of the study, approximately 10% of all trials were point-and-click trials. In these trials, participants had to click on the location in the images where they spotted the differences using a joy-stick. As a consequence, cheating was not possible on the point-and-click trials. Participants always knew prior to the start of a trial whether it was a point-and-click trial indicated by a screen requesting participants to click on the image. This ensured that participants would not refrain from cheating on all other trials, while still reducing the suspicion about the real purpose of the study. Participants were told that only 10% of trials were point-and-click trials because it would take too much time to point out the differences for every pair. Further, participants were instructed that excessive movement by manipulating the joystick would interfere with the brain signal. In sum, there were 144 regular trials (of which 72 cheatable trials) and 12 point-and-click trials. The maximum amount of money earned, in case a participant cheated on all cheatable trials was approximately 35 Euros, whereas in case a participant would not cheat at all he or she would earn approximately 7.50 Euros. After completion of the full study, participants were debriefed that the purpose of the study was to investigate the underlying neural mechanisms of (dis)honest decision-making. They were informed that the number of differences between pictures and level of reward were manipulated to encourage cheating. To be fair to all participants, they were all paid out the maximum amount, irrespective of their actual cheating behavior. In addition, participants received a flat fee of 10 Euros for participation in the scanning session.

Each trial started with a fixation cross which was presented for a variable amount of time between 1-3s (see Figure 2.1). Subsequently, the Level of Difficulty screen was presented for 2 seconds informing the participants about the level of difficulty of the upcoming trial. This screen also displayed how much money could be earned on that trial. As a result, participants were constantly aware of the potential gains of cheating. Next, an image pair was presented for 6s, a length determined by the behavioral pilots, and participants engaged in the visual search. Afterwards, the participants were asked whether they spotted all three differences (yes/no response). On this decision phase screen, again the potential reward for this trial was presented, in order to make the reward more salient and increase cheating behavior. After 3s, the response phase started in which participants' responses were recorded. In the decision phase and the response phase the current balance was also shown, which was done to demonstrate to the participants that if they stated that they had found the three differences, their current balance increased immediately. It was assumed that this direct noticeable effect of behavior on the increase of the current balance, would further motivate participants to cheat.

The decision phase and response phase were separated to isolate the decision from motor responses. This was important for the fMRI analysis as we wanted to isolate the neural mechanisms underlying decision making from possible neural confounds related to button presses. Besides that, the buttons corresponding to "Yes" and "No" were switched across trials to further reduce confounding effects and to reduce the response bias for the dominant hand. Once the participants responded, the choice was highlighted by a blue box for 500ms to indicate that the response was recorded and the trial ended. If no response was made, the trial ended after 3s. In addition, there were five practice trials, in which participants could get acquainted with the task. Stimulus presentation and behavioral data acquisition was performed using

Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

The main advantage of our experimental design is that it allowed tracking on which trials the participants cheated. As we knew how many differences there are in each image pair, we knew precisely whether the participants cheated or not. Further, by varying the number of differences, this design enabled us to assess the magnitude of cheating (i.e, cheating when only 1 vs 2 differences were found). It is therefore the first behavioral paradigm that allowed to assess cheating behavior inconspicuously on a trial-by-trial basis in the scanner.

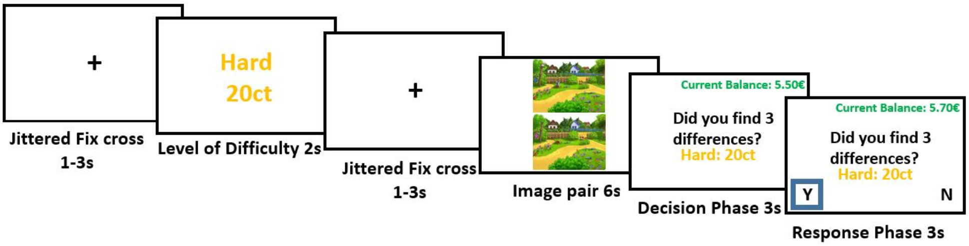


Figure 2.1. One trial of the Spot-The-Differences paradigm. Participants view a screen indicating the difficulty and value of the trial, then the image pair appears for six seconds and then participants have to indicate whether or not they spotted all three differences.

2.3.2.2 Stimuli

Stimuli for the task consisted of 144 Spot-The-Difference image pairs that were downloaded from the Internet. Cartoon images of landscapes containing several objects were selected, to make them engaging and challenging enough for the participants. Landscapes were chosen as they generally satisfied the necessary criterion of containing several different objects. The stimuli consist of pairs of images that are identical apart from a certain number (1-3) of differences that were created using Adobe Photoshop. Differences consisted of objects added to or removed from the landscape picture or changed colors of objects. Differences were fully randomized across all pairs of images, which means that all image pairs could be presented with either one, two or three differences (for example images see Appendix 2.10). To make sure that participants would be able to find the differences between the images in a reasonable amount of time and to minimize the chance of participants believing that they had seen a difference when they had not (false positives), we ran a pilot study on Amazon's Mechanical Turk (N=205) to test the difficulty to spot the differences between the images and to determine the optimal duration of picture presentation (see Appendix 2.2).

2.3.3 Experimental procedure

Before the experiment started, participants were introduced to the cover story, the tasks and the scanner environment and they signed the informed consent form. They were then informed about and checked on the safety requirements for MRI scanning and completed practice trials for both visual search tasks outside of the scanner. Subsequently they were guided into the scanner and completed the simple visual search task (5 min) followed by the Spot-The-Difference task which took approximately 45 minutes. Scans for the Spot-The-Difference task were acquired in six separate runs to allow participants short periods of rest in order to reduce head movements during the scans and to ensure that participants were attentive throughout the task. After completing the two tasks in the scanner, participants were taken to a separate room in absence of the experimenter and filled-out a short questionnaire including questions about their thoughts on the purpose of the task.

2.3.4 FMRI acquisition

The functional magnetic resonance images were collected using a 3T Siemens Verio MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in descending interleaved order (3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 75°). TE was 30ms and TR was 2030ms. A T1-weighted image was acquired for anatomical reference (1.0 × 0.5 × 0.5

mm resolution, 192 sagittal slices, flip angle = 9°, TE = 2.26ms, TR = 1900ms).

2.3.5 fMRI analysis

2.3.5.1 Preprocessing

The fMRI data was preprocessed using fMRIPrep version 1.0.8, a Nipype based tool (Gorgolewski et al., 2011). The reason for choosing fMRIPrep was that it addresses the challenge of robust and reproducible preprocessing as it automatically adapts a best-in-breed workflow to virtually any dataset, enabling high quality preprocessing without the need of manual intervention (Esteban et al., 2019). Each T1w volume was corrected for intensity non-uniformity (INU) and skull-stripped. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Esteban et al., 2016) was performed through nonlinear registration, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w. Fieldmap distortion correction was performed by co-registering the functional image to the same-subject T1w image with intensity inverted (Caballero-Gaudes & Reynolds, 2017) constrained with an average fieldmap template (Tustison et al., 2010). This was followed by co-registration to the corresponding T1w using boundary-based registration (Smith, 2002) with 9 degrees of freedom. Motion correcting transformations, field distortion correcting warp, BOLD-to-T1w transformation and T1wto-template (MNI) warp were concatenated and applied in a single step using Lanczos interpolation. Physiological noise regressors were extracted applying CompCor (Cox & Hyde, 1997).

Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Frame-wise displacement (Treiber et al., 2016) was calculated for each functional run using the implementation of Nipype. For more details of the pipeline see https://fmriprep.readthedocs.io/en/latest/workflows.html.

2.3.5.2 Statistical analyses

For each participant we estimated a general linear model (GLM) using regressors for onsets of the decision phase for cheated trials, honest trials, cheatable trials (trials with less than three differences) and non-cheatable trials (trials with three differences). The duration of the epoch for the decision phase was three seconds and the beginning of the decision phase was used as onset times. The decision phase was used as it provides all the necessary information to make the decision and is free of brain activity related to motor responses. In addition, regressors were added for the onsets of the Level of difficulty phase with a separate regressor for each level of reward. For the level of difficulty phase the duration was two seconds. This phase was used to test whether participants are indeed sensitive to differences in potential gains, as it provided information about the possible reward without any moral conflict. Besides that, in order to ensure that there were no significant differences in engagement or motivation in the Spot-The-Difference task between conditions or subjects, regressors were added for the onsets of the visual search phase in which the image pairs were presented on the screen. The duration of the visual search phase was six seconds (see *Figure 2.1*). Lastly, regressors for the button presses were added. Average background, white matter and cerebrospinal fluid (CSF) signal, framewise displacement, six head motion regressors and six aCompCor regressors, all obtained from fMRIprep, were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. A smoothing kernel of 5 mm (FWHW) was applied. Linear contrasts were computed between honest and cheating decisions and between cheatable and non-cheatable trials. These contrasts were then subjected to a random effects analysis to compute main effects (one sample t-test), and to regression analyses with behavioral data (i.e., total amount of cheating for each participant) as regressors.

2.3.5.3 Cheatable vs. non-cheatable trials

To identify the neural correlates associated with the opportunity to cheat, we contrasted the neural activation during cheatable trials (trials with less than three differences), against activation in non-cheatable trials (trials with three differences) in both directions. Subsequently, using the contrast images obtained for each subject, one sample t-tests were conducted on the group level to explore the average effect of being exposed to the opportunity to cheat across participants. We also added the cheat count, which is a measure how often each participant cheated in total on the Spot-The-Difference task, as a group level covariate to explore whether there are individual differences in the neural mechanisms when exposed to the opportunity to cheat, between individuals who cheat a lot vs. those who rarely cheat. The threshold applied to the group level statistical maps was a voxel-wise false discovery rate of $p < 0.05$ (FDR) to correct for multiple comparisons. Clusters of activation resulting from the thresholding were characterized in terms of their peak voxels in the MNI coordinate space.

2.3.5.4 Honest decisions vs. cheating

To explore the neural mechanisms underlying the decision to cheat, we contrasted neural activation in the decision phase on trials on which participants cheated against trials in which they did not, in both directions. For each of these contrasts we then conducted one sample t-tests on the group level to explore the average effects of each of these contrasts across participants. In addition, we also entered the total cheat count for each participant as covariate on the group level to investigate the correlation between behavior and neural activation in the contrasts of interest. Based on the resulting beta images, second-level random-effects group contrast maps were then created in both directions (i.e., positive and negative correlation between activation and cheat count). The threshold applied to the group level statistical maps was a voxel-wise false discovery rate of $p < 0.05$ (FDR) to correct for multiple comparisons. Clusters of activation resulting from the thresholding were characterized in terms of their peak voxels in the MNI coordinate space. Due to the fact that participants engaged in spontaneous, voluntary and deliberate cheating, the proportion of cheated and honest trials was not balanced for most of the participants. To account for possible confounding statistical effects of this imbalance, we under sampled the majority class for each participant to create a perfect balance when estimating the contrasts (Liu, Wu, & Zhou, 2008).

2.3.5.5 Single trial activation estimation

An important contribution of our task is that it allows us to assess cheating behavior on a trial-by trial-basis. That is, we are able to assess why a person who is generally honest, decides to cheat on some trials, and why a cheater might refrain from cheating on some occasions. To explore which neural mechanisms underlie this within subject variability, we extracted the neural activation from the ROIs identified in the analyses described above during decision making for each trial for each subject. These trial-by-trial activations could then be fed into multilevel models to explore which neural mechanisms may explain within subject variability.

To obtain single trial neural activations for the trial-by-trial multilevel models, individual time series were modeled using a double γ hemodynamic response function in a single trial GLM design using FSL's FEAT. Specifically, one GLM fitted a hemodynamic response function (HRF) for each trial, following the Least-Squares all (LSA) approach (Mumford, Turner, Ashby, & Poldrack, 2012), using the decision phase and level of difficulty phase of each trial, resulting in parameter estimates of sustained activations for each trial for each participant. The resulting β-values were converted to t-values (Misaki, Kim, Bandettini, & Kriegeskorte, 2010), resulting in a whole-brain map of t-values for each trial. The duration of the epoch for the decision phase was 3 seconds and 2 seconds for the level of difficulty phase. As for the previous analyses, average background, white matter and CSF signal, framewise displacement, six head motion regressors and six aCompCor regressors, all obtained from fMRIprep, were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. Multilevel modelling was conducted with custom R scripts in combination with the ‘lme4’ package for linear mixed-effects models (Bates et al., 2015) and the ‘glmmlasso’ package for variable selection for generalized linear mixed models by L1-penalized estimation (Groll & Tutz, 2014). fMRI analyses were conducted using custom Python scripts, which will be made publicly available.

2.3.5.6 Beta-series correlations

In order to further explore how the different areas resulting from the different contrasts described above interact with each other during decisions to cheat, we investigated the functional connectivity between these areas during the decision phase of the Spot-The-Difference Task. To avoid the problem of activation-induced correlations we implemented beta-series correlations (Rissman, Gazzaley, & D'Esposito, 2004). We used the single trial activations obtained as explained above by fitting a model that includes a separate regressor for each trial. We then correlated the parameter estimates from these regressors (the “beta series”) for honest decisions and cheated decisions separately between all the regions found to be significantly related to our contrast of interest, in order to examine the degree to which they show similar trial-by-trial activations, as is expected when these regions were functionally connected. The beta-series model is particularly useful in event-related fMRI studies where the spacing between trials is relatively long (more than 8-10 seconds), which is the case in our paradigm (Poldrack, Mumford, & Nichols, 2011). After obtaining the correlation matrix for each of the participants for honest and cheated decisions, we then also correlated the functional connectivity between each of the regions with the cheat count (individual differences in total cheating) in order to examine how functional connectivity differed for cheaters and more honest participants. To compare functional connectivity between honest and cheated decisions, correlations were transformed to z-values using the Fisher r-to-z transformation. Significance was estimated by means of permutation testing where the cheat count was randomly shuffled at each iteration (N=5000). The resulting empirical p-values were then corrected for multiple comparisons at FDR < 0.05.

2.4 Results

2.4.1 Behavioral results

Forty participants completed the Spot-The Difference task (for detailed description see Methods) inside the MRI scanner. In the Spot-The-Difference task, participants were presented with pairs of images and were told that there were always three differences between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain only one, two, or three differences. Participants were asked to find three differences between the images. Because reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were actually present in the images). Participants were told that the purpose of the study was to investigate the underlying neural mechanisms of visual search for marketing purposes such as searching for a product in an assortment or information on a webpage.

Large individual differences in the total amount of cheating were observed (Mean= 26%, Median=14%, SD=26%; see Figure 2.2): some participants cheated only on one or two trials (17.5% of participants), whereas others only missed one or two opportunities to cheat (5 %). Participants who cheated relatively often in the Spot-The-Difference task are from now on referred to as cheaters, whereas participants who were more honest are referred to as honest individuals. To assess suspicion about the real purpose of the study, participants were asked what the goal of the experiment was. Participants mentioned marketing research, consumer decision-making and visual search as our general cover story suggested that visual search is important for quickly locating one’s favorite brand or product in a supermarket. Importantly, none of the participants mentioned dishonesty, moral decision making or related concepts, which indicates that none of the participants were suspicious of the real goal of the study.

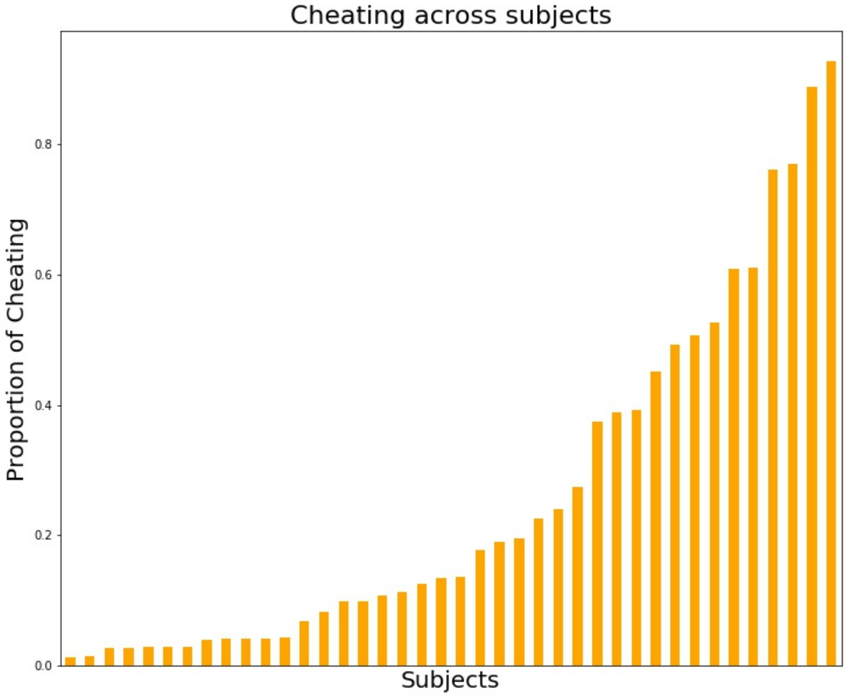


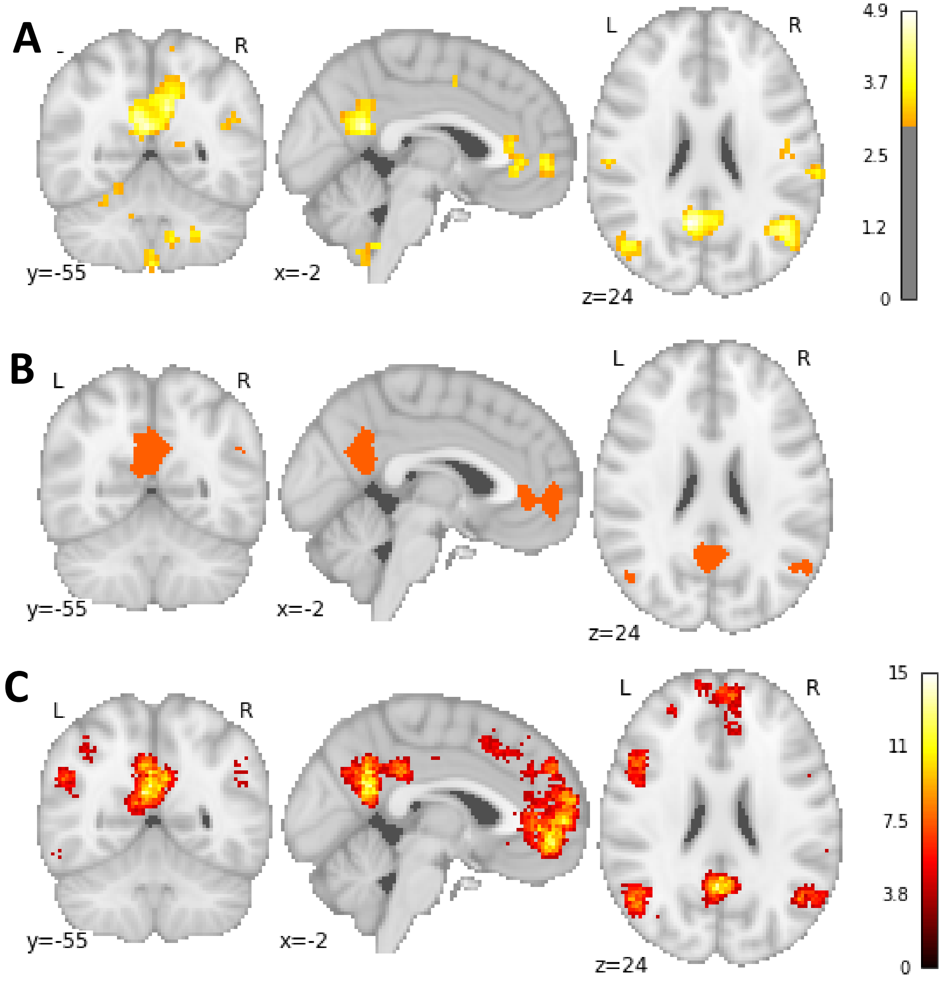
Figure 2.2. Individual differences in proportion of cheating (0-1) on the Spot-The-Difference task. N = 40.

We also explored how the task characteristics of the Spot-The-Difference task influenced cheating behavior. Given the nested structure of our data (trials within different number of differences and rewards within participants), we conducted a multilevel analysis for our behavioral data. This analysis was conducted for the cheatable trials only, so all trials with three differences between the images were removed. The dependent variable was the binary response (cheating vs. honest) with a logit link (cheating = 1, honest = 0). The number of differences and level of reward served as trial level predictors. The model allowed for random intercepts and random slopes within participants. This analysis revealed a significant effect of the number of differences (excluding three differences trials) on cheating behavior (*b*=2.13, *SE*=0.44, *z*=4.85, *p*<0.001). This shows that participants cheated more when the crime is smaller (that is, they indicated to have found three differences more often when there were two differences as compared to when there was only one). Specifically, when there were two differences, participants cheated on 36% of the trials, whereas participants only cheated on 16% of trials with only one difference (t=3.28, p=0.002). No effect of reward magnitude (see Methods) on cheating behavior was observed, and no significant interaction effects between number of differences and reward were found. We also tested for possible fatigue or habituation effects by using trial number as a trial level predictor to see whether cheating behavior increased or decreased over the course of the experiment. No effects of time were observed.

2.4.2 Neural mechanisms associated with the opportunity to cheat

As a first step of our fMRI analysis we explored the neural activation in response to the opportunity to cheat. In order to do so, we contrasted neural activity on trials in which participants had the opportunity to cheat against trials in which they did not have this opportunity (see Methods for details). To explore whether there are individual differences in the neural response to this opportunity, participants' cheat count was added as a group level covariate. The whole brain analysis revealed that more honest participants (compared to those who cheated more) exhibited greater activation in the posterior cingulate cortex (PCC), the medial prefrontal cortex (MPFC) and the bilateral temporoparietal junction (TPJ) when exposed to the opportunity to cheat (p_{FDR}<0.05; see Figure 2.3A and Appendix 2.3 for table with clusters).

As the activated network in our group-level results highly resembled the self-referential thinking network, we conducted a conjunction analysis with a meta-analytically derived self-referential thinking mask obtained from Neurosynth, FDR corrected for multiple comparisons at p<0.01 (using the term “self referential”; Wager, Nichols, Van Essen, Poldrack, & Yarkoni, 2011, See Figure 2.3 B and Appendix 2.4) to test whether there is indeed neural overlap. Neural overlap was found in the PCC (overlap (mm^3) = 4600), in the MPFC (overlap (mm^3) = 4072), in the right TPJ (overlap (mm^3) = 869) and the left TPJ (overlap (mm^3) = 608), see Figure 2.3C.


A
R
y=-55
x=-2
L
R
z=24
4.9
3.7
2.5
1.2
0
B
L
R
y=-55
x=-2
L
R
z=24
C
L
R
y=-55
x=-2
L
R
z=24
15
11
7.5
3.8
0

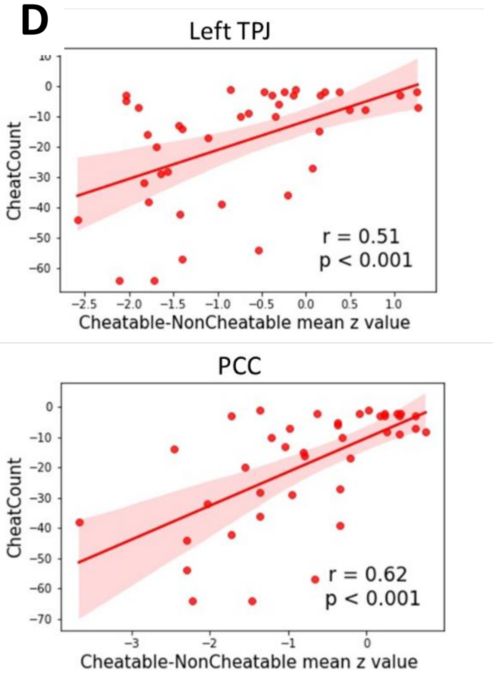
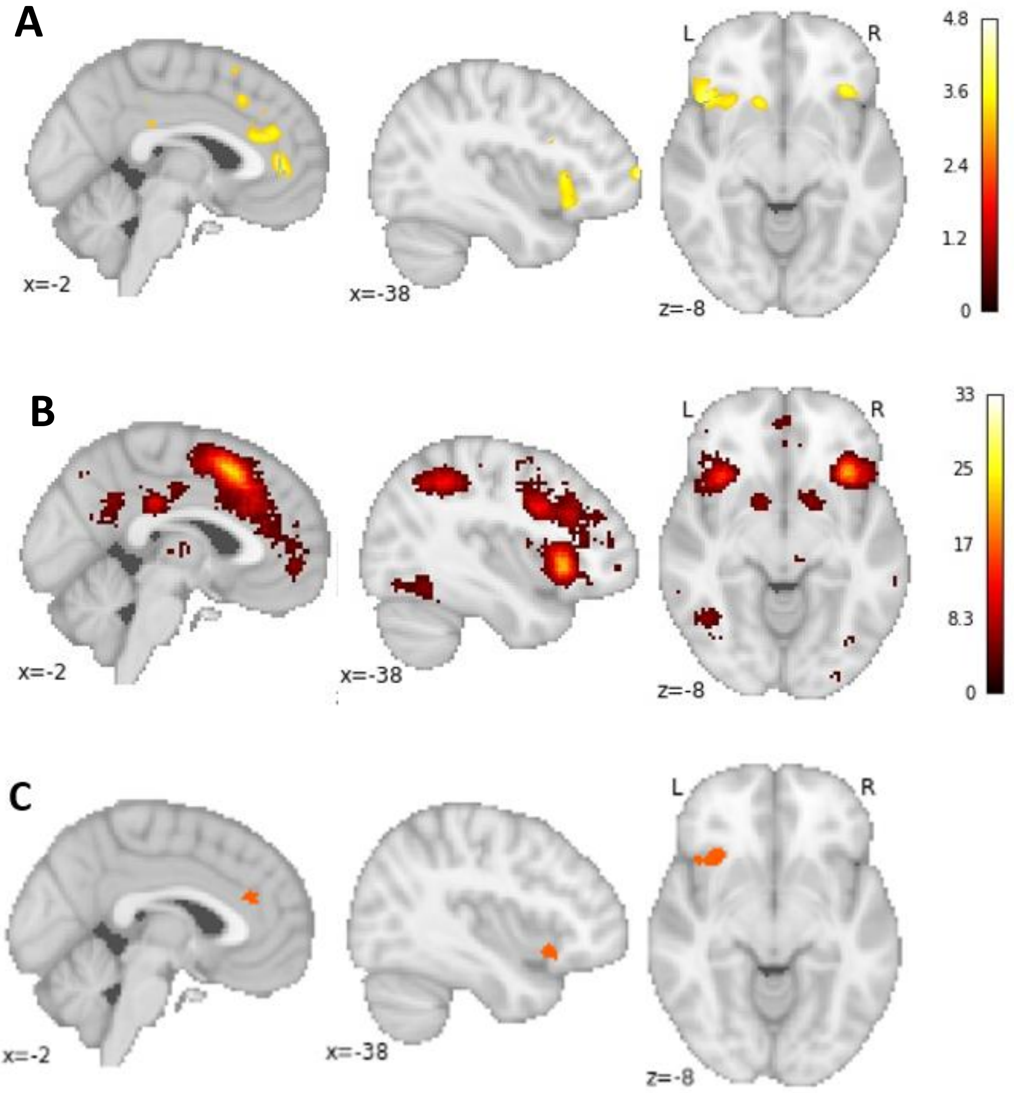


Figure 2.3. Honest participants engage the self-referential thinking network more strongly than cheaters when exposed to the opportunity to cheat. (A) More honest participants exhibit greater activation in the PCC, MPFC and bilateral TPJs when exposed to the opportunity to cheat. (B) Self-referential thinking mask obtained from Neurosynth. (C) Neural overlap between group level results for cheatable vs. non-cheatable trials correlated with cheat count and the self-referential thinking mask obtained from Neurosynth (D) The correlation between the level of honesty (reversed cheat count) and neural activation when participants were exposed to the opportunity to cheat as contrasted to no opportunity trials, for the PCC, bilateral TPJs and the MPFC, respectively (using ROIs obtained from the conjunction analysis).

2.4.3 Neural mechanisms underlying the decision to cheat

Next, we explored which neural mechanisms underlie the decision to cheat or not, when given the opportunity. To answer this question, we contrasted the neural activation of trials where participants had the opportunity to cheat but decided to be honest, against trials on which participants decided to cheat. As before, to explore whether there are individual differences in the neural processes underlying honest as compared to dishonest decisions, participants' cheat count was added as a group level covariate.

We found that participants who cheated more, showed higher activity in the anterior cingulate cortex (ACC) and the inferior frontal gyrus (IFG) when they made the decision to be honest ($p<.001$, uncorrected; see Figure 2.4 A, and Appendix 2.5 for table with clusters). Stated differently, cheaters engage their ACC and IFG more than honest participants when refraining from cheating. As the activated network in our group-level results highly resembled regions within the cognitive control network, we conducted a conjunction analysis with a meta-analytically derived cognitive control mask obtained from Neurosynth (Wager, Nichols, Van Essen, Poldrack, & Yarkoni, 2011, See Figure 2.4 B and Appendix 2.4) to test whether there is indeed neural overlap. Neural overlap was found in the ACC (overlap (mm^3) = 168) and in the left IFG (overlap (mm^3) = 1256), see Figure 3C.


A
x=-2
x=-38
L
R
z=-8
4.8
3.6
2.4
1.2
0
B
x=-2
x=-38
L
R
z=-8
33
25
17
8.3
0
C
x=-2
x=-38
L
R
z=-8

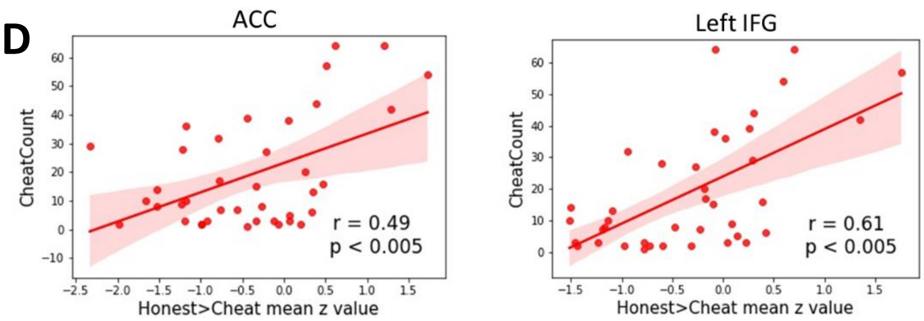


Figure 2.4. Cheaters exhibit higher activation in the ACC and left IFG when deciding to be honest. (A) A whole-brain analysis revealed that participants who cheat more, exhibit higher activation in the ACC and left IFG when deciding to be honest. (B) Cognitive Control network derived from Neurosynth.(C) Neural overlap between group level results for honest vs. cheated trials correlated with cheat count and the cognitive control mask obtained from Neurosynth in the Left IFG and ACC. (D) The correlation between cheat count and neural activation when participants decided to be honest as contrasted to a decision to cheat, for the left IFG and the ACC (using ROIs from the Conjunction Analysis).

2.4.4 Neural correlates of the sensitivity to reward are associated with cheating

2.4.4.1 Level-of-difficulty phase

Although we did not find any effects of reward on cheating on the behavioral level, we did want to test whether the participants responded to the reward on the neural level, as previous research has alluded to the relevance of reward anticipation in explaining individual differences in cheating (Abe & Greene, 2014, Seuntjes et al., 2019). Here, we investigated whether participants were motivated by the possible rewards that could be obtained on each trial and whether participants differentiated between the different magnitudes, 5ct, 20ct and 40ct, of reward on the neural level. We conducted a parametric modulation analysis where we used the onsets of the level of difficulty phase of each trial and added the magnitude of reward at each trial as a parametric modulator on the first level. The analysis revealed that the magnitude of reward modulated the activity in the bilateral Nacc significantly ($p_{FDR}<0.05$; see Figure 2.5 A and Appendix 2.6 for table with clusters).

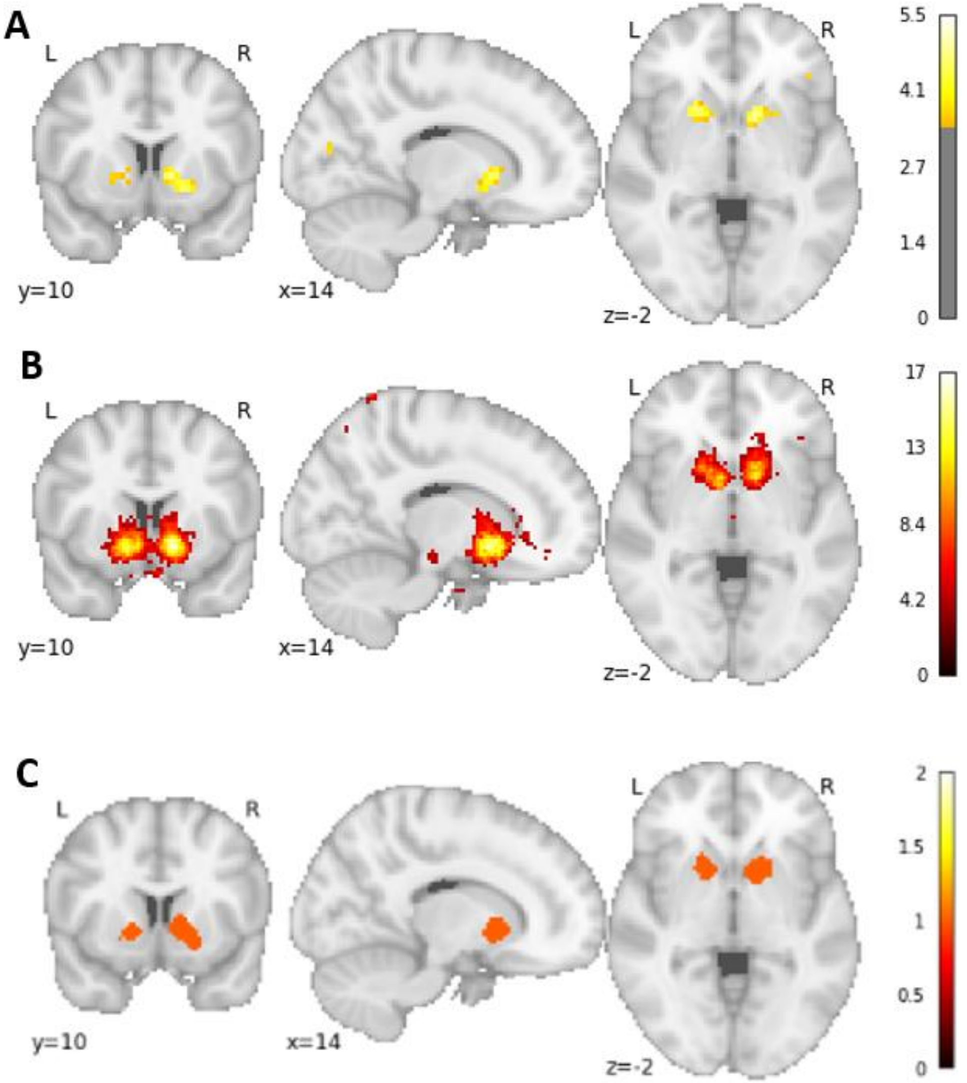
As the Nacc is well known for its role in processing the anticipation of reward (Ballard & Knutson, 2009; Knutson, Adams, Fong, & Hommer, 2001; Oldham et al., 2018) this suggests that the participants were indeed motivated by the potential rewards presented at the beginning of the trial. Further, differences between levels of reward magnitude were reflected in different levels of activity in the Nacc, suggesting that participants were indeed differentiating between the different reward magnitudes.

As the activated network in our second level results highly resembled the reward anticipation network, we conducted a conjunction analysis with a meta-analytically derived reward anticipation mask obtained from Neurosynth with false discovery rate (FDR) corrected for multiple comparisons at $p<0.01$ (Wager, Nichols, Van Essen, Poldrack, & Yarkoni, 2011, See Figure 2.5 B and Appendix 2.4) to test whether there is indeed neural overlap. Neural overlap was found in the right Nacc (overlap (mm^3) = 2040) and left Nacc (overlap (mm^3) = 840), see Figure 2.5 C. We also conducted an additional second level analysis, in which we added the cheat count as a covariate, in order to explore whether reward sensitivity in the Level-of-Difficulty phase differed between subjects. However, no significant differences were observed, indicating that participants were equally sensitive to the rewards, independent of how often they cheated.

2.4.4.2 Decision phase

To explore how the effect of reward anticipation, as represented by activity in the Nacc, on cheating differs for cheaters and more honest participants, we then used the ROIs derived from the conjunction analysis between our parametric modulation analysis and the Neurosynth map for reward (see Figure 2.5 C) and regressed mean Nacc activity per subject during the anticipation and decision phase against the cheat count. This analysis revealed that average Nacc activity significantly predicted cheat count

($b=18.29$, $SE=7.01$, $p<0.05$, see Figure 2.5 D) during the decision phase, whereas no significant effect was found during the level of difficulty phase ($b=-8.89$, $SE=14.2$, $p=0.54$). This suggests that participants are equally sensitive to reward during the level of difficulty phase when there is no moral conflict, however, when making the decision to cheat (or be honest), participants who cheat more seem to be driven more strongly by anticipation of reward.



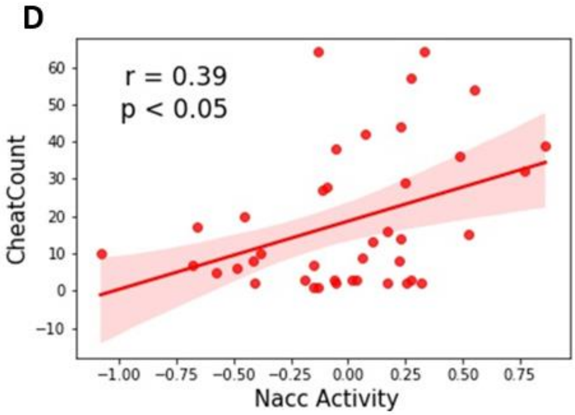


Figure 2.5. Cheaters exhibit higher activity in the Nacc when making (dis)honest decisions. (A) The left and right nucleus accumbens are parametrically modulated by the magnitude of reward. (B) Reward network derived from Neurosynth. (C) Neural Overlap between the parametric modulation analysis of the magnitude of reward and the reward anticipation network derived from Neurosynth. (D) Mean Nacc activity during the decision phase predicts cheat count.

2.4.5 Investigating within-subject variation in cheating: Trial-by trial analysis

In order to further explore how self-concept**,** reward and cognitive control influence decisions to cheat, we conducted a trial-by-trial analysis, which allowed us to investigate the neural mechanisms that determine why the same person may cheat on some occasions and remain honest on others. As a first step, we extracted average trial-by-trial activation from individual regions within the reward, cognitive control and self-referential thinking network, respectively, where we used the conjunction between our second level results and the Neurosynth maps (see Figures 2.3 C, 2.4 C & 2.5 C), resulting in one data matrix where the rows represent trials and the columns represent the regions of interest. Given the nested structure of our data (trials within different number of differences and rewards within participants) we then conducted a multilevel analysis for each of the networks (self-referential thinking, cognitive control and reward). The dependent variable was the binary response with a logit link (cheating = 1, honest = 0). The averaged activity within the obtained regions of interest served as trial level predictors, whereas the cheat count served as a subject level predictor. The models allowed for random intercepts and random slopes within participants.

2.4.5.1 Assessing the relative importance of the networks

To investigate which of the networks is most important in predicting cheating on the trial level, we performed variable selection for generalized linear mixed models by means of L1-penalized estimation. This was implemented using the 'glmmlasso' package in R, which implements a gradient ascent that allows to maximize the penalized log-likelihood, yielding models with reduced complexity (Groll & Tutz, 2014). The lasso regression adds a penalty term to the equation which shrinks coefficients in the model to zero and thus reduces complexity of the model and multicollinearity of predictors (Tibshirani, 1996). In this way it also selects the most important predictors in the model. This analysis revealed that the ACC (b = 0.13, SE=0.06, p=0.02), the left IFG (b = 0.42, SE=0.06, p<0.001), the cheat count (b = 1.59, SE=0.07, p<0.001) and the interaction effect between the left IFG and the cheat count (b = -0.38, SE=0.06, p<0.001), were most important in predicting cheating. These results suggest that the cognitive control network is most important in predicting cheating on the trial level. Inspecting the plot of the interaction effect (see Figure 2.6), we see that for participants who cheat a lot (light blue lines), higher levels in the left IFG are associated with lower probabilities of cheating, whereas for more honest participants (dark blue lines), higher activity in the left IFG is associated with higher probability of cheating. These findings suggest that the effect of the left IFG on cheating depends on whether a participant has the general tendency to cheat or to be honest.

Table 2.1. Multilevel Logistic Regression Model using the Cognitive Control Network to predict Cheating

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.582	0.069	-22.793	<0.001
ACC	0.132	0.06	2.306	0.02
L IFG	0.422	0.061	6.908	<0.001
Cheat count	1.60	0.07	23.735	<0.001
L IFG x Cheatcount	-0.382	0.062	-6.192	<0.001

* The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. ACC= Anterior Cingulate cortex, IFG= inferior frontal gyrus, R= right; L = left;

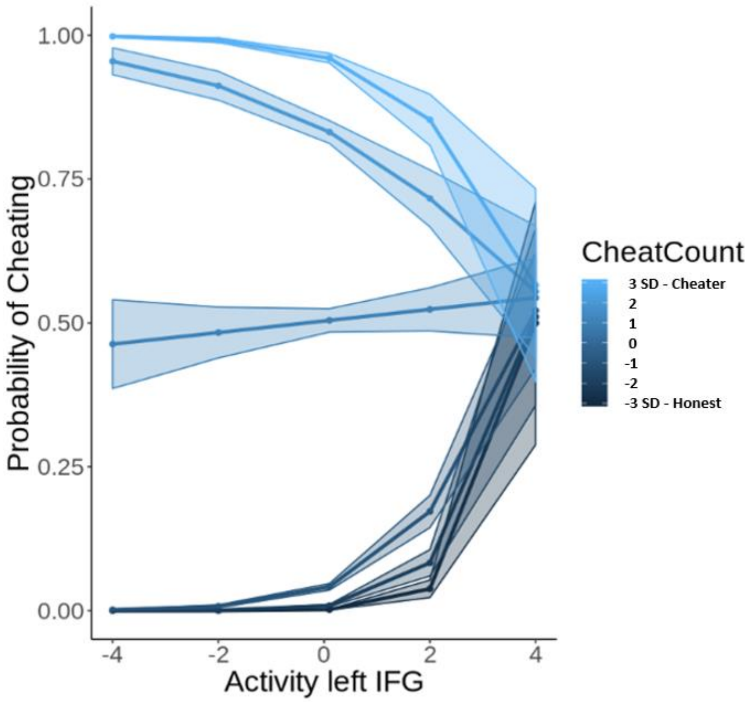


Figure 2.6. Interaction effect between cheat count and the left IFG in predicting the probability of cheating. The lines that are shown are the fitted values for participants 3SD (lightest blue), 2 SD (light blue) and 1 SD (blue) above the mean of the cheat count and participants 1 SD (dark blue), 2SD (darker blue) and 3 SD (black) below the mean of the cheat count.

2.4.5.2 Testing the predictive accuracy of the model

As the cognitive control regions were found to be most predictive of cheating, we used these predictors to test the prediction accuracy of our model. In order to do this, we used the trial level activation in the ACC and left IFG, excluding the cheat count, obtained from the conjunction analysis and trained a multilevel logistic regression model, with random slopes and intercepts, on a training set (70% of the data). Subsequently, we tested the model on the left out 30% of the data. As the dependent variable, cheating, was imbalanced, we used two accuracy metrics that are insensitive to the class imbalance, namely the area under the curve (AUC) and the F1-score, which is the harmonic mean of the precision and recall. Statistical significance was estimated using permutation tests where the dependent variable (cheating) was permuted 5000 times and the classification metrics were estimated based on random permutations. We found that we were able to significantly predict cheating based on unseen data from the cognitive control network (AUC=76%, F1=89%, $p<0.001$).

2.4.6 Individual differences in functional connectivity during decision making

2.4.6.1 Connectivity within the self-referential thinking network

In order to further explore how the different areas resulting from the different contrasts described above interact with each other during decisions to cheat, we investigated the functional connectivity between these areas during the decision phase of the Spot-The-Difference Task. To avoid the problem of activation-induced correlations we implemented beta-series correlations (Rissman, Gazzaley, & D'Esposito, 2004, see Methods for details). The beta-series correlation analysis revealed that functional connectivity within nodes of the self-referential thinking network were more strongly connected for honest participants than for cheaters when making honest decisions. Specifically, correlations between honesty and functional connectivity were found between the PCC and left TPJ ($r=0.51$, $\mathrm{p}_{\mathrm{adj}}<0.05$) and between PCC and MPFC ($r=0.55$, $\mathrm{p}_{\mathrm{adj}}<0.05$; see Figure 2.7). No significant correlations between honesty and functional connectivity were found for cheated decisions. In addition, the correlation between honesty and functional connections between PCC and left TPJ and between PCC and MPFC during honest decisions were significantly different from the correlation during cheated decisions (both comparisons $z > 2$, $p_{adj}<0.005$). Thus, the nodes within the self-referential thinking network, particularly between MPFC, left TPJ and PCC, seem to be more intimately connected to promote honesty particularly for honest participants, whereas when the connectivity between these nodes breaks down, honest participants tend to cheat.

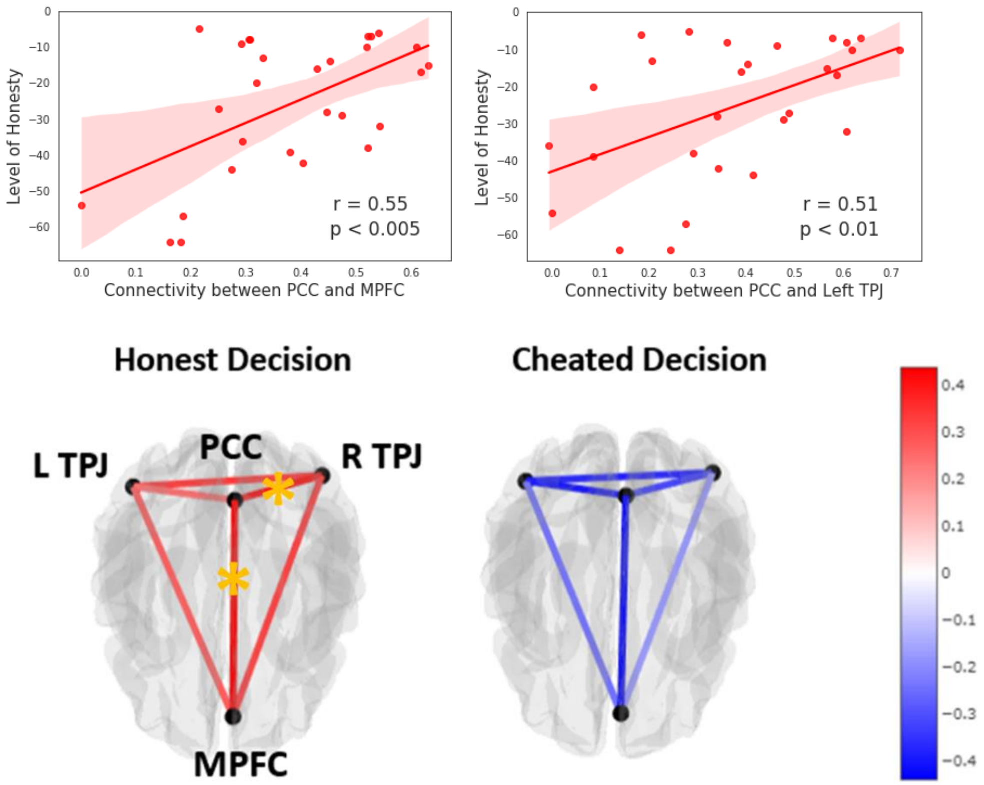


Figure 2.7. Top row: Correlation between level of honesty (reverse cheat count) and functional connectivity between PCC and MPFC and PCC and left TPJ. *Bottom row:* Connectome showing the correlation between level of honesty and the functional connectivity within the self-referential thinking network during honest decisions (left) and cheated decisions (right). The red lines represent a positive correlation between regions, whereas the blue lines represent a negative correlation. Darker color represents stronger correlation. Orange stars indicate a significant correlation after correction for multiple comparisons (FDR at p<0.05).

2.4.6.2 Classification of cheaters versus honest participants based on functional connectivity patterns

To test whether there is sufficient information in the connectivity patterns within the self-referential thinking network reported above to predict individual differences in honesty, a support vector classifier (Cox & Savoy, 2003; Mitchell, 2004) with linear kernel (C=1) was trained on the functional connectivity patterns of each participant to determine whether a participant was a cheater or an honest participant (categorized by median split). In order to avoid overfitting and inflated prediction accuracy (Vul, Harris, Winkielman, & Pashler, 2009) this was done using 8-fold cross validation (see Figure 2.8). Significance was estimated using permutation testing (N=5000). The classification analysis revealed that we could significantly classify an unseen participant as a cheater or an honest individual based on the connectivity patterns within the self-referential thinking network (Accuracy=71%, F1=75%, p<0.05).

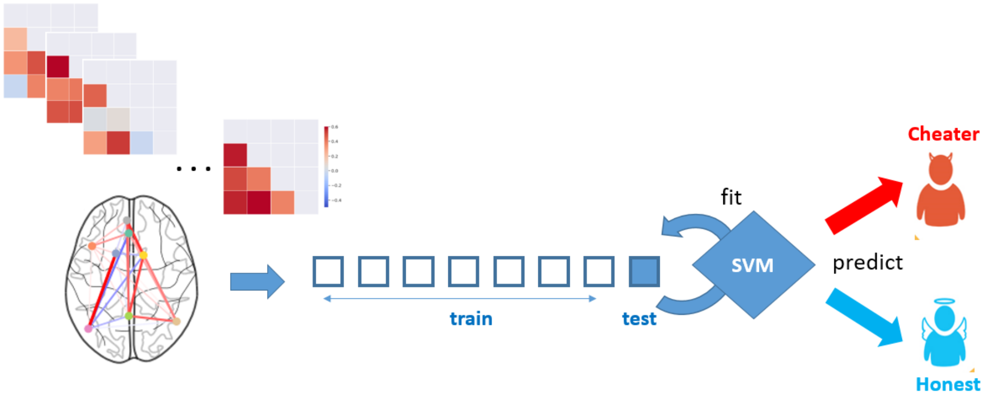


Figure 2.8. Using participants' connectivity patterns within the self-referential thinking network during decision making to classify participants as cheaters or honest participants using support vector classifiers implemented with 8-fold cross-validation.

2.5 Discussion

In this study we explored how neural mechanisms associated with reward anticipation, self-referential thinking and cognitive control determine the (dis)honesty of individual decisions. Using the newly developed Spot-The-Difference task to study trial-by-trial cheating behavior we found that the effect of cognitive control depends on a participants' inclination to be honest or dishonest, in other words, on their moral default.

We found that more honest participants particularly engaged a network of brain regions associated with self-referential thinking when exposed to the opportunity to cheat. Particularly, participants who were generally honest, exhibited higher activity in the self-referential thinking network comprised of the PCC, the bilateral TPJs and the MPFC. We provided further evidence that our results indeed reflect self-referential thinking processes by means of a conjunction analysis with Neurosynth data. Exploring the functional connectivity within the self-referential thinking network, we also found that more honest participants exhibited stronger connectivity during honest decisions between all nodes in this network, whereas this connectivity within the self-referential network broke down during cheated decisions. Collectively, these findings highlight the importance of our moral self-concept and related self-referential thinking processes in promoting honesty.

In line with previous research (Abe & Greene, 2014; Seuntjens et al., 2019), we found that cheaters exhibited stronger sensitivity to reward during decision-making. Our results revealed that all participants were anticipating reward and were sensitive to differences in magnitude of reward during the initial phase of the trial, where the potential reward for finding the differences between the two images is presented, without any presence of moral conflict. However, cheaters, as compared to more honest participants, were more strongly driven by reward when making the decision whether to cheat or not. Specifically, cheaters exhibited higher neural activation in the Nacc, which is an area that has been consistently linked to reward anticipation (Ballard & Knutson, 2009; Knutson, Adams, Fong, & Hommer, 2001; Oldham et al., 2018), during the decision phase. Thus, whereas all participants are sensitive to differences in the magnitude of reward in the absence of moral conflict, particularly the cheaters are driven by the anticipation of reward when making the decisions to cheat.

Importantly, our study is the first to suggest that the function of cognitive control depends on a person's moral default. Particularly, we found that for honest participants, more cognitive control, as represented by higher activity in the left IFG, was needed to cheat, whereas for participants who cheated frequently, control was needed in order to be honest. While honest participants needed cognitive control to overcome their inclination of being honest in order to cheat, cheaters had to exert control to override their greedy tendencies in order to be honest. Thus, our analyses indicated that the role of cognitive control depends on a person's moral default.

In the literature there has been a debate between proponents of the "Will" hypothesis and the "Grace" hypothesis. Research supporting the "Will" hypothesis (Gino et al., 2011; Mead et al., 2009; Welsh & Ordonez, 2014) suggests cognitive control is needed to be honest. In direct opposition to this, another stream of research has accumulated evidence in favor of the "Grace" hypothesis (for meta-analyses, see Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; Verschuere, Köbis, Bereby-Meyer, Rand, & Shalvi, 2018; Carparo, 2017; Spence et al., 2001; Greene & Paxton, 2009), advocating that cognitive control is required for dishonesty.

Our findings help reconcile this conflict as they suggest that people are distributed along a continuum, from individuals who are generally honest to participants who can be considered cheaters. Participants on one side of the spectrum have a default inclination to be honest which is associated with more self-referential thinking when given the opportunity to cheat. In contrast, individuals on the other side of the spectrum have a default inclination for dishonesty and their decisions seem to be driven more strongly by rewards. In order to achieve and maintain a subjectively justifiable balance where one can occasionally profit from cheating but still maintain a positive self-image, people on both sides of the spectrum sometimes need to overcome their initial impulse and default behavior. A generally honest person will need to overcome the default of being honest in order to profit from cheating from time to time, whereas a cheater needs to inhibit the predominant selfish response in order to occasionally be honest and maintain their self-concept.

Thus, it appears that the effect of cognitive control depends on our moral default. For honest people the “Grace” hypothesis applies: honesty results from the absence of temptation and response inhibition is needed to cheat. In contrast, for cheaters the predictions of the “Will” hypothesis apply and active resistance of temptation in the form of inhibition is needed to be honest. Extending findings from cognitive psychology to the social/moral domain, our results suggest that cognitive control seems to serve the purpose of overriding our default behavior. We show that cognitive control processes, well established through previous research on response inhibition (MacLeod, 1991; Eriksen & Eriksen, 1974; Simon & Wolf, 1963), can serve different purposes for different people in the context of (dis)honest decision-making. Our study thus contributes to the reconciliation of the controversy on the role of cognitive control in moral decision making.

In addition, our findings also point to the importance of self-referential thinking processes and the maintenance of a positive self-concept. Whereas previous neuroimaging research has mainly focused on the role of cognitive control and reward sensitivity in cheating behavior, our study is the first to find neural evidence in favor of the self-concept maintenance theory (Mazar et al., 2008). Our results indicate that besides reward and control processes, self-referential thinking as represented by activation in the PCC, MPFC and bilateral TPJs, was engaged, particularly in honest participants, when they were tempted to cheat and more strongly functionally connected with each other when making honest decisions. Thus, our neural evidence suggests that when exposed to an opportunity to cheat, particularly honest people do value their moral self-concept and its maintenance enough to forgo potential financial gains.

Our interpretation of the activation in the PCC, MPFC and bilateral TPJs in terms of self-referential thinking was supported by the observed overlap with a meta-analytically derived activation map. In addition, it has been found that functional connectivity between these areas (Gusnard et al., 2001; Meffert et al., 2013) during rest is associated with higher-level metacognitive operations such as self-reflection and introspection (Qin and Northoff, 2011). Moreover, resting-state functional connectivity between these areas correlates positively with ratings of internal awareness (Vanhaudenhuyse et al., 2011). Based on these findings, we can be fairly confident that the observed network of brain areas can indeed be viewed as neural correlates of the self and self-referential thinking (Carhart-Harris et al., 2014).

Nonetheless, it has to be noted that the same network of regions has also frequently been associated with other psychological processes, most notably theory of mind (Gallagher and Frith, 2003; Ciaramidaro et al., 2007; Schurz et al., 2014) and mental time travel (Buckner et al., 2008; Raichle and Snyder, 2007). While these two processes are distinct from self-referential thinking, they may nonetheless share the common underlying mechanism of perspective taking, as they all require to project one's self to an alternative perspective of time or person (Spreng & Grady, Buckner & Carroll, 2007). In our study it seems unlikely that the activity in this network of regions reflects theory-of-mind processes as our task was designed so that there is no identifiable victim of the dishonest decisions that can be empathized with. Similarly, it seems less plausible that (particularly honest) participants were recollecting past experiences or made plans for the future while making honest decisions. Therefore, we surmise that in our experiment, the observed areas reflect perspective-taking that is required for self-referential thinking.

To examine the generalizability of our findings, we also tested the predictive power of the cognitive control regions in predicting cheating on a trial-by-trial basis using cross-validation. We found that we could significantly predict with high accuracy on unseen data whether on a given trial participants would be honest or would cheat. Moreover, to assess whether connectivity patterns between the different networks contained relevant information about individual differences in honesty, we used support vector classifiers trained on participants' connectivity patterns to discriminate cheaters from honest participants and found that we could indeed accurately classify whether a participant is a cheater or not. Combining the two models did not significantly increase trial-by-trial prediction of cheating (see Appendix 2.7). From the perspective of scientific rigor, cross-validation is a more conservative way to infer the presence of a brain-behavior relationship as compared to correlation or regression, as it is designed to protect against overfitting by testing the strength of the association in a new sample. This increases the probability of successful replication in future studies.

Nevertheless, participants who cheated or were honest on our task, may not act similarly in a different context. Their (dis)honesty, as measured in our task, may thus not reflect a stable moral trait. A recent study using resting state fMRI, however, seems to support the notion that the observed individual differences in honesty may generalize (Speer, Smidts & Boksem, 2020). That study shows that (dis)honesty, as observed in our task, is represented in stable functional connections at rest, within the self-referential thinking network and between the self-referential thinking and the reward and cognitive control network. Resting-state functional connectivity has been shown to be robust and reliable over time and across tasks, which may suggest that the observed correlation with cheating behavior may generalize as well. Further, these individual differences in cheating were also found to be strongly correlated with stable personality characteristics related to impulsivity.

In order to rule out alternative explanations for our findings, we conducted several control analyses. First, in order to test whether neural differences during the decision phase were not driven by differences in levels of engagement with the task, we explored the neural processes during the visual search phase of each trial. As expected for a visual search task, we found that participants showed increased activation in areas related to visual and cognitive processing, working memory and navigation while searching for the differences (see Appendix 2.8). Importantly, no significant differences in neural activations during visual search were found between honest participants and cheaters. This eliminates the possibility that our neural findings were confounded by processes related to differences in engagement or effort during visual search. Second, we also conducted an exploratory factor analysis, which revealed that regions of interest used in our trial-by-trial and functional connectivity analyses indeed belonged to three separate networks that could be clearly identified as the control, reward and self-referential thinking network, respectively (see Appendix 2.9).

In reference to previous neuroimaging research on moral decision-making, our findings align with the early work using hypothetical moral dilemmas (Greene et al., 2004), instructed lying paradigms (Spence et al., 2001, Langleben et al., 2002) and work using the coin-flip task (Greene & Paxton, 2009) in highlighting the importance of the cognitive control network, including areas such as the ACC and IFG, in moral decision-making. As stated above, our findings are also in line with those of Abe and Greene (2014), converging on the conclusion that a more sensitive and responsive reward network is associated with higher levels of cheating.

It is worth noting, however, that reward does not always increase cheating. While higher payoffs were found to increase cheating in sender-receiver games and to some extent in the coin-flip task, they did not have an effect on the die-roll or matrix task (for review see Gerlach et al., 2019). An explanation for this inconsistency may be that reward size may have different, perhaps opposing effects for different individuals (Hilbig & Thielmann, 2017). This aligns with our findings, suggesting that, while for cheaters higher rewards mainly increase the temptation to cheat, for honest participants they may increase the moral cost and associated guilt of cheating, due to the increase in magnitude of the transgression (Mazar, Amir, & Ariely, 2008; Gerlach et al., 2019).

With regard to neural processes linked to self-referential thinking, an fMRI study by Greene and colleagues (2001) found that a network of regions including the MPFC, PCC and bilateral TPJ were involved in making judgements about more personal as opposed to abstract hypothetical moral dilemmas, which they attributed to general emotional processes. More recently, a meta-analysis on neuroimaging research on moral decision-making, conducted by Lisofsky and colleagues (2014) reported that experimental deception paradigms that involved an identifiable victim and consequently perspective taking, were associated with increased activation in the right temporal parietal junction and the bilateral temporal pole, which have been associated consistently with Theory of Mind (ToM) processes (Bahnemann, Dziobek, Prehn, Wolf, & Heekeren, 2009) as compared to less interactive deception and cheating studies. Based on these findings, Lisofsky and colleagues argue that, particularly in studies involving social interaction and an identifiable victim, not only control processes but also perspective-taking and moral reasoning processes are important. Our findings add to their conclusion by demonstrating that also in contexts without an identifiable victim, a similar network of regions, involving the TPJ but also the MPFC and PCC, is crucial in determining the outcome of moral decisions. This suggest that similar neural mechanisms may underlie self-referential thinking and perspective taking processes in the context of moral decision-making.

To conclude, we used a task that is the first to allow measuring cheating on the trial level in an fMRI environment. Using this novel task, we found that not only reward sensitivity but also the extent to which someone engages self-referential thinking processes determines whether someone is a cheater or tends to be honest most of the time. Importantly, we also found that the role of cognitive control in (dis)honesty depends on a person's moral default. These findings may prove to be useful for developing interventions targeted at reducing cheating and dishonesty. Considering the huge economic costs caused by dishonest behavior, such as tax evasion, music piracy or business scandals such as the Volkswagen emission fabrications, reducing dishonest behavior effectively is of great relevance to policy makers and our economy in general.

Taken together, we showed that the neural mechanisms engaged in (dis)honest decisions, ranging from neural activation in reward, self-referential thinking and control networks to functional connectivity patterns, differ fundamentally between honest and dishonest participants. Specifically, we found that cognitive control overrides a person's moral default. Cognitive control allows honest people to cheat at times, whereas it enables cheaters to sometimes be honest. These insights contribute to a deeper understanding of the neural correlates of individual differences in moral decision making. Future research may explore whether neural markers associated with dishonesty are also observable in more stable neural measures such as resting state functional connectivity or structural brain differences.

Appendix Chapter 2

Appendix 2.1 – Visual search task

To further increase the credibility of our cover story on brain processes underlying visual search, we also included the visual search task introduced by Treisman and Gelade (1980) at the beginning of our experiment. Specifically, participants were told that the experiment would start with a simple visual search task and then proceed to visual searches in more complex visual stimuli in the second task. In this first task, the goal was to determine whether a specific target was present or absent. In each trial participants were presented with colored letters presented in random locations on the screen. If the target was present, then participants had to press the left button as quickly as possible. If no target was present, then they had to press the right mouse button as quickly as possible. For this task, participants had to search for a green T. Participants were instructed to answer as quickly as possible while still being as accurate as possible. The task took approximately 5 minutes and was also completed in the scanner while localizer scans were obtained to ensure that scanning noise was audible, so participants would believe this task was indeed part of the study**.** This task was not analysed as it was included solely for the purpose of increasing the credibility of our cover story.

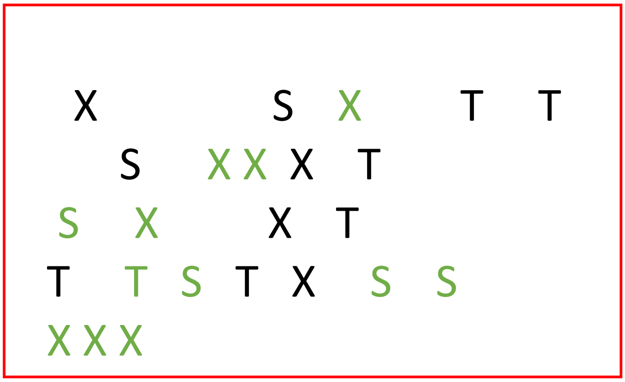


Figure S1. One trial of the simple visual search task. Participants had to indicate whether a green T was among the letters on the screen.

Appendix 2.2 – Validation of the picture set

Stimuli for the task consisted of 144 Spot-The-Difference image pairs that were downloaded from the Internet. Cartoon images of landscapes containing several objects were selected, to make them engaging and challenging enough for the participants. Landscapes were chosen as they generally satisfied the necessary criteria of containing several different objects, which made the task of spotting differences more challenging and engaging. The stimuli consist of pairs of images that are identical apart from a certain number (1-3) of differences that were created by the experimenter using Adobe Photoshop. Differences consisted of objects added to or removed from the landscape picture or changed colors of objects.

To make sure that participants would be able to find the differences between the images in a reasonable amount of time and to reduce the chance of participants believing that they have seen a difference when they have not (false positives), we ran a pilot study on Amazon's Mechanical Turk with 205 subjects using 180 pictures to test the difficulty to spot the differences between the images and to determine the optimal duration of picture presentation. Participants were presented with cartoon image pairs, presented horizontally next to each other, containing three differences and were asked to click on the differences identified in the image on the right hand side. They were given 15 seconds to give their response. Using the heatmap function provided by Qualtrics, regions of interest were defined around the locations of the differences in the image on the right hand side and response times for each of the clicks were recorded. This allowed us to test whether participants were able to find all differences in an image pair, which differences were particularly difficult to find, and how long it took to identify all differences. Based on the responses of these 205 participants, 36 image pairs that took too long or had differences that were too difficult or too easy to spot, were removed, resulting in 144 images that took 92% participants less than 6s to find all three differences (M=5.4s, SD =1.5s). This high success rate of finding the three differences also points to the low ambiguity of the differences, which reduces the chance of false positives. While we cannot completely rule out the chance of false positives it has to be noted that these false positives are unlikely to contribute anything other than noise to the data.

Appendix 2.3 – Cluster statistics for the second-level results for cheatable vs non-cheatable trials

Table S2. Regions more activated during cheatable trials as compared to non-cheatable trials for honest participants as compared to cheaters

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
PCC	1	-9	-57	23.79	492.67	10014
R TPJ	2	45	-60	23.79	445.75	4138
Hippocampus	3	24	-18	-18.33	440.28	3632
(v)MPFC	4	-6	54	-4.29	388.30	3538
Cerebellum	5	0	-54	-60.45	379.48	3222
MFG	6	-30	24	44.85	407.82	3032
Cerebellum	7	-27	-48	-25.35	429.12	2811
Left Frontal Pole	8	-18	39	44.85	421.57	2337
MPFC	9	-6	30	6.24	39.30	2053
L TPJ	10	-45	-69	23.79	38.96	1674
R. Postcentral Gyrus	11	30	-42	65.91	382.64	1547
R Supramarginal Gyrus	12	66	-30	27.3	448.58	1263
L Supp motor area	13	0	0	48.36	365.03	1232
L C	14	-18	-42	-49.92	411.61	1105
R Cerebellum	15	12	-45	-11.31	429.17	1105
R Hippocampus	16	21	-39	6.24	466.65	1105
L OFC	17	-42	36	-14.82	431.2	1042

Appendix 2.4 – Regions extracted for ROI analyses

Depicted here are the tables showing the regions extracted from Neurosynth.

Table S1. ToM and cognitive control masks link for download

Network	Studies	Date of	Link to download
Self Referential	166	03.06.2019	http://neurosynth.org/analyses/terms/self%20referential/
Cognitive Control	598	03.06.2019	http://neurosynth.org/analyses/terms/cognitive%20control/
Reward Anticipation	92	03.06.2019	https://neurosynth.org/analyses/terms/reward%20anticipation/

Appendix 2.5 – Cluster statistics for the second-level results for cheated vs honest decisions

Table S3. Regions more activated during honest decisions as compared to cheated decisions for cheaters than for honest participants

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
L IFG	1	-46	21	-5	41	4156
R ACC	2	7	36	21	384	2797
L ACC	3	-7	41	7	450	1922
R Insula	4	37	26	-6	387	762
L Frontal Pole	5	-34	62	5	476	704
L Supp Motor Area	6	-11	23	63	398	639
L Nacc	7	-14	19	-7	372	326
L SFG	8	-4	20	43	356	272
R Cingulate Gyrus	9	1	-28	29	330	237
R Angular Gyrus	10	54	-51	45	331	200

Here we also find the left Nacc to be activated, which seems inconsistent with the other findings. However, it has to be noted that these activations, including the Nacc, were further tested in the trial-by-trial analysis. In this trial-by-trial analysis we investigate which of the previously identified regions is most important in predicting trial-by-trial cheating. This analysis includes the NAcc as well as the ACC and the IFG, and there the Nacc was not found to be a significant predictor of the decisions to be honest for cheaters. This suggests that the cluster reported here may have been a false positive. An alternative explanation could be that cheaters experience a warm glow effect (Andreoni, 1990), which proposes that people behave selflessly or morally because they are compensated by the warm glow of knowing they have acted prosocially. Honest participants may intuitively act honesty without further thinking about it whereas cheaters may do so more rarely and when they do so they experience the warm glow which is represented in the Nacc. This is, however, very speculative and further research would be needed to confirm these speculations.

Appendix 2.6 – Cluster statistics for the second-level results of the parametric modulation analysis for the level of reward

Table S4. Regions parametrically modulated by level of reward during the level of difficulty phase of the Spot-The-Difference task

Region	cluster_id	peak_x	peak_y	peak_z	peak_value	volume_mm
Left Cuneus	1	-9	-78	16.77	546	1611
R Nacc	2	12	12	-0.78	493	1232
L Nacc	3	-21	15	-0.78	47	568
L Cuneus	4	-6	-96	27.3	414	315

Appendix 2.7 - Classifying cheaters versus honest participants and predicting trial by trial cheating

Due to the fact that we found that we could classifiy cheaters and honest participants based on the functional connectivity patterns during decision-making, we wanted to see whether average activation within a subject in the ROIs from the three networks of interest (cogntive control, reward & self referential thinking) could be used to classify participants as cheaters or honest participants (categorized by median split). In order to do this, we average the trial-by-trial estimates within participants, resulting in one observation for each subject, which represents the average activation in each ROI across the whole task.

In order to test this, we employed a support vector classifier (Cox & Savoy, 2003; Mitchell, 2004) with a linear kernel (C=1), trained on average activations in the ROIs of each participant to determine whether a participant was a cheater or an honest participant (categorized by median split). To avoid overfitting and inflated prediction accuracy (Vul et al., 2009), this was done using 8-fold cross validation. Significance was estimated using permutation testing (N=5000). The classification analysis revealed that we could significantly classify an unseen participant as a cheater or an honest participant based on the average activation in the ROIs (F1=70%, AUC=77%, $p<0.05$). Using activations from honest trials only an even higher classification accuracy was found (AUC=84%, $p<0.05$). Classification was not significant using cheated trials only.

We also tested whether combining the model predicting trial-by-trial cheating, using the trial-by-trial activation from the ACC and IFG, could be improved by adding the output from the model classifying cheaters versus honest participants based on the participants connectivity patterns. In order to increase statistical power, instead of using a support vector machine trained on participants that were median split on cheatcount, we used a support vector regression approach to predict the cheatcount of an unseen participant based on participants' connectivity patterns. This allowed us to use the full range of the participants' cheatcounts. Specifically, as in the model reported in the manuscript, we used 8-fold cross validation to train a support vector regression (SVR) model on the connectivity patterns of our participants to predict the cheatcount of an unseen participant. The predictions from the SVR model correlated significantly with the cheatcount ($r=0.73$, $p<0.05$), demonstrating the predictive accuracy of the SVR model. In a direct model comparison, adding the output from the SVR to the multilevel model with ACC and left IFG led to a significantly improved fit ($\chi^2=14.1$, $p<0.05$). However, when testing the model using 8-fold cross validation, no substantial improvement in predictive accuracy was found (AUC=79%, F1=85% as compared to AUC=76%, F1=89%). This could be due to the intercept of the multilevel model already accounting for individual differences in moral default that are similarly explained by the connectivity in the self-referential thinking network.

To test this conjecture, we also trained support vector machines without an intercept capturing individual differences, on the trial-by-trial data with the activity from the control regions and with or without the output from the connectivity model. This analysis revealed that when using only the control regions, a considerably lower predictive accuracy (AUC=68%) was found as compared to the model with the output from the connectivity model included (AUC=75%). It can thus be concluded that the intercept in the multilevel model indeed captures individual differences in moral default, that are also explained by the output of the model trained on connectivity patterns. In this sense, adding output from the connectivity model increases the interpretability of the model as individual variation in moral default is explicitly captured by variation in connectivity between regions in the default mode network.

Alternatively, or additionally, the ACC, which is already included in the model predicting trial-by-trial cheating, may encode individual differences in moral default that are similarly captured by connectivity within the self-referential thinking network. Whereas we found that higher activity in the IFG increases the probability of cheating in honest participants and decreases the probability of cheating for cheaters, no such effect was found for the ACC. The ACC has been frequently associated with conflict monitoring and conflict detection (Carter & Van Veen, 2007) and may encode individual differences in moral default to some extent. Stated differently, the extent to which honest participants monitor and detect moral conflict may differ from cheaters and may reflect individual differences in moral default.

Appendix 2.8 – Levels of engagement during visual search

In order to test whether our findings may be confounded by different levels of engagement during the visual search phase, we tested whether there were differences in neural activation during the visual search phase between more honest participants and cheaters. First, we ran a univariate analysis in which we contrasted neural activity during the visual search against baseline activation. The analysis revealed that a large cluster in the visual cortex showed higher activation during search as compared to baseline activation, which is expected as participants were engaged in visual search. In addition, several regions related to working memory, cognitive processing and navigation, such as the dmPFC and the MFG were more strongly activated during visual search (see Table S5 for table with cluster statistics).

To explore whether there are individual differences in level of engagement during visual search, participants' cheat count was added as a group level covariate. The whole brain analysis revealed that there are no significant differences between more honest participants and cheaters during the visual search phase. In addition, we also tested whether differences in neural activation during visual search between cheatable and non-cheatable trials were more strongly expressed in cheaters or honest participants. In order to do so, a univariate analysis was run in which we contrasted neural activation during visual search in cheatable trials against activation during visual search in non-cheatable trials. Again, these contrast maps were then correlated with cheat count on the group level. The whole brain analysis did not reveal any significant effects. These findings suggest that there are no significant differences in level of engagement or motivation during visual search between more honest participants and cheaters.

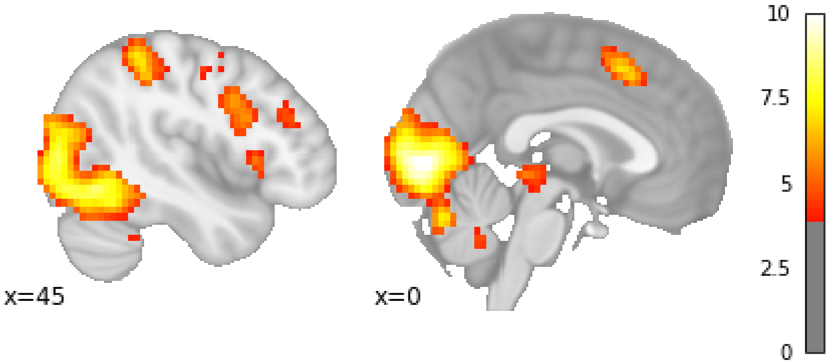


Figure S2. The visual cortex, dMPFC and left and right dlPFC are more activated during visual search as compared to baseline

Table S5. Regions more activated during visual search as compared to during rest

Region	**cluster_id**	**peak_x**	**peak_y**	**peak_z**	**cluster_mean**	**volume_mm**
Occipital Cortex	1	0	-84	2.73	643.202	301369
dmPFC	2	0	15	48.36	514.447	16490
MFG	3	24	6	51.87	515.819	6981.39
R dlPFC	4	45	6	30.81	488.557	6160.05
R Insula	5	30	27	-0.78	553.421	5907.33
R dlPFC	6	-48	0	30.81	448.594	3601.26
L Insula	7	-33	21	-0.78	514.818	3064.23
Cerebellum	8	-18	-42	-46.41	490.497	1769.04
R lPFC	9	51	36	27.3	439.172	1674.27
Cerebellum	10	-30	-69	-53.43	494.043	663.39

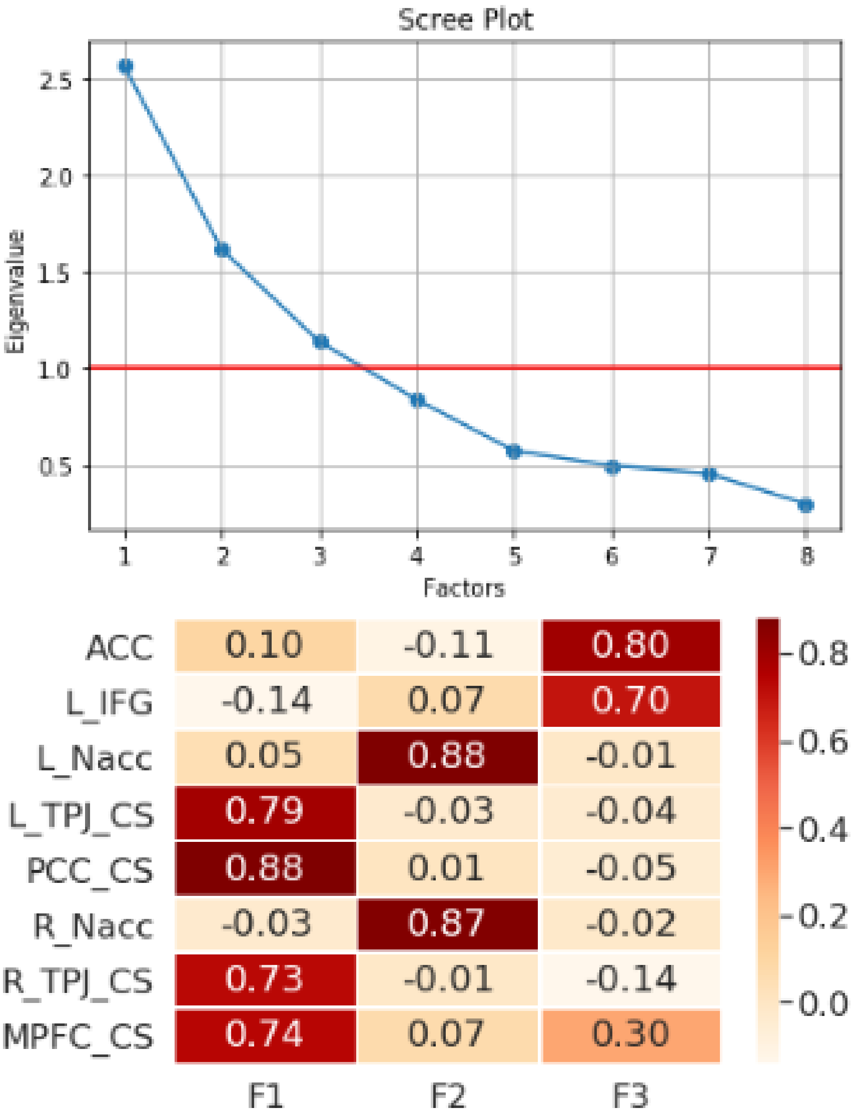
To corroborate this neural evidence with behavioral data, we tested whether there were significant differences in accuracy on the simple visual search task (see Appendix 1) between honest participants and cheaters. The analysis revealed that there were no significant differences between honest participants and cheaters in accuracy on the simple visual search task (t = 1.17; p = 0.25; participants were categorized in groups by median split). Assuming participants were honest on three differences trials, we could also compare the behavioral accuracy between cheaters and honest participants on the Spot-The-Difference task. We performed this analysis and found no significant differences (t=1.54, p=0.16) in how often cheaters or honest participants (as categorized by median split) reported to have found three differences when there were actually three differences. Collectively, these findings suggest that there were no significant differences in levels of engagement during the visual search of the Spot-The-Difference task.

Appendix 2.9 - Factor analysis to confirm validity of networks

To test whether the regions we are analyzing indeed belong to three separate networks, we conducted an exploratory factor analysis with promax rotation (Hendrickson & White, 1964), which is an oblique rotation method which allows for correlation between latent factors. Specifically, the goal of this factor analysis was to determine the most important latent factors underlying all the regions resulting from our conjunction analyses, namely the left IFG and ACC (cognitive control network), the PCC, bilateral TPJs and MPFC (self-referential network), and the bilateral Nacc (reward network).

We used the single trial activations obtained as explained above by fitting a model that includes a separate regressor for each trial from each of the regions as input for the factor analysis. Before conducting the factor analysis, we first checked whether the regions intercorrelated at all using Bartlett's test of sphericity, which tests the observed correlation matrix against the identity matrix. Bartlett's test indicated that the null hypothesis can be rejected and there is significant correlation between variables justifying a factor analysis ($\chi2$ =10582, $p<0.001$). In addition, the Kaiser-Meyer-Olkin (KMO) test was conducted which determines the adequacy of the observed variables by estimating the proportion of variance among all the observed variables. The KMO test revealed an overall estimate of 0.69 which indicates that the observed variables are adequate for a factor analysis.

Next, we determined the number of factors with the help of the Kaiser criterion (choosing factors with an eigenvalue >1). This resulted in three latent factors, where the first factor represented the self-referential thinking network with the bilateral TPJs, PCC and the MPFC loading highly on this factor. The second factor clearly represents the reward network as only the bilateral Nacc show high factor loadings. Lastly, the third factor clearly represents the cognitive control network as only the ACC and the left IFG load highly on this component. This exploratory factor analysis clearly indicates that the regions of interest used in our trial-by-trial and functional connectivity analysis indeed belong to three separate networks.



	F1	F2	F3
ACC	0.10	-0.11	0.80
L_IFG	-0.14	0.07	0.70
L_Nacc	0.05	0.88	-0.01
L_TPJ_CS	0.79	-0.03	-0.04
PCC_CS	0.88	0.01	-0.05
R_Nacc	-0.03	0.87	-0.02
R_TPJ_CS	0.73	-0.01	-0.14
MPFC_CS	0.74	0.07	0.30

Figure S3. Left: Scree plot showing the eigenvalues for each factor. Right: the loadings for each of the factors

Appendix 2.10 – Example image pairs for the Spot-the-Difference task

To provide a better sense of how difficult it was to spot the difference between images if there were indeed three differences, three example pairs are shown. For the sake of space, image pairs are presented horizontally (next to each other), whereas in the actual Spot-The-Difference task the image pairs were presented vertically (on top of each other). All images are also available in the publicly available repository.

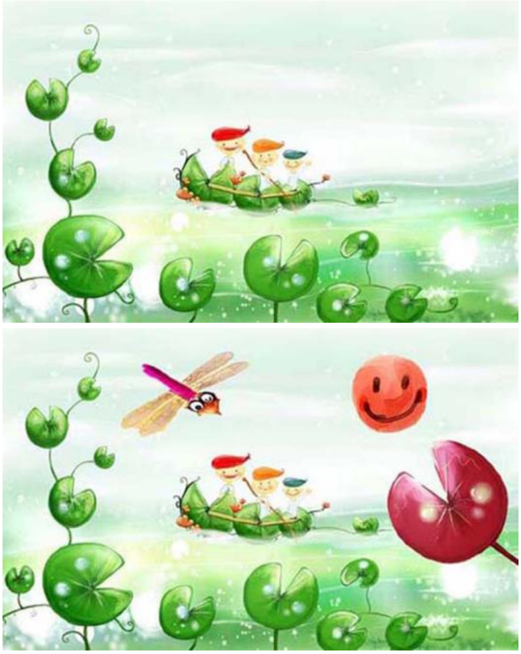


Figure S4. Example image pair used in the Spot-The-Difference task. The right image contains a dragonfly (top left) a red sun with a smiley (top right) and a red flower (middle right), which are not present in the left image.

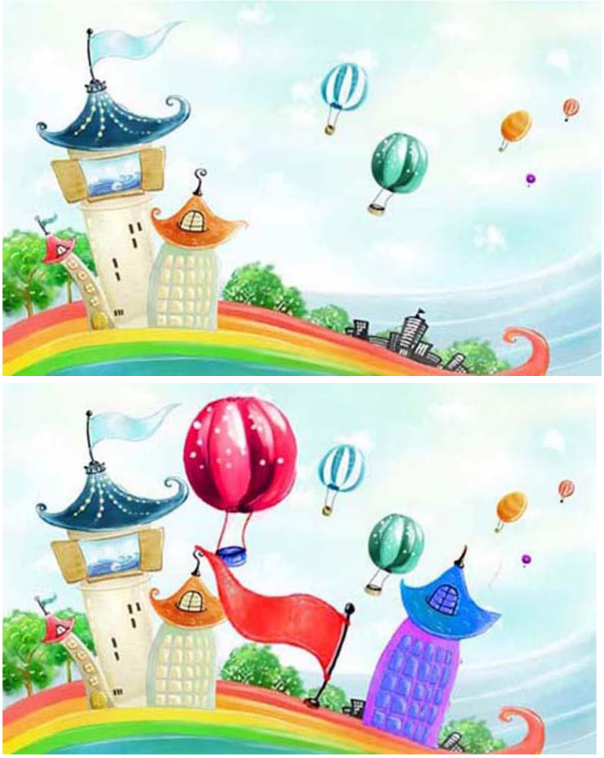


Figure S5. Example image pair used in the Spot-The-Difference task. The right image contains a hot air balloon (top left) a red flag (centre) and a purple building(right), which are not present in the left image.

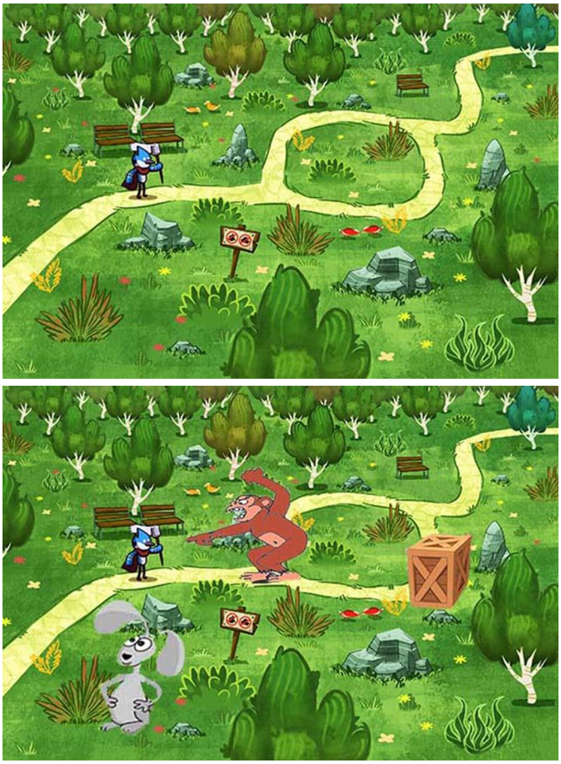


Figure S6. Example image pair used in the Spot-The-Difference task. The right image contains a rabbit (left) a monkey (centre) and a wooden box (right), which are not present in the left image.

Chapter 3

Replicating the contingency of cognitive control on moral default[2]

3.1 Abstract

Cognitive control is crucially involved in making (dis)honest decisions. However, the precise nature of this role has been hotly debated. Is honesty an intuitive response or is willpower in the form of cognitive control needed to override an intuitive tendency to cheat? Recently, a reconciliation of these conflicting views proposed that cognitive control enable dishonest participants to be honest, whereas it allowed cheating for those who are generally honest. Thus, cognitive control does not promote (dis)honesty per se; it depends on one's moral default. In the present study, we further tested this proposal using EEG in combination with an external localizer task to mitigate the problem of reverse inference. Our analysis revealed that the same neural signature evoked by cognitive control demands in the Stroop task can be used to estimate (dis)honest choices in an independent cheating task, providing converging evidence that cognitive control can indeed help honest participants to cheat, whereas it facilitates honesty for cheaters.

3.2 Introduction

Dishonesty imposes a lasting social toll by undermining personal relationships as well as sabotaging trust in social institutions. Although dishonesty may be ubiquitous, it is nonetheless apparent that not everybody is a cheater: there are considerable individual differences in our moral flexibility. Previous research has found that, when given the opportunity, individuals differ considerably in how often they cheat (Gino et al., 2012,

[2] A revised version of this chapter has been accepted for publication at the Journal of Neuroscience.

2014; Speer, Smidts & Boksem, 2020; for a meta-analysis on dishonesty see: Gerlarch, Teodorescu & Hertwig, 2019).

When exposed to the temptation to cheat, evidently, the (monetary) incentive is important in driving whether someone will succumb (Becker, 1968, Allingham & Sandmo, 1972; Seuntjens et al., 2019; Abe & Greene, 2014, Speer, Smidts & Boksem, 2020). However, most individuals do not only consider (financial) rewards when tempted to cheat, which explains why we observe behaviors such as altruism, reciprocity and honesty. Converging evidence indicates that our self-concept, our perception of ourselves as moral beings (Aronson 1969; Baumeister 1998; Bem 1972), motivates us to be honest (Mazar, Amir, & Ariely, 2008; Speer, Smidts & Boksem, 2020).

Recently, an fMRI study (Speer, Smidts & Boksem, 2020) investigated how the brain accomplishes this task of arbitrating between reward and self-concept. The study showed that brain regions associated with cognitive control helped dishonest participants to be honest, while they enabled cheating for more honest participants. Therefore, the findings indicated that cognitive control is not needed to be honest or dishonest per se but that it is contingent on an individual's moral default.

Despite these compelling results, it has become clear over the last decade that a single finding cannot be taken as sufficient support for any novel theory or hypothesis. The importance of conducting replications has been highlighted as several authors have questioned the validity of numerous findings in neuroscience (e.g. Ioannidis, 2005; Button, et al., 2013; Botvinik-Nezer et al., 2020). If the reliability of a considerable proportion of neuroimaging results is questionable, this has grave consequences for the scientific community as future work is built on a fragile fundament. The most promising way of addressing these issues is to attempt to replicate findings from previous studies, as the probability of a finding being true increases with the number of replications.

An additional issue with most neuroimaging studies on dishonesty thus far is that they have used reverse inference to infer the neurocognitive processes determining the outcome of (dis)honest decisions, where the cognitive operations are inferred from activation in an observed region of the brain. While reverse inference can be informative, it should be interpreted with caution depending on how selective these areas of interest are activated by a specific cognitive process (Poldrack, 2006). Specifically, a given brain area may be involved in several different cognitive processes, which makes it difficult to conclude that activation observed in this area represents one specific function.

Here, we attempt to replicate previous findings (Speer, Smidts & Boksem, 2020) with a different neuroimaging method in combination with an additional localizer task to circumvent reverse inference problems. Using a different modality has the potential to substantially increase the convergent validity of the previous finding, as it allows to capture different aspects of the underlying neural processes due to the higher temporal resolution of EEG. We focus on cognitive control specifically, since it has been identified as the central process in determining the outcomes of (dis)honest decisions (Speer, Smidts & Boksem, 2020). Here, we elicited the neural patterns associated with cognitive control by means of recording the electroencephalogram (EEG) of participants while they completed the Stroop task, which is a well-established task to elicit cognitive control. Subsequently, participants engage in the Spot-the-Difference task (Gai & Puntoni, 2021; Speer, Smidts & Boksem, 2020), which allows participants to cheat repeatedly, deliberately, and voluntarily while recording their EEG. If we observe the same neural patterns in both tasks, we can deduce with empirical support that neural patterns we observe in the Spot-The-Difference task do indeed reflect cognitive control processes, relating to response inhibition. Our study thus contributes by more rigorously testing the neural mechanisms underlying dishonest decisions and reducing issues of reverse inference.

3.3 Methods

3.3.1 Participants

EEG recordings were obtained from 35 participants. One participant completed the Spot-The-Difference task, but not the Stroop task, so for that task there were 34 participants. The data of 2 participants had to be discarded because for these participants 27 or more channels (out of 64) were identified as bad channels by the Autoreject algorithm (see preprocessing section; Mainak et al., 2017), which categorized them as outliers (IQR rule, 2 standard deviation rule). The reported analyses are based on the remaining 33 (32 for the Stroop task) participants (18 females; age 18 to 29; M = 21, SD = 2.6), recruited from an online community for university students, where students can sign up for experiments. An initial screening interview ensured that all participants were right-handed with normal or corrected to normal vision, spoke English fluently, were not on any psychoactive medication influencing cognitive function, and had no record of neurological or psychiatric illness. The study was approved by the internal review board and was conducted according to the Declaration of Helsinki.

3.3.2 Task and stimuli

3.3.2.1 Spot-The-Difference task

As described in Chapter 2, in the Spot-The-Difference task, participants were presented with pairs of images and were instructed that there were always three differences present between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were requested to find three differences between the images. Since reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design allowed and encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were present in the images). For details regarding this task and the stimuli. I refer to Chapter 2 paragraphs *2.3.2.1* & *2.3.2.2*.

3.3.2.2 Stroop task

In order to localize the neural signature of cognitive control the Stroop task was used. In the Stroop task participants view color names presented in different ink colors and are instructed to read out the color of the ink as fast as possible. In incongruent trials, color names and the color of the ink are not matching, while on congruent trial the name and the color of the ink match. On the incongruent trials, participants must inhibit the predominant response of reading the word and just focus on the ink of the color. Thus, the overlearned response of reading interferes with successful task performance.

To allow conducting the Stroop task without participants speaking, which would cause artefacts in the EEG signal, we adapted a version developed by Zysset and colleagues (2007). Participants were instructed that they would see two words vertically arranged (see Figure 3.1) and they would have to decide, by means of a button-press (yes, no), whether the color of the top word corresponds to the color name written at the bottom row. There were 72 trials in total. Half of the trials were corresponding, meaning that the color of the top word matches the meaning of the bottom word. There were 36 congruent and 36 incongruent trials. In congruent trials the color of the upper word was consistent with the word meaning of the upper word. For incongruent trials this was not the case. Each trial starts with a fixation cross with a duration jittered between one and two seconds. Subsequently, to prevent subjects from focusing on the lower word and blurring out the top word, the top word was presented 150ms before the lower word. As a result, visual attention was shifted automatically to the top word and participants were forced to read the top word enabling interference. Next, the bottom word and a white fixation cross were presented. The white fixation cross turned orange as soon as a response was made, to signal to the participant that the response has been recorded. After the words were on the screen for 3 seconds the feedback was presented (see Figure 3.1).

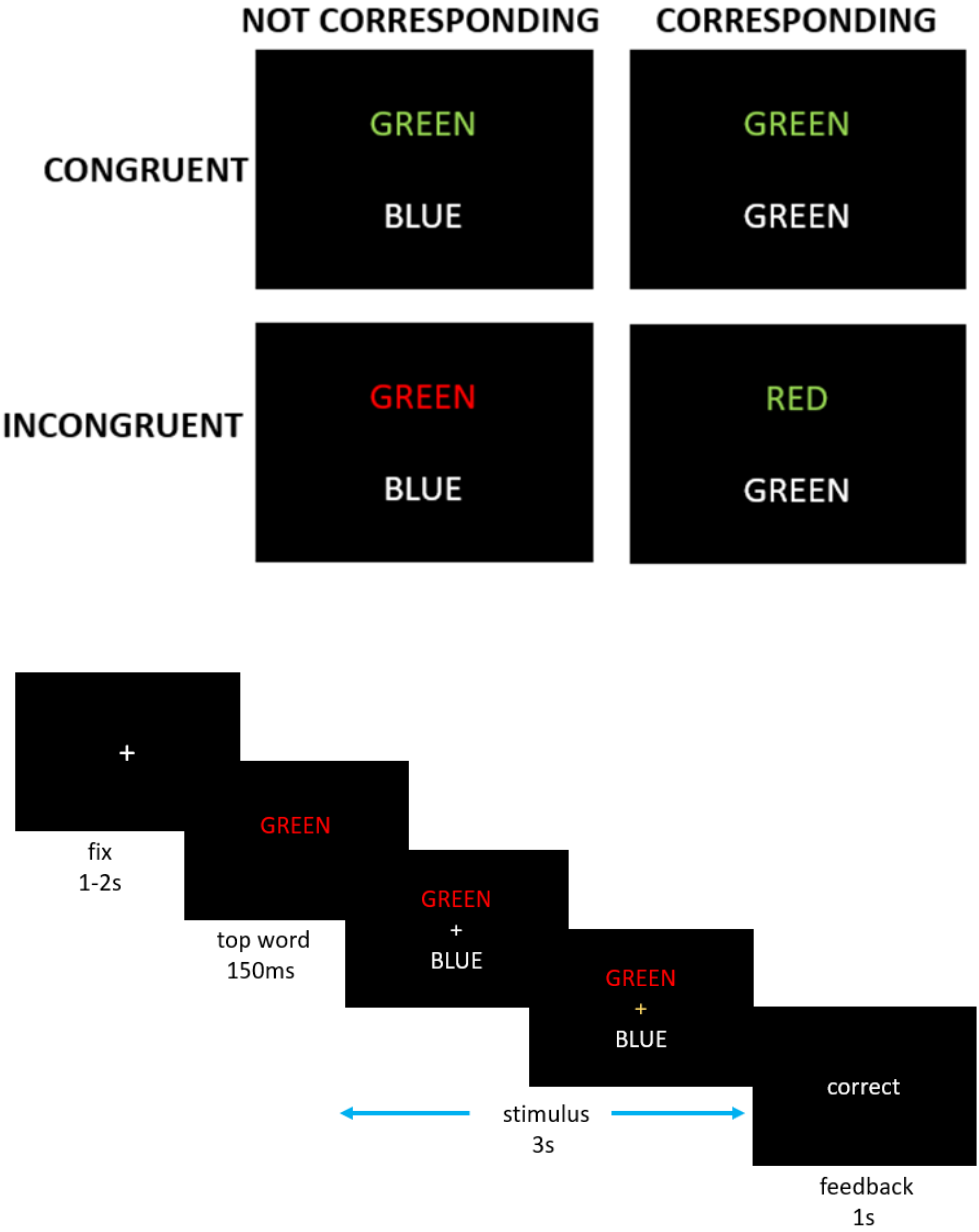


Figure 3.1. An example trial of the Stroop task. Participants are asked whether the color of the upper word corresponds with the meaning of the lower word. Top: The four conditions of the Stroop Task. Bottom: The sequence of a trial in the Stroop task.

3.3.3 Experimental procedure

Before the experiment started, participants were introduced to the cover story, the tasks and they signed the informed consent form. During the EEG data collection, participants were seated in a slightly reclining chair positioned in front of a 19-inch PC monitor in a sound-attenuated, electrically shielded, dimly lit room. Participants first completed practice trials for both visual search tasks. Afterwards, the participants completed the simple visual search task (5 min) followed by the Spot-The-Difference task which took approximately 40 minutes. Subsequently, participants completed practice trials for the Stroop task and the actual Stroop task which lasted around seven minutes. After completing all tasks, participants left the EEG room and filled-in a short questionnaire including questions about their thoughts on the purpose of the task on a computer in a separate testing room.

3.3.4 EEG acquisition and preprocessing

3.3.4.1 Acquisition and preprocessing

We recorded EEG from 64 active scalp electrodes using a Biosemi Active Two system (Biosemi, The Netherlands). Additional flat type electrodes were placed on the right and left mastoid, and in the eye region in order to record eye movements or electro-oculograms (EOG): Electrodes were placed below and above the left eye in line with the pupil to record vertical EOG, and at the outer canthi of both eyes to record horizontal EOG. The EEG and EOG signals were sampled at a rate of 512 Hz. All preprocessing was done using the MNE package in python (Gramfort et al., 2013). EEG data was filtered with a low cutoff filter of 1Hz to remove slow drifts and a notch filter of 50Hz to remove line noise. Subsequently, bad and noisy channels were detected by means of correlation, checking how well a given channel is correlated with all other channels, the robust z-score deviation aggregates per channel, the robust z-score estimates of high frequency noise per channel and the RANSAC channel correlations , which is the correlation for each channel with itself across the original data versus the RANSAC predicted data, as implemented in the PREP pipeline (Bigdely-Shamlo et al., 2015). After detection, these channels were removed from the data and subsequently interpolated (i.e., estimated from surrounding channels). The EEG data was then re-referenced to the average signal across channels. As a next step, ocular artefacts were removed by performing an independent component analysis (ICA) on the data and then correlating the resulting components with the EOG channels to see which of the components represents the ocular artefacts. The component that correlated the highest with the EOG channels was then removed from the EEG data.

3.3.4.2 Epoching & artefact rejection for the Spot-The-Difference task

The EEG data from the Spot-The-Difference-Task was then segmented into 4 second epochs, time-locked to the onset of the decision phase. The epochs were baseline corrected using the last second of the fixation period preceding the presentation of the image pair, which occurred seven to six seconds before the decisions phase. The resulting epochs were then subjected to *Autoreject*, an automated artefact detection algorithm based on machine learning classifiers and cross-validation to estimate the optimal peak-to-peak threshold (Jas et al., 2017). On average 3% of trials (~4 trials out of 144 trials, SD=5%) were rejected. This algorithm was implemented to remove artefacts not identified by previous preprocessing steps and, depending on the number of bad sensors for a given trial, either repairs the trial based on interpolation or excludes it from further analysis. The preprocessed data were then submitted to a morlet wavelet analysis to transform the data into the time-frequency domain with 18 log scaled frequency bins ranging from 4Hz to 40Hz in order to have higher sensitivity in lower frequency ranges such as the theta band. To optimize both spectral and temporal resolution, the number of cycles to include in the sliding time window were defined by dividing each individual frequency by two. After transform the data to the time-frequency domain the data was decimated by a factor of 4 (sampling every 4th timepoint) to increase computational efficiency.

3.3.4.3 Epoching & artefact rejection for the Stroop task

Following the same preprocessing as described above, the EEG data from the Stroop task was segmented into two second epochs time-locked to the presentation of second word on the screen with one second preceding the onset and one second following the onset. The epochs were baseline corrected using the second preceding the onset of the word. The same artefact rejection as described above was applied to the Stroop task and in this task on average 2% (SD = 4%) of trials for each subject was rejected. The same morlet wavelet analysis was used to transform the data to the time frequency domain.

3.3.5 Statistical analysis

3.3.5.1 Contrasting conditions in the Stroop task

To localize the neural correlates of cognitive control related to interference in the Stroop task, we conducted a whole brain multilevel analysis contrasting the conditions of the Stroop task with the highest behavioral difference (congruent/corresponding vs. congruent/not corresponding; see *Task and Stimuli* for details) from EEG data from all electrodes and spectral points. At the first level (i.e., the participant level), we computed the averaged time-frequency maps for each of the two conditions. We then tested the resulting average maps at the second level for significant group effects, using a paired sample t-test. We used cluster-based permutation testing as a stringent control for multiple comparisons (Maris and Oostenveld 2007). Specifically, for every sample within a channel, we quantified the experimental effect by a t-value. We selected samples for which the t-value was larger than a given threshold ($p < .05$) for potential inclusion in a cluster. It has to be noted that the threshold used does not affect the false alarm rate of the final statistical test; it only sets a threshold for considering a sample as a candidate member of a cluster. We subsequently clustered selected samples in connected sets based on temporal and spectral adjacency, and we computed cluster-level statistics by taking the sum of the t-values within every cluster. Subsequently, we performed permutation testing utilizing the Monte Carlo method to compute the posterior significance probability of our observed effect (Maris and Oostenveld, 2007). Note that the cluster-based permutation test was conducted separately for each channel, to explore which channels contained significant clusters.

3.3.5.2 Contrasting cheatable vs. non-cheatable trials

Using a mask, consisting of the significant channels and frequencies from the Stroop task, we then tested at what point in time, and in which channels, cheatable trials differ from non-cheatable trials. Specifically, we computed the averaged time-frequency maps of cheatable trials and non-cheatable trials for each subject. We then applied the mask from the Stroop task (i.e. the spectral points within the channels found to be significant in the analysis of the Stroop effect) to the resulting average maps. For group-level analysis we then conducted the cluster-based permutation testing procedure described above to the masked contrast maps.

3.3.5.3 Trial-by-trial analysis

To test whether activation in the selected channels in the theta band could be used to estimate the actual decision to cheat or to be honest, we conducted a trial-by-trial analysis. Based on the findings from the analysis above, we focused particularly on the channels, time window and frequency range which most strongly differentiate cheatable and non-cheatable trials. We used this time window and frequency range to test whether we can estimate (dis)honesty on each trial. We extracted trial-by-trial power from these channels, time window, and frequencies of interest, by extracting the first factor from an exploratory factor analysis. A factor analysis (without rotation) was used to extract the most amount of variance from the time window of interest. The advantage of using a factor analysis over using the mean is that different items (timepoints) can be given different weights to maximize the correlation between timepoints in this time period (DiStefano et al., 2009). This resulted in one data matrix where the rows represent trials and the columns represent the channels of interest.

Given the nested structure of our data (trials within participants) we then conducted a multilevel analysis to estimate trial-by-trial cheating. The dependent variable was the binary response with a logit link (cheating = 1, honest = 0). The power in the theta range within the time window for each of the channels of interest served as trial-level regressors. In addition, the average cheatcount (number of times a participant cheated) was added as a subject-level regressor. To reduce multicollinearity between the regressors of interest (see Appendix 3.1) and to explore which of the channels is most important in estimating cheating on the trial-level, we performed variable selection for generalized linear mixed models by means of L1-penalized estimation. This was implemented using the glmmlasso package in R, which applies a gradient ascent that enables us to maximize the penalized log-likelihood, yielding models with reduced complexity (Groll & Tutz, 2014). The lasso regression adds a penalty term to the equation which shrinks less important coefficients in the model to zero and thus reduces complexity of the model and multicollinearity of regressors (Tibshirani, 1996). In this way it also selects the most important regressors in the model.

3.4 Results

3.4.1 Behavioral results

Substantial individual differences in the total amount of cheating were observed (Mean= 41%, Median=30%, SD=29%; see Figure 3.2): some participants cheated no more than three times out of 72 trials (6% of participants), whereas others only missed one or two opportunities to cheat (also 6 %).

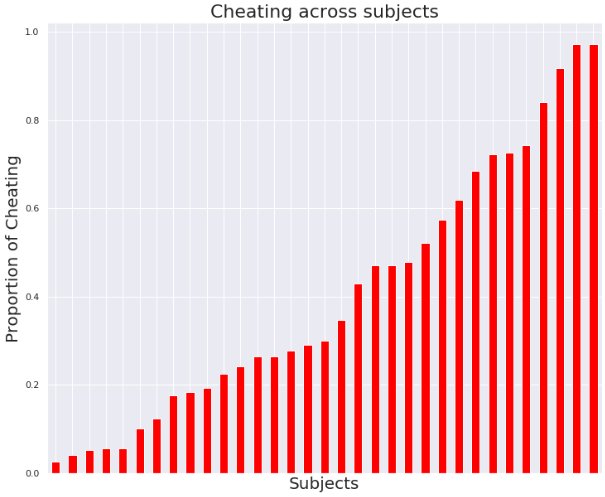


Figure 3.2. Individual differences in proportion of cheating on the Spot-The-Difference task. N = 33.

We investigated how different task characteristics of the spot-the-difference task influenced cheating behavior. Due to the nested structure of our data (trials within different numbers of differences and rewards within participants), we applied a multilevel logistic regression analysis on our behavioral data. This analysis considered cheatable trials only. The dependent variable was the binary response (cheating vs. honest) with a logit link (cheating = 1, honest = 0). The number of differences, trial number and level of reward served as trial level regressors. The model allowed for random intercept within participants. Replicating findings from a previous study (Speer, Smidts & Boksem, 2020), this analysis revealed a significant effect of the number of differences on cheating behavior ($b = 1.02$, $SE = 0.08$, $z = 12.86$, $P < 0.001$). This indicates that participants cheated more frequently when the crime is smaller (they claimed to have found three differences more often when there were two differences as compared to when there was only one). As in the previous study (Speer, Smidts & Boksem, 2020), no significant effect of reward magnitude (see Methods) on cheating behavior was found. Further, no significant effects of reaction time on cheating were found, which is probably due to the fact that participants had three seconds to make up their mind and form a decision before the response options appeared and a response could be made (see Methods). Also the trial number regressor, testing for possible fatigue or habituation effects, resulted in no significant effects.

3.4.2 Effects of congruency and correspondence on reaction time in the Stroop task

In order to test whether there were significant differences in cognitive control demands between the different conditions of the Stroop task, we tested the effects of the conditions, namely congruency (whether the color of the top word matches the meaning of the top word) and correspondence (whether the color of the top word matches to the meaning of the bottom word), and their interaction on reaction time. Here, the assumption is that longer reaction times reflect higher cognitive control demands. Due to the nested structure of the data, a multilevel level regression model was used. The analysis revealed that both Congruency ($b = 282.09$, $SE = 23.55$, $t = 11.99$, $P < 0.001$)

and Correspondence (b = 306.20, SE = 21.27, t = 14.39, P < 0.001) had a significant effect on reaction time. In addition, the interaction effect between Congruency and Correspondence was found to be significant (b = -354.98, SE = 25.09, t = -14.15, P < 0.001). The results show that incongruent as well as non-corresponding trials resulted in significantly higher reaction times than their congruent and corresponding counterparts. The combination of incongruency and not corresponding words also led to significantly higher reaction times (see Figure 3.3). Thus, congruent and corresponding trials had significantly lower cognitive control demands than any of the other conditions (see Figure 3.3). For parsimony, we will from now on only focus on the difference between congruent/corresponding and congruent/not -corresponding trials, since this is the contrast where the largest difference in reaction time between conditions was observed. On congruent/non-corresponding trials, the meaning of the top word is congruent with color of the top word, leaving participants particularly inclined to categorize this trial as a corresponding trial. In order to respond accurately, they must override this intuitive impulse, which demands a high degree of cognitive control, resulting in the longest reaction times. We also repeated the neural analysis for the congruency contrast (incongruent-corresponding trials vs. congruent-corresponding trials) and found similar results (see Appendix 3.2).

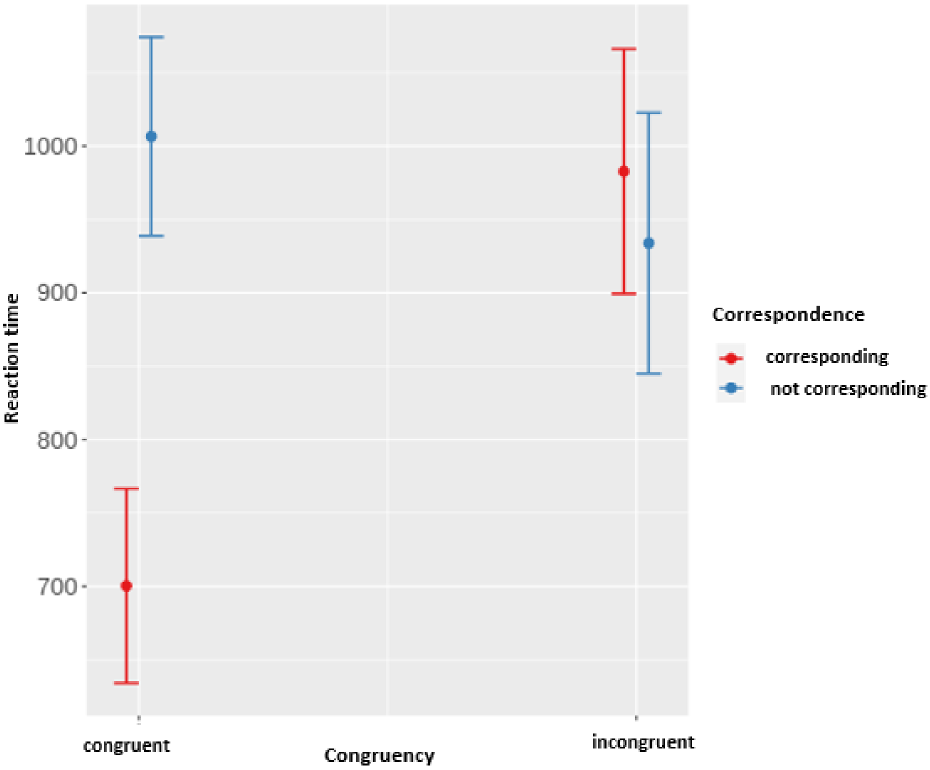


Figure 3.3. The effect of congruency and correspondence on reaction time in the Stroop task.

3.4.3 The neural correlates of cognitive control in the Stroop task

To test in which regions and frequency bands cognitive control processes can be observed, we conducted a whole brain analysis contrasting congruent/corresponding trials against congruent/non-corresponding trials, using EEG data from all electrodes, time- and spectral points throughout the whole decision period (see *Method* for details). The analysis revealed that there was a significant increase in power in several midfrontal channels including Fz, F2, FCz, FC2, in the theta band (4-8hz) starting at around 500 ms after the onset of the lower word (see Figure 3.4). Previous studies have consistently reported that response conflict modulates EEG activity in these channels and in this frequency range (Cavanagh et al. 2012; Cohen and Cavanagh 2011; Cohen et al. 2008; Nigbur et al. 2012; Wang et al. 2005;

Womelsdorf et al. 2010). The timing of our findings also aligns well with previous studies using time-frequency analysis that found effects of cognitive control appear about 470-1000ms after stimulus onset (Hanslmayr et al., 2008, Kovacevic et al., 2012, Tang et al., 2013). Consequently, it can be assumed that this difference in power in the theta band reflects cognitive control processes related to the processing of cognitive conflict. To test the robustness of these findings the analysis was repeated for the congruency effect (contrasting all congruent vs incongruent trials), revealing similar results (see Appendix 3.2).

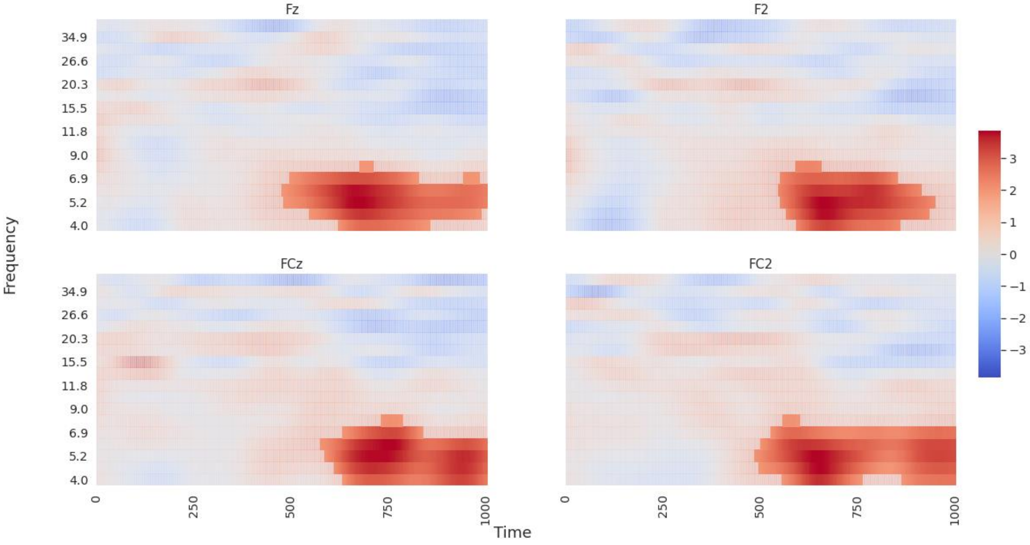


Figure 3.4. Significant clusters in the midfrontal channels can be observed in theta band (4-8 Hz) between 500-1000ms after stimulus onset. The clusters plotted in solid colors are significant at p<0.05 (corrected for multiple comparisons using a cluster-based permutation test N_{perm}=1000 within channels). The transparent colors are insignificant. The legend represents t-values.

3.4.4 Conflict processing when exposed to the opportunity to cheat

Using a mask consisting of the channels (Fz, F2, FCz, FC2) and frequencies (4-8Hz) as derived from the analysis of the Stroop task, we then contrasted cheatable against non-cheatable trials (see *Method* for details). The analysis revealed that there were significant clusters in all of the channels in the theta band in the first 500ms of the decision phase (see Figure 3.5), suggesting that conflict associated with the temptation to cheat is detected early on in the decision phase.

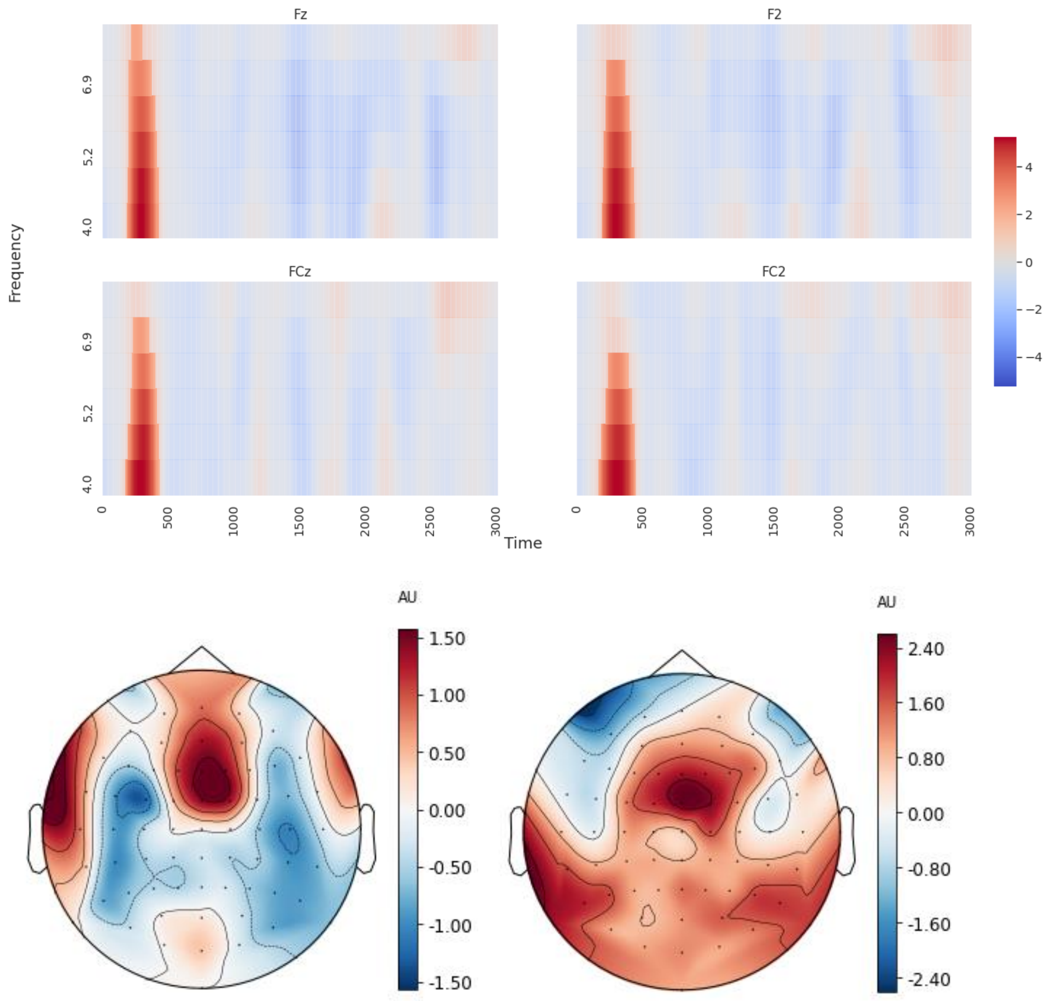


Figure 3.5. Top*:* Significant clusters in the midfrontal channels can be observed in theta band (4-8 Hz) in the first half second of the decision period. The clusters plotted in solid colors are significant at p<0.05 (corrected for multiple comparisons using a cluster-based permutation test N_{perm}=1000, corrected for multiple comparisons across channels using FDR at p<0.05) the transparent colors are insignificant. Bottom: Topoplots of the first half second of the Spot the difference task (left) and the last 250 ms of the Stroop task (right) in the theta band.

3.4.5 Power in theta Band is associated with trial-by-trial cheating

Entering the power in the theta range within the first 500ms for each of the channels of interest together with participants' cheatcount into a multilevel lasso model, the lasso model revealed that the cheatcount (b =1.61, SE = 0.07, p < 0.001) and the interaction effect between the cheatcount and theta power on Fz (b =-0.15, SE = 0.07, p < 0.05) was significantly associated with trial level cheating. Scrutinizing the plot depicting the interaction effect (see Figure 3.6), we see that for participants who cheat a lot (light blue lines), higher power on Fz was associated with lower probabilities of cheating, whereas for more honest participants (dark blue lines), higher theta power on Fz is associated with higher probability of cheating. These findings suggest that the effect of mid-frontal theta on cheating is contingent on whether a given participant is more inclined to be honest or dishonest. We also conducted robustness checks for these results, where we averaged over all channels (Fz, F2, FCz, FC2) and over the central (Fz, FCz) and right (F2, FC2) electrodes resulting in similar results (see Appendix 3.3).

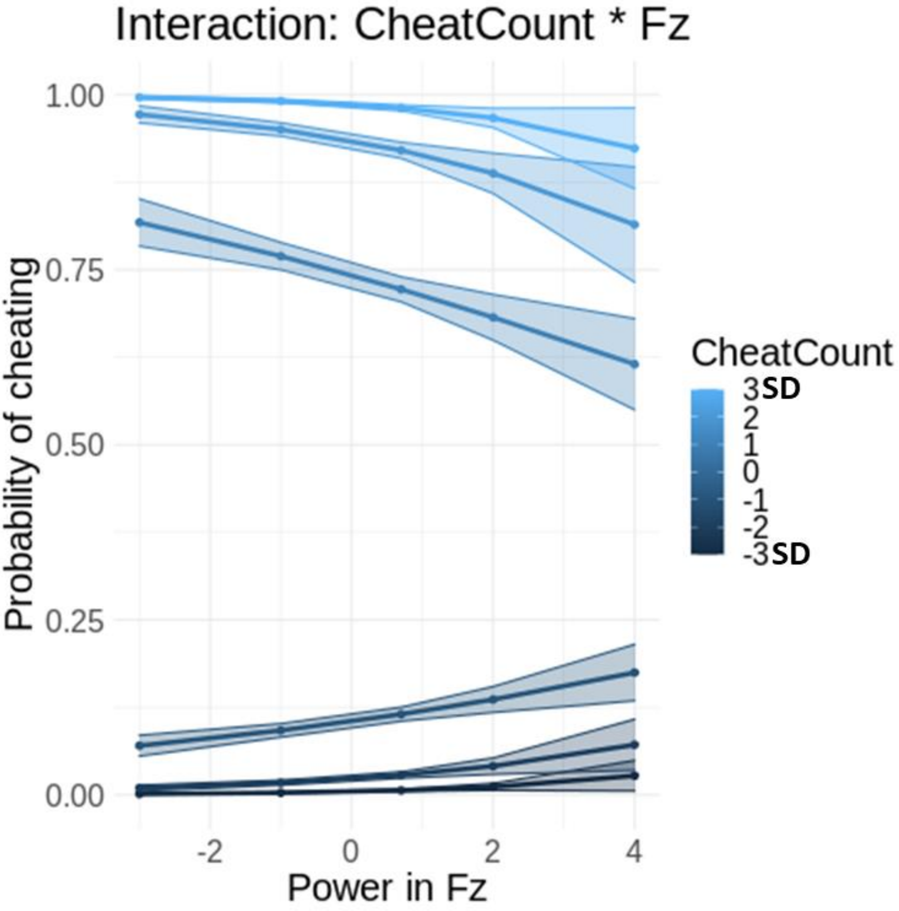


Figure 3.6. Interaction effect between cheatcount and theta in Fz in estimating the probability of cheating. The lines that are shown are the fitted values for participants 3 SD (lightest blue), 2 SD (light blue), and 1 SD (blue) above the mean of the cheat count and participants 1 SD (dark blue), 2 SD (darker blue), and 3 SD (black) below the mean of the cheatcount.

3.5 Discussion

In the current study we replicated previous findings (Speer, Smidts & Boksem, 2020) that the function of cognitive control in (dis)honest decisions depends on an individual's moral default. Specifically, combining EEG with the Stroop task as a localizer we identified the neural signature of cognitive control in the theta band (4-8) in midfrontal channels (Fz, F2, FCz, FC2). Subsequently, we observe higher power in these channels and this frequency range in the first 500 ms of the decision-making phase when participants are exposed to the opportunity to cheat as compared to the absence of this temptation. Finally, we find that higher theta power in these channels in the first 500ms is associated with a higher probability of cheating for honest participants, whereas it increased the probability of honesty for cheaters.

Our findings align well with the literature on the electrophysiological signature of cognitive control, which has consistently found that cognitive control demand modulates power in the theta range over the midfrontal cortex (Cavanagh et al. 2012; Cohen and Cavanagh 2011; Cohen et al. 2008; Nigbur et al. 2012; Wang et al. 2005; Womelsdorf et al. 2010). Furthermore, our findings from the Stroop task are in accordance with previous research showing that cognitive control demands related to the Stroop interference can be observed around 470-1000ms after stimulus onset (presentation of both words; Hanslmayr et al., 2008, Kovacevic et al., 2012, Tang et al., 2013). Using source localization analysis, Hanslmayer and colleagues (2008) showed that a source in the anterior cingulate cortex (ACC) contributed most to this effect. This dovetails elegantly with previous fMRI research, which found that especially the Stroop task activates frontal brain areas such as the ACC and the dorsolateral prefrontal cortex (dlPFC; van Veen & Carter, 2005; Kerns et al., 2004; Zysset, Mueller, Lohmann, & von Cramon, 2001; MacDonald, Cohen, Stenger, & Carter, 2000).

Observing the same increase in power over the midfrontal channels in the theta band when participants are tempted to cheat further supports the role of cognitive control in (dis)honest decisions. It has to be noted that previous research has deduced the involvement of cognitive control in moral decisions through relating observed activations to those observed for cognitive control tasks in previous studies (Greene & Paxton, 2009; Abe & Greene, 2014) or with the help of meta-analytic evidence (Speer, Smidts & Boksem, 2020) from Neurosynth (Yarkoni et al., 2011). This approach, which relies on reverse inference, must be interpreted with caution because any given brain area may be involved in several different cognitive processes, which makes it difficult to conclude that activation observed in this area represents one specific function (Poldrack, 2006). Here, we extend prior research by providing more rigorous evidence by means of explicitly eliciting cognitive control in a separate localizer task and then demonstrating that this same neural signature can be identified in the Spot-The-Difference task when participants are exposed to the opportunity to cheat. This approach provides strong evidence that cognitive control processes are recruited when participants are tempted to cheat. The observation that power modulations in these midfrontal channels represent activation in the ACC (Hanslmayr et al., 2008), aligns well with the observed involvement of the ACC in previous research using the same task (Speer, Smidts & Boksem, 2020), and implies that these effects reflect cognitive control processes related to conflict detection and resolution (Kerns et al., 2004).

In addition, we observed that the detection of cognitive conflict occurs in the first 500ms of the decision phase, irrespective of whether participants were honest or dishonest. This suggests that moral conflict is detected rapidly early in the decision process. We also tested whether differences in cognitive control between conditions could already be detected during the visual search period. This was done to rule out the possibility that differences in cognitive control can be attributed to differences in mental effort exerted during visual search rather than moral conflict. No significant differences between conditions (cheatable vs noncheatable) were found (see Appendix 3.4).

Importantly, in this EEG study we replicate the previous results from an fMRI study providing further support for the hypothesis that cognitive control is not needed to be honest or dishonesty per se but that its function in (dis)honest decision making is contingent on a person's moral default. This replication is of significance, since recently the reproducibility of many findings in (cognitive) neuroscience have been questioned (e.g. Ioannidis, 2005; Button, et al., 2013; Botvinik-Nezer et al., 2020). Concerns about the validity of insights gained from neuroimaging studies, and psychology and neuroscience more generally, may stem from substantial variability in analysis pipelines (Carp, 2012, Aurich et al., 2015, Botvinik-Nezer et al., 2020), low power (Button et al., 2013) and "p-hacking" (inappropriately controlling for the overall risk of error; Simonsohn et al., 2014, Simmons et al., 2011). In addition, errors in code or in data manipulation are easy to make (Donoho et al., 2009) and hard to detect post-hoc in neuroimaging analysis, especially if data and code are not made available online publicly (Nichols et al., 2017). To avoid building further research on fragile ground, replications are necessary to validate novel insights and hypotheses (Barch & Yarkoni, 2013; Botvinik-Nezer et al., 2020). In this study we succeed at replicating the interaction effect of cognitive control and individual differences in the tendency for dishonesty on trial-by-trial cheating behavior, across modalities and samples. Specifically, by means of using EEG instead of fMRI, evoking cognitive control by means of a well validated localizer, and using a different sample, we provide further evidence to validate the notion that the effect of cognitive control on (dis)honesty depends on an individual's moral default. What is more, the data and code is publicly available in order to allow other labs to reproduce our findings or facilitate replication in new samples.

While this study contributes by replicating and extending previous findings, some limitations are worth addressing. Here, we focused mainly on more rigorously investigating the effect of cognitive control on dishonesty decisions by applying a localizer specifically designed to investigate the neural mechanisms underlying conflict detection and response inhibition. In a Chapter 2 it was revealed that cognitive control helps to reconcile the conflict between the anticipation of reward and the maintenance of a positive self-concept. Therefore, future studies may benefit from also designing localizer tasks to elicit the neural processes underlying the anticipation of reward and self-referential thinking, to more rigorously explore the interaction between these processes.

Furthermore, although neuroimaging can provide insights into which neural processes are *associated* with (dis)honest decisions it cannot determine whether there is a *causal* relationship. In order to test the causality of the interaction effect of cognitive control on dishonesty, methods such as transcranial direct current stimulation or transcranial magnetic stimulation could be used to directly manipulate activity in the cognitive control network in honest participants and cheaters to test whether the interaction effect can be causally induced. Similarly, the causal relationship between moral default, cognitive control and cheating could be examined with the help of behavioral interventions to induce cognitive load or evoke acute stress test (Wood et al., 1984), which have been found to reduce cognitive control capacity (Shiv & Fedorikhin, 1999; Schwabe & Wolf, 2009; Yu, 2016).

To conclude, the current study shows that the role of cognitive control does not enable cheating or honesty per se but is contingent on a person's moral default. By using a well-established localizer task for cognitive control, we provide further validation of the differential effect of cognitive control on (dis)honesty. Moreover, the evidence obtained in this study suggests that the interaction effect between moral default and cognitive control on (dis)honesty is not driven by heterogeneity in latency of detection of moral conflict. Future studies may extend this work by causally manipulating cognitive control while exposing participants to the opportunity to cheat.

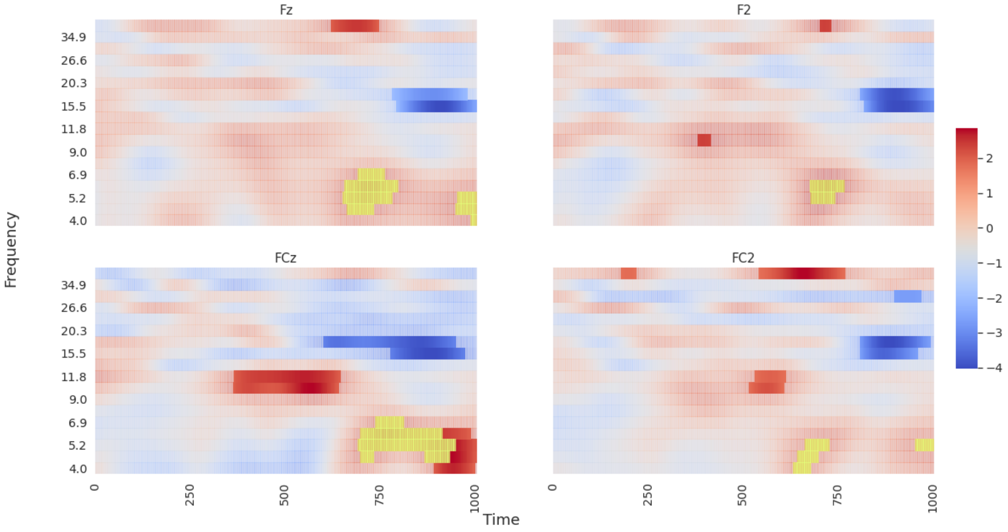
Appendix Chapter 3

Appendix 3.1: Multicollinearity between regressors for the model estimating trial-level cheating

To test whether there is multicollinearity between the regressors derived from the channels associated with cognitive control from the first half second of the decision, we correlated the average activation in each of these channels with each other. No significant correlations were found.

Appendix 3.2: Robustness check neural signature of incongruency effects from the Stroop task

To assess the robustness of the neural effects of the Stroop task, we also explored the effect of congruency on power across channels, time and frequencies. In this analysis, we contrasted incongruent-corresponding trials with congruent-corresponding trials, to isolate the effect of congruency. The analysis revealed very similar effects (see Figure S2).



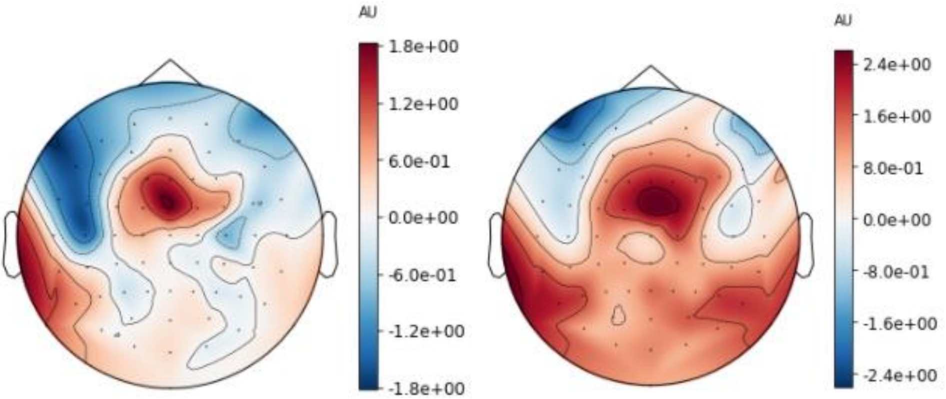


Figure S2. Top: Significant clusters in the midfrontal channels can be observed in theta band (4-8 Hz) around 600ms after stimulus onset. The clusters plotted in solid colors are significant at $p<0.05$ (uncorrected) the transparent colors are insignificant. The yellow clusters indicate overlap with the results from the contrast reported in the main text (congruent-not corresponding > congruent-corresponding). Bottom: Topoplots of the last 250 ms for the effect of congruency (incongruent-corresponding > congruent-corresponding; left) and the effect of correspondence (congruent-not corresponding > congruent-corresponding; right) of the Stroop task (right) in the theta band.

Appendix 3.3: Robustness check trial-by-trial analysis

Instead of applying a regularization, in form of a lasso model, to reduce multicollinearity between channel, we tested the robustness of the findings reported in the main text by averaging over the channels selected by the Stroop task. As a first step, we averaged over all channels resulting in one trial-level regressor summarizing all electrodes and the cheatcount as a subject-level regressor. Entering these two regressors into a multilevel model revealed that there was a significant main effect of cheatcount (b =1.60, SE = 0.07, $p < 0.001$) and a significant interaction effect between the average of all electrodes and the cheatcount (b =-0.26, SE = 0.10, $p < 0.05$).

To increase spatial specificity, we also averaged over the central (Fz & FCz) and right (F2 & FC2) electrodes and subsequently both as trial-level regressors along with the cheatcount as subject-level regressor into the multilevel model. Here, the model revealed that there was a significant main effect of cheatcount (b =1.61, SE = 0.07, $p < 0.001$) and an effect of the central electrodes (Fz & FCz; b =0.20, SE = 0.09, $p < 0.05$, see Figure S4). In concert, these robustness checks convey that our results are robust across different ways of dealing with multicollinearity and that particularly the central electrodes appear to drive the interaction effect between cheatcount and power in the theta band.

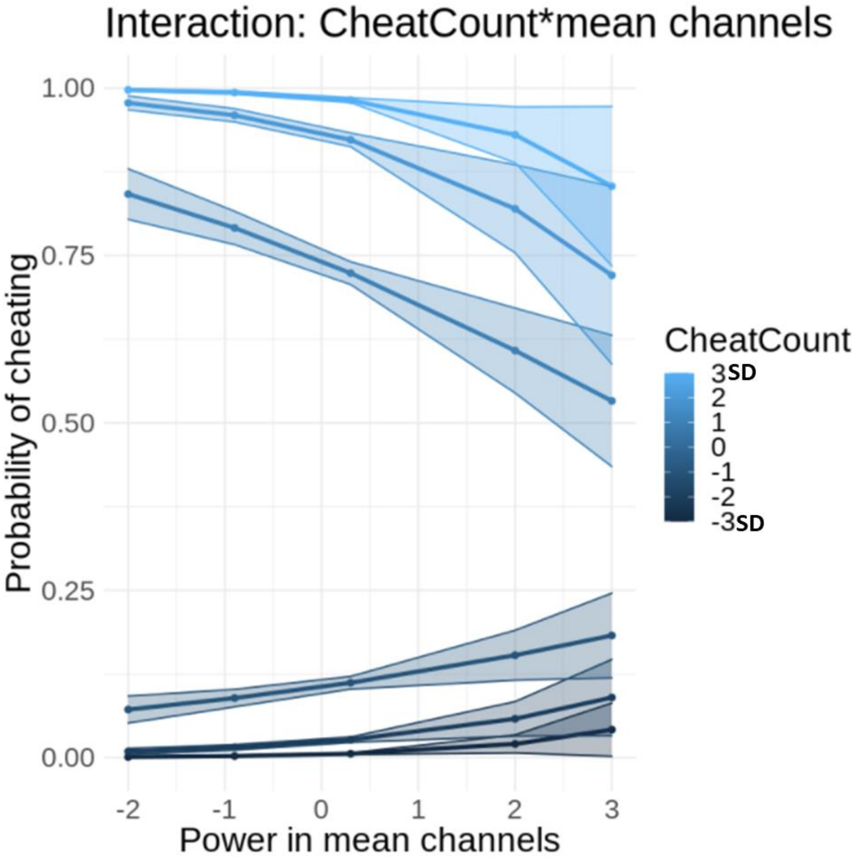


Figure S4. Interaction effect between cheatcount and the central channels in estimating the probability of cheating. The lines that are shown are the fitted values for participants 3 SD (lightest blue), 2 SD (light blue), and 1 SD (blue) above the mean of the cheat count and participants 1 SD (dark blue), 2 SD (darker blue), and 3 SD (black) below the mean of the cheatcount.

Appendix 3.4: Testing for early conflict detection during visual search

To assess whether participants might already become aware of the moral conflict during the visual search (see Figure 1: 4th screen), we tested whether there were differences in neural activation during the visual search phase between cheatable trials and non-cheatable trials. To this end, the EEG data from the Spot-The-Difference-Task was segmented into 6 second epochs, time-locked to the onset of the decision phase but extending backwards to the onset of the image pair (thus -6s to 0s before the onset of the decision phase). The epochs were baseline corrected using the last second of the fixation period preceding the presentation of the image pair, which occurred seven to six seconds before the decisions phase. Other than that, the same preprocessing steps as before were performed. For the analysis, as was done previously, we used a mask, consisting of the significant channels and frequencies from the Stroop task, to test whether at any point in time during the visual search phase, cheatable trials differ from non-cheatable trials. Specifically, we computed the averaged time-frequency maps of cheatable trials and non-cheatable trials for each subject. We then applied the mask from the Stroop task (i.e. the spectral points within the channels found to be significant in the analysis of the Stroop effect) to the resulting average maps. For group-level analysis we then conducted the cluster-based permutation testing procedure described above to the masked contrast maps. The analysis revealed that there were no significant differences between cheatable and non-cheatable trials during the visual search phase, which suggests that participants did not process the moral conflict before the onset of the decision phase.

Chapter 4

Different neural mechanisms underlie non-habitual honesty and non-habitual cheating[3]

4.1 Abstract

There is a long-standing debate regarding the cognitive nature of (dis)honesty: Is honesty an automatic response or does it require willpower in form of cognitive control in order to override an automatic dishonest response. In a recent study (Speer, Smidts & Boksem, 2020) we proposed a reconciliation of these opposing views by showing that activity in areas associated with cognitive control, particularly the inferior frontal gyrus (IFG), helped dishonest participants to be honest, whereas it enabled cheating for honest participants. These findings suggest that cognitive control is not needed to be honest or dishonest per se but that it depends on an individual's moral default. However, while our findings provided insights into the role of cognitive control in overriding a moral default, they did not reveal whether overriding honest default behavior (nonhabitual dishonesty) is the same as overriding dishonest default behavior (nonhabitual honesty) at the neural level. This speaks to the question as to whether cognitive control mechanisms are domain-general or may be context specific. To address this, we applied multivariate pattern analysis to compare neural patterns of nonhabitual honesty to nonhabitual dishonesty. We found that these choices are differently encoded in the IFG, suggesting that engaging cognitive control to follow the norm (that cheating is wrong) fundamentally differs from applying control to violate this norm.

[3] This chapter has been published in Frontiers in Neuroscience (Speer, Smidts & Boksem., 2021).

4.2 Introduction

In a recent study (Speer, Smidts & Boksem, 2020) we found that areas associated with cognitive control, particularly the inferior frontal gyrus (IFG), helped dishonest participants to be honest, whereas it enabled cheating for those who are generally honest. These findings suggest that honest participants needed cognitive control to overcome their inclination for being honest in order to cheat whereas cheaters had to exert control to override their greedy tendencies in order to be honest. Based on these findings we argued that cognitive control is not needed to be honest or dishonest per se but that it depends on an individual's moral default.

Our results help reconcile the long-standing debate between proponents of the Will hypothesis and the Grace hypothesis. Research supporting the Will hypothesis (Gino et al., 2011; Mead et al., 2009; Welsh & Ordonez, 2014) suggests cognitive control is needed to be honest. In direct opposition to this, a separate stream of research has accumulated evidence in favor of the Grace hypothesis (for meta-analyses, see refs. Greene & Paxton, 2009; Shalvi, Eldar & Bereby-Meyer, 2012; Suchotzki et al., 2017; Verschure et al, 2018; Caparo, 2017), advocating that cognitive control is required for dishonesty. Our findings suggested that people are distributed along a continuum, from individuals who are generally honest to cheaters. Individuals on one end of the continuum are inclined to be honest, which is associated with more self-referential thinking when given the opportunity to cheat. In contrast, individuals on the other end of the spectrum have an inclination for dishonesty, and their decisions are driven more strongly by rewards. In order to achieve a subjectively justifiable and desirable balance where one can occasionally profit from cheating but still maintain a positive self-image, people on both sides of the spectrum sometimes need to override their moral default. We show that the cognitive control network may orchestrate both honesty for people who can be considered cheaters and dishonesty for the more honesty inclined and thus provide potential reconciliation for this long-standing paradox.

In a commentary on our work, Abe (2020) astutely noted that, while our findings provide insights into the role of cognitive control, they do not reveal the exact nature of control-related activity. Although our study demonstrated that activity in the IFG is required when participants override their moral default, our analyses did not reveal whether activity in the IFG for a cheater's decision to be honest is actually identical to activity in the IFG for an honest person's decision to cheat. Therefore, the question arises whether, on the neural level, overriding the default to be honest in favor of cheating is identical to overriding the default to be dishonest in favor of honesty. Answering this question would substantially improve our understanding about the nature of the cognitive control processes that enable us to override our moral default, because it may reveal whether the IFG has access to the moral valence (overriding the 'good' or 'bad' default) of a given decision. To investigate whether neural patterns of activity associated with nonhabitual dishonesty can be distinguished from neural patterns underlying nonhabitual honesty, Abe suggested the use of multivariate pattern analysis (MVPA, Norman et al., 2006). In this brief report, we apply MVPA to compare neural patterns of nonhabitual honesty to nonhabitual dishonesty.

4.3 Methods

4.3.1 Participants

The reported analyses are based on 40 participants (30 females; age 18 to 35 y; M = 23.7, SD = 3.2) recruited from an online community for university students. All participants were right-handed with normal or corrected to normal vision, spoke English fluently, were not on any psychoactive medication influencing cognitive function, and had no record of neurological or psychiatric illness. The study was approved by the Erasmus Research Institute of Management (ERIM) internal review board and was conducted according to the Declaration of Helsinki.

4.3.2 Task and stimuli

4.3.2.1 Spot-the-Difference task

As described in Chapter 2, in the Spot-The-Difference task, participants were presented with pairs of images and were instructed that there were always three differences present between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were requested to find three differences between the images. Since reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design allowed and encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were present in the images). For details regarding this task and the stimuli refer to Chapter 2 paragraph *2.3.2.1* & *2.3.2.2.*

4.3.3. FMRI acquisition & preprocessing

The same data and preprocessing pipeline were used as in Chapter 2. For details please refer to paragraphs 2.3.4 & 2.3.5.1.

4.3.4 Statistical analyses

For each participant, we estimated a general linear model (GLM) using regressors for onsets of the decision phase for cheated trials and honest trials. The duration of the epoch for the decision phase was three seconds and the beginning of the decision phase was used as onset times. The decision phase was used as it provides all the necessary information to make the decision and is free of brain activity related to motor responses. In addition, regressors for the button presses were added. Average background, white matter and cerebrospinal fluid (CSF) signal, framewise displacement, six head motion regressors and six aCompCor regressors, all obtained from fMRIprep, were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. A smoothing kernel of 5 mm (FWHW) was applied. Linear contrasts were computed between honest and cheating decisions. Neural patterns were then extracted from the resulting t-maps (cheat > honest & honest > cheat) using the left IFG mask derived from the conjunction analysis between the Neurosynth map for cognitive control and the results from the second level analysis investigating the neural mechanisms underlying the decision to cheat in Speer, Smidts and Boksem (2020). Due to the fact that participants engaged in spontaneous, voluntary, and deliberate cheating, the ratio of dishonest and honest trials was not perfectly balanced for most of the participants. In order to account for potential statistical confounds resulting from this imbalance, we under-sampled the majority class for each participant to create a perfect balance when estimating the contrasts (Liu, Wu & Zhou, 2009).

To test whether nonhabitual honesty differs from nonhabitual dishonesty on the neural level, we conducted a classification analysis on the neural patterns in the IFG (see Figure 4.1) associated with nonhabitual honesty in cheaters and nonhabitual dishonesty in honest participants. Specifically, we used the t-maps derived from contrasting cheated decisions against honest decisions (cheat > honest) for honest participants and contrasting honest decisions against cheated decisions (honest > cheat) for cheaters as input for the classification analysis. Participants were categorized as cheaters or honest participants based on a median split (median = 10 cheated decisions). The choice of using the honest decision as baseline condition for honest participants and cheated decisions as a baseline for cheaters was motivated by the fact that these decisions represent the default behaviors for the two groups of participants respectively. We then trained and tested a logistic lasso regression classifier to decode whether a given contrast map represented the neural pattern associated with a nonhabitual decision for honesty or dishonesty, using 4-fold cross-validation (see Figure 4.1). The 4-fold cross validation refers to training the classifier on 30 participants and testing on the 10 remaining participants at each iteration. Importantly, in case the neural patterns associated with nonhabitual honesty and dishonesty are the same, the classifier should not be able to accurately categorize cheaters or honest participants based on their neural patterns.

4.4 Results

Substantial individual differences in cheating were observed (mean = 26%, median = 14%, SD = 26%): some participants cheated only once or twice (17.5% of participants), while others only missed one or two chances to cheat (5%). Participants who cheated relatively often are from now on referred to as cheaters and more honest are referred to as honest individuals.

The multivariate classification analysis revealed that we can indeed successfully classify whether a participant engaged in either nonhabitual honest or nonhabitual dishonest behavior (Accuracy = 78%, p<0.01, Npermute=1000; Cross-validation accuracy scores per fold: 70%, 70%, 80%, 90%; see Figure 4.2). That is, neural patterns of overriding habitual honesty in favour of cheating differ significantly from the patterns underlying a cheater's decision to be honest. The distribution of activity across voxels within the IFG (Figure 4.3) suggest that inhibiting habitual honesty (nonhabitual dishonesty) is associated within relatively high activity in voxels situated more ventrally in the IFG, while overriding habitual dishonesty (nonhabitual honesty) is associated with relatively high activity in more dorsal voxels within the IFG.

In order to make sure that the successful classification can be attributed to differences between nonhabitual honesty and nonhabitual dishonesty, we ruled out several alternative explanations. First, to test whether the classification accuracy was not just driven by individual differences in honesty irrespective of choice, we trained and tested a logistic lasso regression model to classify cheaters and honest participants using the same contrast for all participants. For both cases, using the honest>cheat and the cheat>honest contrast, the classification accuracy was not significant, indicating that the classification accuracy observed in the previous analysis cannot be attributed to individual differences in honesty alone. Secondly, we explored whether the decoding accuracy mostly resulted from simply using a different contrast (different choices), irrespective of participants' moral default (individual differences in honesty). To rule this out, we randomly assigned half of the participants to one of the contrasts (honest>cheat) and the other half to the other contrast (cheat>honest) and then used the logistic lasso regression model to decode which participants was assigned to which contrast. The whole procedure was repeated 1000 times. This approach was adopted as it would ensure that the mean level of honesty in the two groups would be the same and thus render the classification of contrasts independent of the participants' moral default. This analysis resulted in insignificant classification accuracy, which shows that the classification accuracy of the main analysis is not merely driven by the differences in the contrast used.

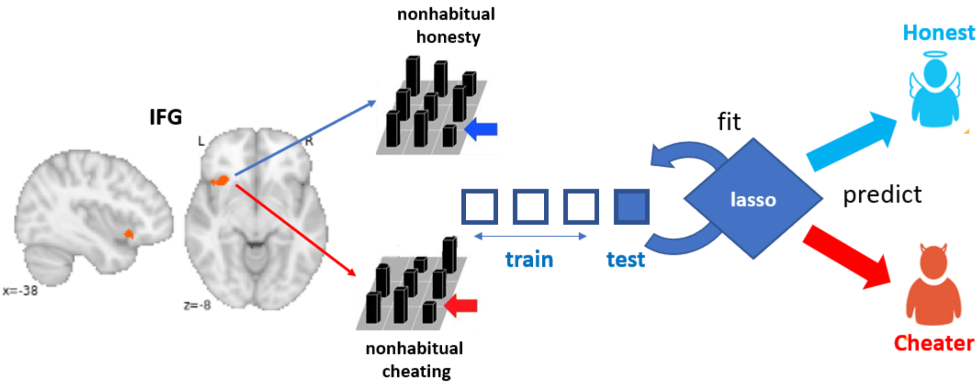


Figure 4.1. For honest participants the neural pattern in the IFG was derived from the contrast cheat>honest and for cheaters the neural pattern was obtained from the contrast honest>cheat. These patterns were then fed to a logistic lasso regression classifier, which was trained and tested using 4-fold cross validation to decode whether a given pattern belonged to a cheater or to an honest participant.

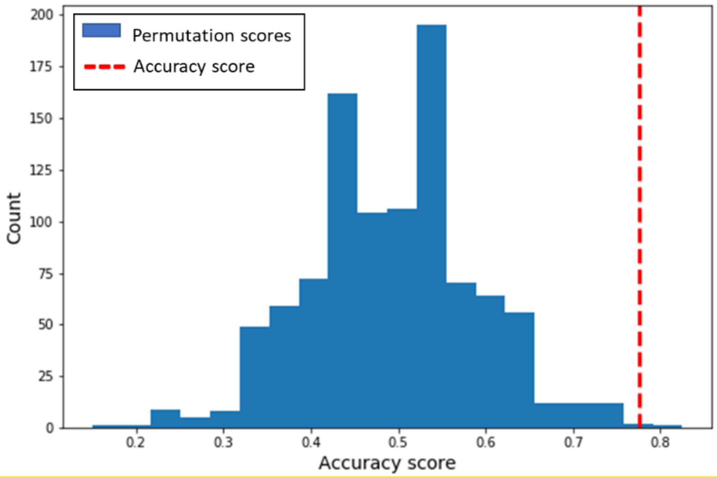


Figure 4.2. Distribution of classification performance. The blue bars indicate the predictions from the permutation test (N=1000). The red dashed line represents the empirical accuracy score of the model (78%).

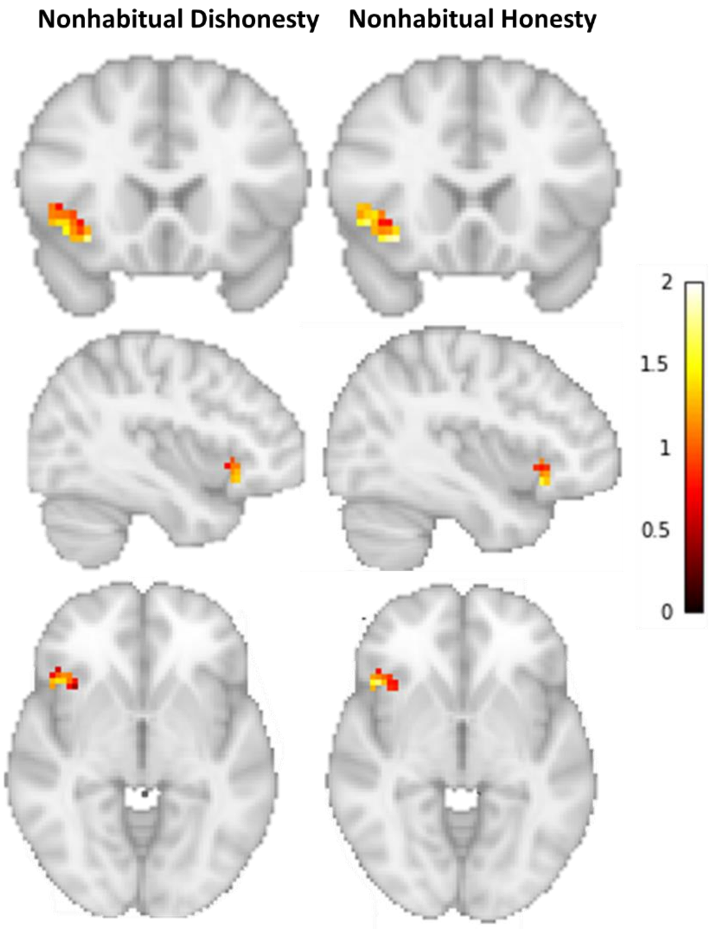


Figure 4.3. Average activation patterns for nonhabitual honesty and nonhabitual dishonesty in the left IFG.

4.5 Discussion

Using the spot-the-difference task to study trial-by-trial cheating behavior we previously found (Speer, Smidts & Boksem, 2020) that the effect of cognitive control depends on a participants' inclination to be honest or dishonest, in other words, on their moral default. The follow-up, analysis presented here revealed that, whereas the level of average activation across all voxels in the IFG is the same for honest participants and cheaters when engaging in a nonhabitual (dis)honest decision, the information encoded in the distributed pattern across voxels differs. Specifically, our results hint at differential involvement of dorsal and ventral IFG in nonhabitual honesty and dishonesty, respectively.

These results provide deeper insights into the nature of the cognitive control processes that enable us to override our moral default, as they may suggest that the IFG has access to the moral significance of the decision at hand. Engaging cognitive control to follow the norm that cheating is wrong appears to be represented differently in the IFG as compared to applying control to violate this norm. This may suggest that, even though individuals have idiosyncratic default responses in morally ambiguous situations, the underlying moral norm, that cheating is wrong, may nonetheless be universal across individuals. Alternatively, the neural patterns may differ because the specific cognitive processes that need to be inhibited differ. For dishonest participants, the motivation to obtain (monetary) reward needs to be inhibited, while for honest participants the motivation to maintain a positive self-concept needs to be inhibited.

In the field of cognitive neuroscience there has been an enduring interest to refine the constructs of cognitive control (Baddeley et al., 1996; Cuthbert & Insel, 2013; Diamon, 2013). Particularly, it has been a long-standing challenge to determine which cognitive control processes should be considered domain-general, thus commonly engaged by different types of tasks, and which cognitive control processes are domain or even context specific. The current results provide some initial answers to this question in the context of cognitive control in the form of response inhibition applied to moral decisions. Our findings suggest that, at least in the context of moral choice, inhibiting prepotent responses to cheat are indeed different from inhibiting a default of behaving honestly.

Chapter 5

Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest[4]

5.1 Abstract

Many of the economically most costly forms of unethical behavior such as tax evasion, stock manipulations and music piracy relate to acts of (dis)honesty. Measurement of the determinants of socially undesirable behaviors, such as dishonesty, are complicated and obscured by social desirability biases. To address this issue, we used connectome-based predictive modelling (CPM) on resting state functional connectivity patterns in combination with a novel task, which measures voluntary cheating inconspicuously, to investigate whether task-independent neural patterns within the brain at rest can be used to predict (dis)honesty at a later point in time. The combination of resting state connectivity and an inconspicuous cheating task granted direct access to the neurocognitive determinants of (dis)honesty. Our analyses revealed that particularly functional connectivity between brain networks linked to self-referential thinking (vmPFC, temporal poles, and PCC), and the caudate nucleus, strongly associated with reward processing, can reliably and accurately predict, in an independent sample, whether a participant is honest or a cheater. Participants who cheated the most also scored highest on several impulsivity measures, which highlights the ecological validity of our task. Notably, when comparing neural and self-report measures, the neural measures were found to be significantly better at predicting cheating behavior.

[4] A revised version of this chapter is under second-round review at *NeuroImage.*

5.2 Introduction

Cheating and dishonesty, manifested in diverse behaviors such as financial fraud, scientific misconduct and software piracy, is ubiquitous and represents one of the economically most costly forms of unethical behavior. However, it is evident that not everybody is a cheater: there are substantial individual differences in (dis)honesty, ranging from people who embody integrity and remain honest even when it comes at their own cost, such as Abraham ‘honest Abe’ Lincoln, to individuals such as Jordan ‘Wolf of Wallstreet’ Belfort, who greedily engaged in fraudulent stock market manipulations that led to investor losses of more than 200 million US dollars.

Both such anecdotal evidence as well as scientific studies have shown that, when given the opportunity, individuals indeed differ considerably in how frequently they cheat (Gino et al., 2012, 2014; Speer, Smidts & Boksem, 2020). For example, research in social and personality psychology has found that greedy people find a variety of moral transgressions more acceptable and engage in such unethical behaviors more frequently than less greedy people do (Seuntjens et al., 2019). Yet, people do not only care about their own financial gains, which is evident from the omnipresence of prosocial behaviors such as altruism, reciprocity and honesty. When exposed to an opportunity to cheat, the way we view ourselves, our self-concept (Aronson 1969; Baumeister 1998; Bem 1972), may motivate us to refrain from cheating. People highly value integrity and honesty in others and also have strong convictions of their own moral standards (Dhar & Wertenbroch, 2012). As a result, individuals often tend to uphold their self-concept even if it means forgoing financial gains (Mazar, Amir, & Ariely, 2008).

Whereas psychological research has suggested several personality variables to be associated with individual differences in unethical behavior and propensity to cheat (Anderman et al., 2009; Tang & Chen, 2008; Gino et al., 2012; Seuntjens et al., 2019), so far the underlying cognitive operations that determine these individual differences in the propensity to cheat have remained elusive, which renders it difficult to predict who will cheat. This may in part be due to the fact that measurement of the psychological determinants of socially undesirable behaviors, such as dishonesty, obtained by self-report methods, may suffer from social desirability bias. Participants may not want to admit or reveal that they are dishonest, impulsive or greedy, and consequently try to conceal their undesirable behavioral tendencies. Such a bias may obscure the psychological and cognitive mechanisms that determine whether someone will cheat or not and consequently make it hard to determine how cheaters may differ from more honest individuals. To address this issue, neuroimaging techniques such as functional magnetic resonance imaging (fMRI), can be combined with behavioral paradigms measuring (dishonesty) to obtain direct and relatively uncontaminated access to the underlying psychological processes.

Neuroimaging studies exploiting this more direct access to the underlying psychological processes tend to support the notion of two opposing forces of greed and upholding a moral self-concept that steer us towards (dis)honesty. In a recent fMRI study (Speer, Smidts & Boksem, 2020), it was found that activity in the nucleus accumbens (Nacc), associated with reward anticipation and greed (Ballard & Knutson, 2009; Knutson, Adams, Fong, & Hommer, 2001; Abe & Greene, 2014), promotes cheating, particularly for individuals who tend to cheat a lot, whereas a network consisting of Posterior Cingulate Cortex (PCC), bilateral Temporoparietal Junctions (TPJ) and Medial Prefrontal Cortex (MPFC), associated with self-referential thinking processes (Gusnard et al, 2001; Meffert et al., 2013; Van Buuren et al., 2010), promotes honesty, particularly in individuals who are generally honest. In addition, numerous studies have proposed that cognitive control is needed to resolve this tension between reward and self-concept (Abe & Greene, 2014; Gino, Schweitzer, Mead, & Ariely, 2011; Greene & Paxton, 2009; Maréchal, Cohn, Ugazio, & Ruff, 2017; Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009). In accordance with these findings, the study by Speer et al. (2020) revealed that activity in cognitive control regions, namely the anterior cingulate cortex (ACC) and the inferior frontal gyrus (IFG; Swick et al., 2008; Carter & Van Veen, 2007) were recruited to resolve the conflict between self-interest and self-image.

Importantly, however, all these studies were measuring neural activity while participants were making cheating decisions which render these findings relatively context dependent. The neural activity associated with (dis)honesty may therefore be specific to the given situation and may not generalize to other situations in which there is an opportunity to cheat. A promising approach to address this issue and explore more stable neural markers of (dis)honesty is to identify neural correlates of dishonesty using resting state functional magnetic resonance imaging (rsfMRI).

Variability in whole brain resting state functional connectivity is substantial across individuals, and research has shown that an individual's functional connectome is robust and reliable across sessions over time and can even be reproduced between task and rest (Cao et al., 2014; Zuo and Xing, 2014; Finn et al., 2015). For instance, Finn and colleagues found that it is possible to identify an individual from a large group of subjects with high accuracy (94-98%) solely based on her connectivity matrix across resting state (and also task) sessions on separate days. RsfMRI has previously been employed to successfully link functional connectivity to individual differences in self-reported personality (Nostro et al., 2018; Cai et al., 2020) and impulsivity in intertemporal choice (Li et al., 2013). In the present study we use rsfMRI as it may grant us unique direct access to the neurocognitive determinants of undesirable social behavior (dishonesty).

The current study examined whether the resting functional connectome could predict an individual's propensity to cheat. In order to do so, we used a sample of 99 participants, consisting of four independent sub-samples, who completed a resting state scan and a task (at a different time) in which participants could cheat repeatedly, deliberately and voluntarily without suspicion of the real purpose of the task (the 'Spot-The-Difference Task': see Gai & Puntoni, 2021; Speer, Smidts & Boksem, 2020). We employed connectome-based predictive modeling (CPM) to investigate whether (dis)honesty in this task can reliably be predicted from an individual's unique pattern of functional connectivity.

CPM has recently been developed to predict individual differences in both cognitive abilities and personality traits from patterns of whole-brain functional connectivity (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017). Importantly, the CPM approach differs from methods implemented in previous rsfMRI studies as it uses out-of-sample prediction, instead of just establishing correlational relationships between brain and behavior. This out-of-sample prediction leads to a more conservative inference of the brain-behavior relationship than does correlation, which increases the probability of successful replication in the future (Shen et al., 2017). More practically, establishing predictive models with high accuracy is required to convert insights from neuroimaging to applications in practice (Gabrieli, Ghosh & Whitfiled-Gabrieli, 2015). Since a complex construct such as (dis)honesty is a product of several interacting psychological processes (e.g., cognitive control, reward, self-referential thinking), it is reasonable to assume that its neural manifestation is most likely similarly sophisticated and distributed across several networks throughout the whole brain. In light of previous (task-based) research, we hypothesize that particularly functional connectivity between regions in the self-referential thinking network, reward network and cognitive control network will be predictive of (dis)honesty. In this study we train a CPM model on a training set of two distinct samples and then evaluate the predictive performance on two independent samples using regions from these three networks (self-referential thinking, reward and cognitive control). Testing the predictive importance on independent samples provides the most accurate estimate of the generalizability of the models.

To provide a benchmark for out-of-sample prediction of (dis)honesty we also acquired several self-reported personality questionnaires measuring traits previously associated with cheating and (dis)honesty, including impulsivity (Anderman et al., 2009), creativity (Gino et al., 2012), greed (Seuntjens et al., 2019) and manipulativeness (Tang & Chen, 2008) amongst others. These self-report measures also assisted in evaluating the ecological validity of the Spot-The-Difference task.

5.3 Methods

5.3.1 Participants

The reported analyses are based on 99 participants (65 females; 24 nationalities; age 18 - 43 years, *M* = 24.3, *SD* = 4.30) from four separate studies. The data were collected in three different scanners. We included data from four studies to increase sample size and diversity. Specifically, we collected data from two big and two small samples. This was done to obtain two independent big samples for the training and two independent samples for the test set. Having two independent samples in the training set increases the generalizability of the models as they are trained on a more diverse sample from different scanners and are less likely to learn idiosyncratic patterns related to a specific sample.The big samples were chosen as training set as they allow for the biggest number of observations to be used for feature selection and training of the models, which is necessary to ensure generalizability of the model and improve prediction accuracy (Sima et al., 2005; 2006).

The first sample of participants consisted of students (N = 40, 29 females; age 18 - 35 years, *M* = 23.8, *SD* = 3.25) from now on referred to as Study 1. The second sample consisted of a general population sample from a different city and neural data was collected in a different scanner (N=41, 23 females; age 18 - 43 years, *M* = 24.8, *SD* = 5.4). The third sample were students at a different university in a different scanner (N=9, 7 females; age 19 - 24 years, *M* = 21.6, *SD* = 1.42). The fourth sample was scanned at the same scanner as Study 1 but consisted of different students (N=9, 6 females; age 22 - 30 years, *M* = 26.9, *SD* = 2.6). Samples 1 and 2 will form the training set and samples 3 and 4 form the test set. Due to the fact that only 91 out of the 99 subjects completed all questionnaires (see below) all the analyses reported in the Results section focus on these 91 subjects, apart from the robustness checks that were done on all 99 subjects (see Results section).

We also assessed whether differences in demographics existed between the samples. Here we focused on the 91 subjects, since these were the ones that were used in all the analyses reported in the main text. Testing the differences in age between the four samples (on the 91 subjects) no significant differences were found (ANOVA, $M_1 = 23.7$, $M_2 = 24.9$, $M_3 = 21.9$, $M_4 = 26.8$, $F = 2.23$, $p = 0.09$). Testing for differences in Gender across the four samples (on 91 subjects) again no significant differences were found ($X^2 = 3.91$, p = 0.27). When comparing the training set (Study 1 & Study 2 using 91 subjects) to the test set (Study 3 and 4 using 91 subjects) again no significant differences were found for age ($t = 0.04$, $p = 0.97$) and gender ($X^2 = 0.31$, $p = 0.58$).

5.3.2 Experimental procedure

All participants were first informed about and checked on the safety requirements for MRI scanning, and they signed the informed consent form. They then completed the resting state scan (8 min). The resting state data was collected while participants were lying in the scanner looking at a fixation cross and instructed to let their mind wander; they did not perform any task during this period. All neural data discussed in this paper is based on the data from this 8-minute rest period.

Before the behavioral tasks started, participants were introduced to the cover story and completed a number of practice trials. Then, the participants completed two experimental tasks: a simple visual search task (5 min) followed by the Spot-The-Difference task which took approximately 40 minutes. See Task and Stimuli below for details about these tasks. For study 1, participants completed the tasks inside the MRI scanner, but only the behavioral data was used in the present study. The task-based neural data recorded during the Spot-The-Difference task (see below) was reported in a previous study (Speer, Smidts & Boksem, 2020). For all other studies the Spot-The-Difference task was performed outside the fMRI scanner. After completing the two tasks, participants filled-in a short questionnaire including questions about their thoughts on the purpose of the task.

After completion of the experimental session, participants received an email with a link to a Qualtrics questionnaire including measures for impulsivity, greed, creativity, manipulativeness and sensitivity to different moral foundations (explained below), which they were allowed to fill out at home.

First, in the test battery, we included four impulsivity scales: a) the Brief Sensation Seeking Scale (BSSS; Hoyle, Stephenson, Palmgreen, Lorch & Donohew, 2002), b) the BisBas scale to assess dispositional inhibition and approach behavior (Carver & White, 1994), c) the short version of the UPPS-P Impulsive Behavior scale (Cyders, Littlefield, Coffey, & Karyadi, 2014) and d) a Risk seeking scale implementing a standard risk preference elicitation method where one can choose between a certain amount of money and a risky gamble. To quantify risk preference, participants were presented with a sequence of binary choices between a certain amount of money for sure or a gamble with a fifty percent chance of winning 30€ for sure and a fifty percent of not winning anything. Whereas the gamble remains the same for each question the amount gained for sure increases at each step. A person's risk preference could thus be established by identifying the amount of money for sure at which the person switches from the gamble to certain payout (example item: "Would you prefer 13€ for sure or 0€ or 30€ with a 50-50% chance"). These impulsivity scales were selected because dishonesty and cheating has been linked to impulsivity as a personality trait (Anderman et al., 2009).

Second, we measured how an individual's sensitivity to different moral foundations, namely Care vs. Harm, Fairness vs. Cheating, Loyalty vs. Betrayal, Authority vs. Subversion, Sanctity vs. Degradation and Liberty vs. Oppression, may influence cheating by including two such measures: a) the Moral Foundations Questionnaire (MFQ; Graham et al., 2011) and b) the Moral Foundations Vignettes (Clifford, Iyengar, Cabeza & Sinnott-Armstrong, 2015).

Third, as greed is assumed to drive cheating behavior (Seuntjens et al., 2019), the Dispositional Greed Scale was added (Seuntjens et al., 2015). Fourth, an individual's creativity was measured by means of three scales: a) the Remote Associates Test (Mednick, 1968), b) Gough's Creative Personality Scale (Gough, 1979), and c) Hovecar's Creative Behavior Inventory (Hovecar, 1979), as it has been found that more creative people tend to cheat more (Gino et al., 2012). Fifth, the MACH-IV test (Christie & Geis, 1970) to measure manipulativeness was added as Machiavellianism has also often been associated with unethical behavior (Tang & Chen, 2008). Participants were informed that they would only receive their payment once they completed the questionnaires.

5.3.3 Task and Stimuli

5.3.3.1 Spot-The-Difference task

As described in Chapter 2, in the Spot-The-Difference task, participants were presented with pairs of images and were instructed that there were always three differences present between the image pairs. Differences consisted of objects that were added to or removed from an image, or objects that differed in color between images. However, images could actually contain one, two, or three differences. Participants were requested to find three differences between the images. Since reward (see below) was contingent on participants reporting that they had found all three differences, without having to point them out, this design allowed and encouraged cheating behavior (i.e., reporting having found all three, even when objectively fewer than three differences were present in the images). For details regarding this task and the stimuli refer to Chapter 2 paragraph *2.3.2.1* & *2.3.2.2.*

5.3.4 fMRI acquisition

For Study 1 and 4, the functional magnetic resonance images were collected using a 3T Siemens Verio MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in descending interleaved order (3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 75°). TE was 30ms and TR was 2030ms. A T1-weighted image was acquired for anatomical reference (1.0 × 0.5 × 0.5 mm resolution, 192 sagittal slices, flip angle = 9°, TE = 2.26ms, TR = 1900ms)

For Study 2, the functional magnetic resonance images were collected using a 3T MRI system (General Electric). Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in ascending interleaved order (3mm slice thickness, 3.5 mm slice gap, 3 × 3 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 75°, TE = 30ms, TR = 2030ms). A T1- weighted image was acquired for anatomical reference (1.0 × 1.0 × 1.0 mm resolution, 160 sagittal slices, TE = 2.35ms, TR = 7.21ms).

For Study 3, the functional magnetic resonance images were collected using a 3T Phillips Achieva MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in descending interleaved order (3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 76°). TE was 27ms and TR was 2000ms. A T1-weighted scan was acquired using 3D fast field echo (TR: 82 ms, TE: 38 ms, flip angle: 8°, FOV: 240 × 188 mm, in-plane resolution 240 × 188, 220 slices acquired using single-shot ascending slice order and a voxel size of 1.0 × 1.0 × 1.0 mm). For all studies functional scans were acquired for 8 minutes.

5.3.5 Preprocessing

Data was preprocessed using the standard pipeline of the CONN toolbox (https://www.nitrc.org/projects/conn) in MATLAB. This pipeline includes realignment of the functional data using SPM12's realign & unwarp procedure (Anderson et al., 2001), where all scans are coregistered and resampled to a reference image (first scan of the first session) using b-spline interpolation. Subsequently, outlier detection was performed from the observed global BOLD signal and the amount of subject-motion in the scanner. Acquisitions with framewise displacement above 0.9mm or global BOLD signal changes above 5 SD are marked as potential outliers. Framewise displacement is computed at each timepoint by considering a 140x180x115mm bounding box around the brain and estimating the largest displacement among six control points placed at the center of this bounding-box faces. Afterwards, both the functional and the anatomical data are normalized into standard MNI space and segmented into grey matter, white matter, and CSF tissue classes using SPM12's unified segmentation and normalization procedure (Ashburner and Friston, 2005). As a last step, functional data is smoothed using spatial convolution with a Gaussian kernel of 8mm full width half maximum (FWHM), in order to increase BOLD signal-to-noise ratio and reduce the influence of residual variability in functional and gyral anatomy across subjects.

As a next step, denoising of the functional data was performed again using the standard pipeline from the CONN toolbox. For each participant, CONN implemented CompCor, a method for identifying principal components associated with segmented white matter (WM) and cerebrospinal fluid (CSF). In a first-level analysis, aCompCor components (Behzadi et al., 2007) and first-order derivatives of motion were entered as confounds and regressed from the BOLD signal. In addition, preprocessing steps included temporal band-pass filtering (0.008 Hz – 0.09 Hz), linear detrending, and regression of outlying functional volumes (>97th percentile in normative sample; global-signal z-value threshold = 5, subject-motion mm threshold = 0.09) identified using the artifact removal toolbox (ART) (https://www.nitrc.org/projects/artifact_detect/).

5.3.6 Functional network construction

To define brain regions of interest, we used dictionary learning (Mensch et al., 2016, Mensch et al., 2018) to extract 80 components from the denoised resting state data from our training set (Sample and 2). Performing the parcellation on the training set only was done to keep the test set (Samples 3 and 4) completely independent from the training set. Dictionary learning is a sparsity-based decomposition method for extracting spatial maps. It extracts maps that are naturally sparse and usually cleaner than ICA (Mensch et al., 2016), and it has been found to be the method that leads to the highest predictive success in a comparison of different connectome-based prediction pipelines (Dadi et al., 2019). In addition, the authors found that 80 components are the optimal amount for predictive performance. Subsequently, a random walk-based extraction of regions from the brain networks obtained by the dictionary learning algorithm was used as proposed in Abraham et al. (2014) resulting in 142 regions.

As we were interested in the contribution of the self-referential thinking, reward and cognitive control network in particular, we then conducted a conjunction analysis between the parcellated regions derived from the dictionary learning algorithm analysis and meta-analytically derived maps associated with, self-referential thinking, cognitive control and reward obtained using Neuroquery (Dockes et al., 2020). Neuroquery is a new meta-analytic tool for human brain mapping that was developed by researchers who were also involved in creating Neurosynth (Yarkoni et al., 2011). The advantage of Neuroquery is that it is focused on producing a brain map that predicts where in the brain a study on a particular cognitive process is likely to report observations, while Neurosynth tests the consistency of observations reported in the literature. Prediction, as opposed to statistical testing, is important because it can be applied *out-of-sample* and is thus more generalizable. To estimate functional connectivity between the resulting 17 regions (see Figure 5.1) efficiently, we use the Ledoit-Wolf regularized shrinkage estimator (Ledoit and Wolf, 2004; Varoquaux and Craddock, 2013), which gives a closed form expression for the shrinkage parameter. For parametrization of the functional interactions, Pearson's correlation was used.

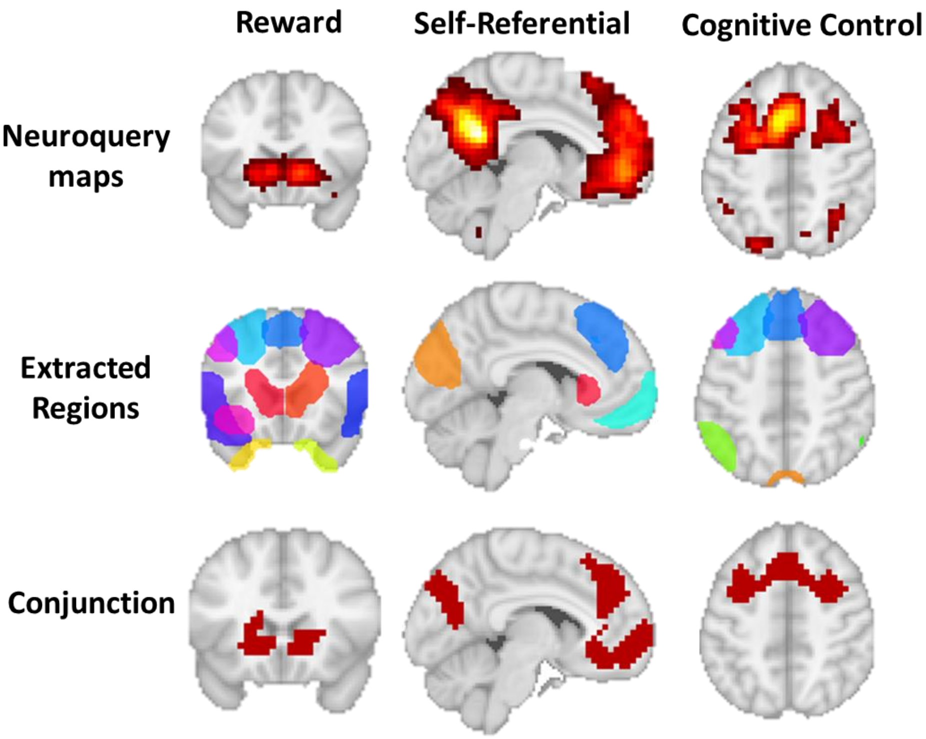


Figure 5.1. The top row depicts the Neuroquery maps for reward, self-referential thinking and cognitive control, respectively. The middle row shows the regions derived from the parcellation that showed overlap with the Neuroquery maps. The bottom row shows the overlap between the top row and middle row regions.

5.3.7 Connectome-based predictive modeling

The main analysis utilized CPM to predict participants' propensity to cheat from whole-brain resting state functional connectivity patterns. CPM is a recently developed tool for identifying functional brain connections related to a behavioral variable of interest, which are then used to predict behavior in unseen participants (i.e., participants whose data were not used in the training of the model; Shen et al., 2017). CPM was introduced and described in recent neuroimaging literature in a series of studies reporting its successful implementation in prediction of cognitive variables such as fluid intelligence, attention control and creativity (Finn et al., 2015; Rosenberg et al., 2016; Shen et al., 2017, Beaty et al., 2018). The MATLAB syntax used for CPM is freely available online (https://www.nitrc.org/projects/bioimagesuite/).

We implemented the CPM procedure in Python since the functional network construction steps were also performed in Python with use of the Nilearn package (Abraham et al., 2014b). We extended the standard procedure, described in detail below. As in the standard CPM approach (Shen et al. 2017), as a first step, the level of honesty which is the reverse of the number of times a participant cheated on the Spot-The-Difference task (the cheatcount), was correlated with each edge (i.e., correlation of mean BOLD signals between a given pair of brain regions) in the functional connectivity matrix of each participant (see Figure 5.2 A&B). Subsequently, a threshold was applied to the connectivity matrix to keep only the edges that were significantly positively or negatively correlated with honesty (see Figure 5.2 C).

In our extended procedure, we then fit a lasso regression model on the significant edges, to derive beta coefficients for each of the edges (see Figure 5.2 D). This allows us to assess the importance of the different edges that are entering the predictive model. Lastly, as in the standard approach, the model was applied to unseen participants in a leave-one-out cross-validation scheme (see Figure 5.2 E). Specifically, all steps described, including correlating edges with honesty and feature selection (thresholding) and model estimation, were implemented on n – 1 participants connectivity matrices and cheatcount scores, and then the fitted model was tested on the left-out participant. Leave-one-out cross-validation is the default scheme used in previous papers (Shen et al., 2017, Beaty et al., 2018) and thus renders our results comparable to previous findings. Because feature selection is performed inside the cross-validation loop, slightly different edges may be selected at each iteration which result in slight variations in the predictive models.

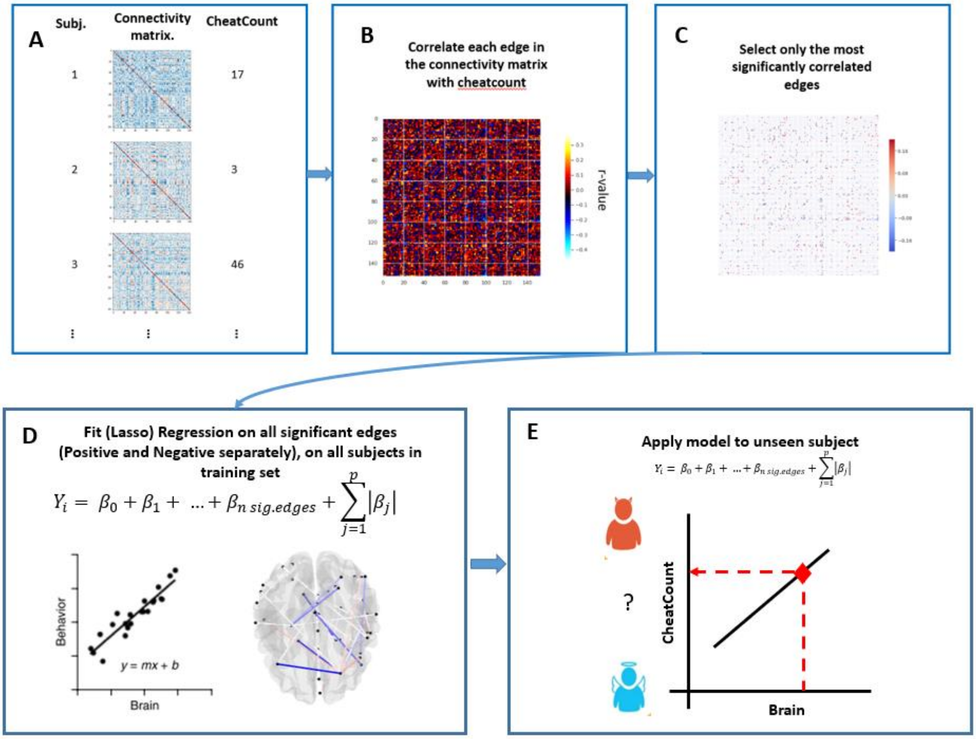

A
Subj.
Connectivity matrix.
CheatCount
1
17
2
3
3
46
B
Correlate each edge in the connectivity matrix with cheatcount
r-value
C
Select only the most significantly correlated edges
D
Fit (Lasso) Regression on all significant edges (Positive and Negative separately), on all subjects in training set
$Y_i = \beta_0 + \beta_1 + \ldots + \beta_{n\,sig.edges} + \sum_{j=1}^{p} |\beta_j|$
Behavior
Brain
y = mx + b
E
Apply model to unseen subject
$Y_i = \beta_0 + \beta_1 + \ldots + \beta_{n\,sig.edges} + \sum_{j=1}^{p} |\beta_j|$
?
CheatCount
Brain

Figure 5.2. Adapted CPM procedure: A-B) As a first step, the correlation between each edge in the connectivity matrix and level of honesty is computed. C) Next, the correlations are thresholded and only the significant edges are retained. D) We then fit a (lasso) regression model using the selected edges to the level of honesty on the participants in the training set. E) As a last step, we use the model to predict the level of honesty of the left-out subject (whether someone is more a saint = honest participant or a devil = person who cheats a lot/ cheater). This figure was adapted from Shen et al. (2017).

The CPM approach detailed here, was applied only to the two big samples (Samples 1 and 2). These two big samples were selected as training set as they allow for the biggest number of subjects to be used for feature selection and training of the model, which is important to ensure generalizability of the model and improve prediction accuracy (Sima et al., 2005; 2006). After executing the CPM procedure on this training set, we identified the significant predictors by running one-sample t-tests for each of the connections to test whether the average beta value across cross-validation folds of a given connection was significantly different from zero (see Step 1 in Figure 5.3). Once the significant connections were identified we used these predictors to train the model on the full training set (Step 2 in Figure 5.3) and subsequently tested the model on the test set (Samples 3 & 4; Step 3 in Figure 5.3). This procedure allowed us to perform feature selection and train models on samples that are completely independent of the samples in the test set. It is important to note that there was no significant difference in cheating propensity between the training set and the test set (t(90) = 0.31, p = 0.75).

The predictive power of the models was assessed by means of the statistical significance of the Pearson correlation between the predicted honesty scores and the observed scores. Statistical significance of this correlation is estimated by means of permutation testing where honesty scores are permuted, and the CPM procedure is repeated 1000 times. The empirical score is then compared against the null distribution to derive a p-value.

Our analysis approach focuses on maximizing relative performance, as we are interested to investigate the neural underpinnings of a propensity to cheat not how much precisely someone cheats on our task. As a consequence, functional connections between regions in the training set were selected based on the Pearson correlation, to maximize relative predictive performance, not absolute predictive performance. Even though the goal was to predict relative cheating propensity, to test for robustness, we also assessed predictive accuracy using the absolute measures of predictive performance, namely the mean absolute error (MAE) and the median squared error (MSE), in addition to the Pearson correlation (see Appendix 5.1).

To test the contribution of the predictors in the different models formally, we used a permutation importance approach (Breiman, 2001) to explore which predictors contribute most to the out-of-sample prediction accuracy. Specifically, the estimation for the given model on the training set (predictors identified on the training set using LOO-CV) was repeated and each of the predictors was permuted 100 times in sequence and the average prediction on the independent test set across the 100 permutations was recorded for each of the predictors. Finally, the reduction in correlation associated with permuting each coefficient (as compared to the correlation including the unpermuted prediction) at each fold was computed. Feature importance was then calculated as the difference in correlation between predicted and actual level of honesty between the baseline model (no permuted predictors) and the permuted model. Consequently, the higher the difference, the more important the predictor was.

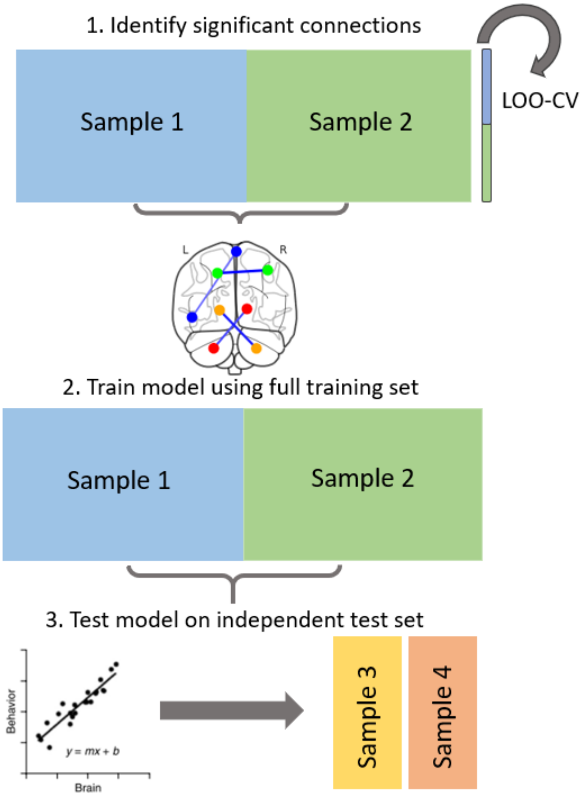

1. Identify significant connections
Sample 1
Sample 2
LOO-CV
L
R
2. Train model using full training set
Sample 1
Sample 2
3. Test model on independent test set
Behavior
Brain
$y = mx + b$
Sample 3
Sample 4

Figure 5.3. Procedure to test the predictive power of our models: First, significant connections were identified using leave-one-out cross validation on the training set, comprised of samples 1 and 2. Secondly, a model was trained using these predictors on the full training set. Lastly, the model was tested on an independent test set, comprised of samples 3 and 4.

5.4 Results

5.4.1 Behavioral Results

Substantial individual differences in the total amount of cheating were observed (Mean = 37%, Median=28%, SD=31%; see Figure 5.4: some participants cheated only on one or two trials (11% of participants), whereas others only missed one or two opportunities to cheat (4%).

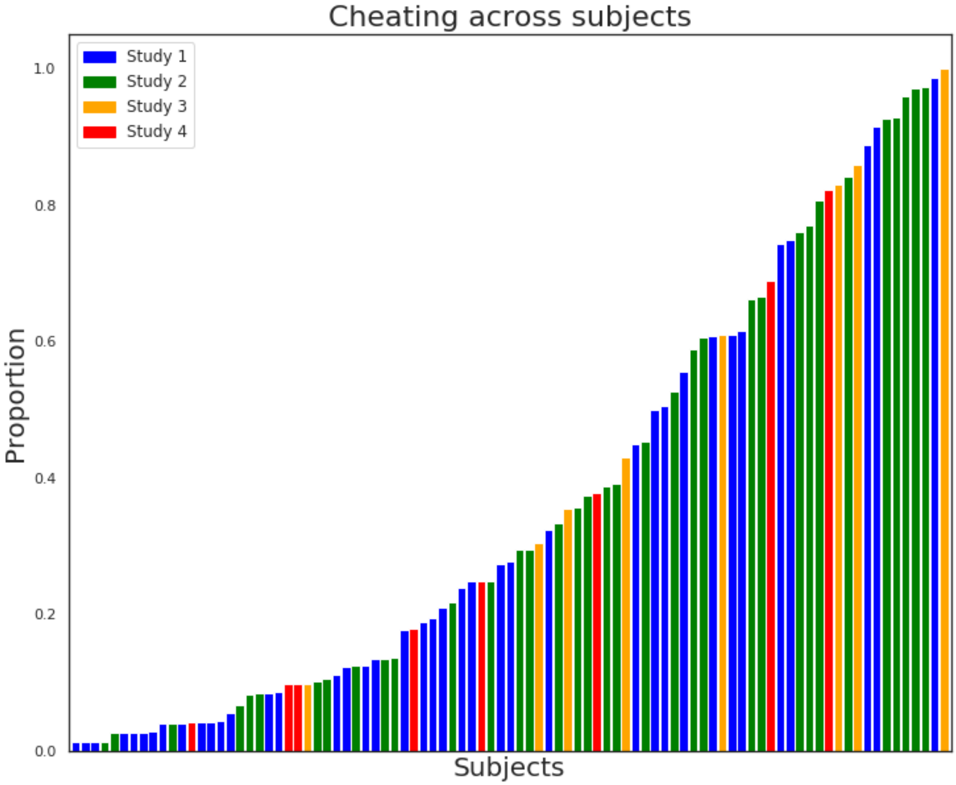


Figure 5.4. Individual differences in proportion of cheating on the Spot-The-Difference task. N = 91.

5.4.2 Cheating correlates with several measures of impulsivity

We tested whether the magnitude of (dis)honesty we measure with our task generalizes to personality characteristics, to explore the ecological validity of the Spot-The-Difference task. To test this, we correlated the individual differences in (dis)honesty with scores from measures of impulsivity, greed, creativity, moral foundations and manipulativeness, respectively (see Methods for the questionnaires). We found that levels of honesty (reverse cheatcount) negatively correlated (for all correlations see Appendix 5.2) with five measures of impulsivity: a) The Brief Sensation Seeking Scale ($r = -0.23$, $p < 0.05$), b) the Sensation Seeking subscale of the short version of the UPPS-P Impulsive Behavior scale ($r = -0.28$, $p < 0.05$), c) the Bas Drive subscale ($r = -0.31$, $p < 0.05$) of the BisBas scale, d) the Fun Seeking subscale of the BisBas scale ($r = -0.21$, $p < 0.05$), and e) the Risk seeking scale ($r = -0.25$, $p < 0.05$). In accordance with previous literature (Anderman et al., 2009), these findings suggest that cheating is associated with greater impulsivity.

Unexpectedly, we also found that more honest participants scored higher on reward responsiveness of the BisBas scale than less honest participants ($r = 0.21$, $p < 0.05$). Honesty also correlated negatively with how sensitive participants reported themselves to be to social norms ($r = -0.32$, $p < 0.05$), and to violations of purity ($r = -0.25$, $p < 0.05$); the more sensitive to social norms and violations of purity they reported themselves to be, the more they cheated. Note that the correlations presented above are not corrected for multiple comparisons, as we merely present them here to illustrate that our measure of cheating relates to these personality measures in the expected ways.

5.4.3 Predicting (dis)honesty using questionnaires

In order to more rigorously test whether the personality measures predict cheating behavior in our task, we investigated whether the questionnaire data could be used to predict honesty out-of-sample (leave-one-out cross-validation on training set to identify significant predictors & prediction on test set of independent samples). Using the significant predictors identified in the training set to test predictive accuracy on the independent test set (Samples 3 & 4) we observed no statistically significant prediction performance ($r = 0.14$, $p_{perm} = 0.3$). Suggesting that, while some personality measures are associated with cheating in our task, this association is not strong enough to allow out-of-sample prediction of cheating based on personality measures alone.

5.4.4 Predicting (dis)honesty using a resting state functional connectome

We subsequently investigated whether the resting state functional connectome could be used to predict honesty out-of-sample (leave-one-out cross-validation on the training set to identify significant predictors and prediction on the test set of independent samples). Using the significant predictors identified in the training set to test predictive accuracy on the independent test set we observed a significant prediction accuracy ($r = 0.40$, $p_{perm} < 0.05$; see Figure 5.6). The permutation importance analysis revealed that particularly the functional connectivity between the ventromedial prefrontal cortex (vmPFC) and the left temporal pole (TP) is important in predicting cheating in the independent sample. In addition, connectivity between the PCC and the left TP, the left caudate and the right TP, the left inferior frontal gyrus (IFG) and the PCC, the medial prefrontal cortex (MPFC) and the vmPFC, the left insula and the PCC, the right and left caudate and the PCC, the right angular gyrus (AG) and the vmPFC, the right caudate and the right TP, the left IFG and the left medial temporal lobe (MTL), and the left angular gyrus and the right dorsolateral prefrontal cortex (dlPFC) were found to be important predictors (see Figure 5.5). This suggests that these connections represent the neural substrate of honesty in the resting brain and can be used to identify cheaters, even in the absence of any temptation to cheat.

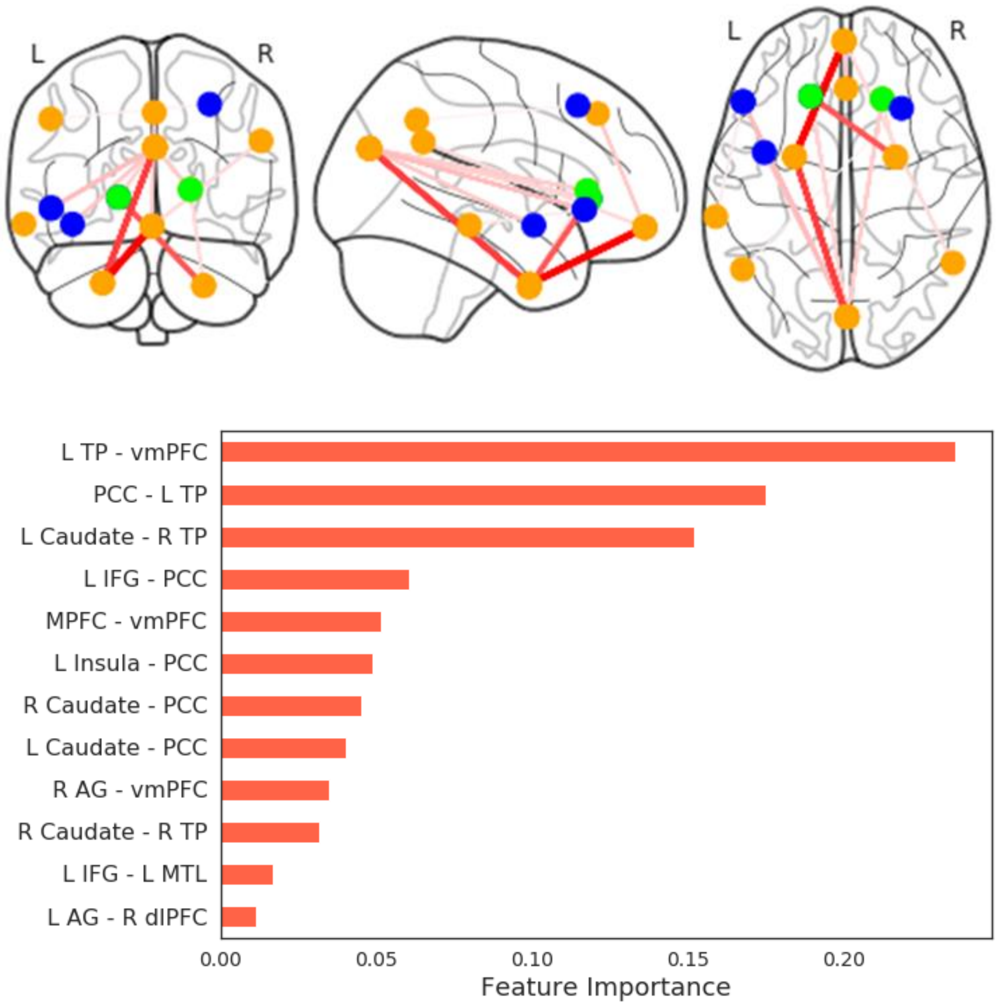


Figure 5.5. Higher functional connectivity between self-referential thinking and reward network is linked to more honest responses (lower cheatcount) in the Spot-The-Difference task. Top: Connectome of the significant predictors (identified on the training set) of honesty (negative predictor of cheatcount) averaged across folds (Red = positive association with honesty; thicker lines indicate stronger association; orange nodes represent regions of the self-referential thinking network, green nodes represent region belonging to the reward network and blue nodes represent cognitive control). Bottom:

Permutation importance for prediction on the test set for each of the predictors (identified on the training set).

5.4.5 Robustness of the neural model

To test for the robustness of our neural model we also applied the same analysis procedure (leave-one-out cross-validation on the training set to identify significant predictors and prediction on the test set of independent samples) on all 99 subjects. Using the significant predictors identified in the training set to test predictive accuracy on the independent test set, we observed a significant prediction accuracy (r = 0.65, p_{perm} < 0.001). This indicates that the model still significantly predicts cheating out of sample when the 8 additional participants are added to the data.

To test for potential effects of using a different composition of participants and samples in the training and test set, we conducted an additional CPM analysis where we randomly selected 75% of participants from all 4 samples as training set and the remaining 25% of participants as test set. It is important to note that again there was no significant difference in cheating propensity between the training set and the test set (t = 0.09, p = 0.92). We again performed the parcellation procedure only on the training set to keep training set and test set independent. We then used the lasso regression CPM approach to identify the most important predictors from the training set as detailed in the method section (see Figure 3) and then tested this model on the independent test set (see Figure 4). The analysis revealed that we are again able to significantly predict honesty from the resting-state functional connectome in the independent sample with similarly high accuracy (r = 0.43, p < 0.05). These findings demonstrate that there does not seem to be a significant effect of composition of the training set on the reported effects, highlighting the robustness of our findings.

5.4.6 Comparing predictive importance of the functional connectome to that of questionnaires

Next, we implemented a combined model where both the neural and questionnaire data were added to the lasso regression model.

After identifying the significant predictors using both the neural and the questionnaire data of the training set, we observed a significant predictive performance on the independent test set (r=0.49, $p_{perm} < 0.05$; see Figure 7). Thus, prediction on the independent samples was only significant using the neural data alone or combining both types of data, suggesting that the neural data is more important in predicting honesty out-of-sample.

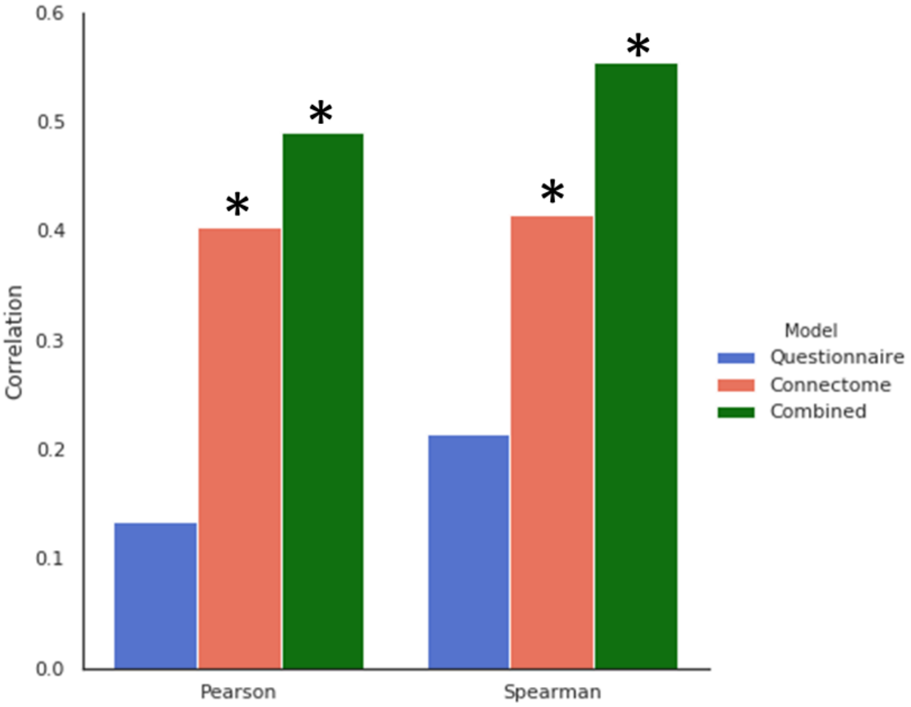


Figure 5.6. Pearson and Spearman correlations between predicted and actual level of honesty for the questionnaire, neural and combined model in comparison. *: significant at p<0.05 established using permutation tests (N=1000).

Here, the permutation analysis revealed that connectivity between the vmPFC and the left TP resulted in a reduction of the correlation of over 0.2, which is four times as large as any of the questionnaire predictors and highlights the importance of this predictor. In addition, connectivity between the right temporal pole and the left caudate, and the connectivity between the left temporal pole and the PCC, were important predictors of honesty (see Figure 5.7). The fact that connections differ slightly from the results reported above (Figure 5.5) is because here questionnaire data was added to the model.

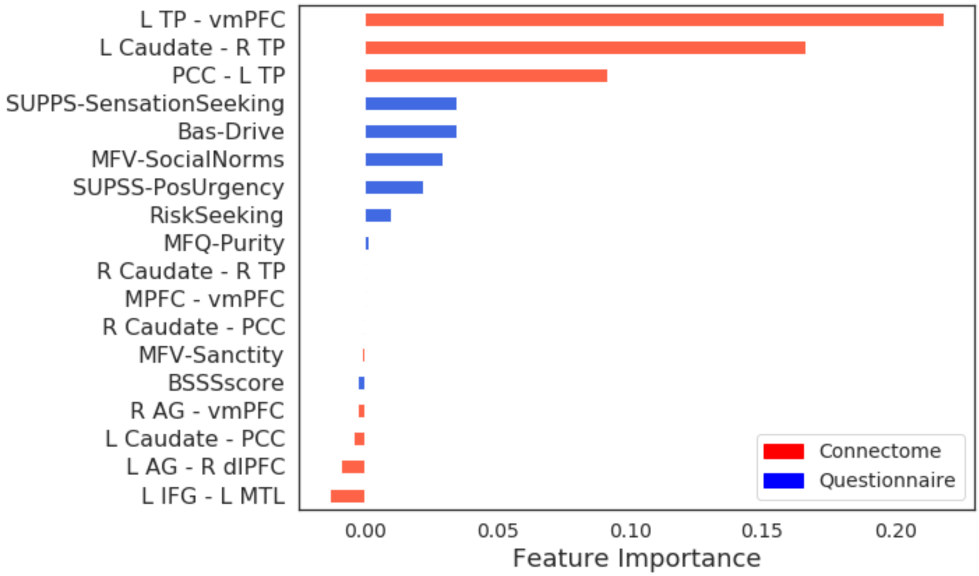


Figure 5.7. Comparing the contribution of the functional connectome and questionnaires as predictors of honesty on the independent test set. Permutation importance of predictors for prediction on independent samples.

5.5 Discussion

Employing connectome-based predictive modelling (CPM) in combination with our innovative Spot-The-Differences task, which allows for inconspicuously measuring cheating, we identified a functional connectome that can predict (dis)honesty in an independent sample. We observed a correlation between out-of-sample predicted and actual cheatcount (r = 0.40) that resides on the higher side of the typical range of correlations (between r = 0.2 and r = 0.5) reported in previous studies employing CPM (Shen et al., 2017). Thus, functional connectivity within the brain at rest accurately predicts whether someone is more honest, or more inclined to cheat in our task.

In light of previous research on moral decisions, the regions we identified in our resting state analysis can be associated with two networks frequently found to be involved in moral decision making. First, the vmPFC, the bilateral temporal poles and the PCC have consistently been associated with self-referential thinking. For example, it has been found that functional connectivity between these areas during rest is associated with higher-level metacognitive operations such as self-reflection, introspection and self-awareness (Gusnard et al, 2001; Meffert et al., 2013, Northoff et al., 2006; Vanhaudenhuyse et al., 2011). Secondly, the caudate nucleus, which has been found to be involved in anticipation and valuation of rewards (Ballard & Knutson, 2009; Knutson, Adams, Fong, & Hommer, 2001; Abe & Greene, 2014) can be considered an important node in the reward network (Bartra et al., 2013). Participants with higher levels of activation in the reward network, in anticipation of rewards, have previously been found to indeed be more dishonest (Abe & Greene, 2014).

It is interesting to note that the evidence from the current resting-state fMRI study and an earlier task-based study (using the same task) investigating cheating behavior (Speer, Smidts & Boksem, 2020), strongly converge. In that study it was found that, while activity in the nucleus accumbens promoted cheating, a network consisting of posterior cingulate cortex, temporoparietal junction, and medial prefrontal cortex promoted honesty. This convergence highlights the robustness and replicability of these findings. In concert with the other fMRI studies measuring dishonesty and fairness mentioned above, this emphasizes the importance of the interaction between self-referential thinking and reward network in determining our propensity for (dis)honesty.

Notably, previous task-based fMRI studies found that regions related to cognitive control, such as the ACC, dlPFC and IFG, also play an important role in determining the outcome of (dis)honest decisions (Abe & Greene, 2014, Speer, Smidts & Boksem, 2020). Specifically, cognitive control helped dishonest participants to be honest, whereas it enabled cheating for honest participants. In the present study we only find limited evidence for the importance of cognitive control in determining the propensity to cheat. A possible explanation may be that the extent to which our moral decisions are governed by our focus on moral self-concept as opposed to being focused on reward, determines our general and stable moral default (i.e., propensity to cheat). Cognitive control is only engaged in specific situations or contexts when we need to override our moral default to create a subjectively justifiable balance between the two opposing forces (reward & self-concept). As a consequence, particularly the resting-state functional connectivity between brain areas related to reward and self-concept, but not cognitive control, is predictive of the propensity to cheat.

Further support for our interpretation that the identified regions indeed belong to the proposed networks, derives from the fact the regions were selected based on a conjunction analysis with meta-analytically derived maps associated with self-referential thinking, reward, and cognitive control, obtained using Neuroquery (Dockes et al., 2020). Selecting regions based on this conjunction supports our interpretation of the observed functional networks and reduces the reverse inference problem (Poldrack, 2006). Our findings thus suggest that, within the brain at rest, higher connectivity within the self-referential thinking network, but also between self-referential thinking and reward regions increases a participant’s propensity for honesty long before being confronted with an actual opportunity to cheat. Stated differently, the inherent reactivity of the identified connectome plays an important role in biasing a person towards (dis)honesty even before being confronted with any decision situation.

We also found that several well-established self-report personality measures of impulsivity correlated significantly with dishonesty on our task. This highlights the ecological validity of the Spot-The-Difference task as a measure of dishonesty, as impulsivity has frequently been associated with other forms of cheating such as academic cheating (Cochran et al., 1998; Anderman et al., 2009), dishonesty more generally (Gino et al., 2011) and unethical behavior (Zimmerman, 2010; Loeber et al., 2014). While we found that these self-reported measures of impulsivity were correlated with dishonesty, the predictive power of these measures was significantly outperformed by the predictive accuracy of the functional connectome.

This higher predictive importance of neural measures could in part be due to the fact that self-report measures may suffer from social desirability bias, particularly in the context of dishonesty (Grimm, 2010). For example, participants may not want to admit that they are impulsive or may not even be aware that they are. This social desirability bias might explain why we unexpectedly observed that participants who report to be more sensitive to social norms and matters of purity also cheat more. Some participants, particularly the dishonest ones, might have overstated how much they care about, and base their decisions on, social norms and moral purity.

The predictive advantage of the neural measures might be attributed to the fact that they are uncontaminated by the biases mentioned above and may represent neuro-cognitive processes that govern our behavior in a way that we often may not be aware of. The functional connectome at rest, when combined with a reliable and valid measure of behavior, appears to provide direct and reliable access to different behavioral tendencies. Importantly, it permits measuring the neural manifestation of socially inadmissible behavioral tendencies such as dishonesty, which could remain undetected by self-report measures. Therefore, rsfMRI in combination with well-designed behavioral tasks inconspicuously eliciting socially undesirable behaviors, may be a powerful tool to unobtrusively and accurately study the neurocognitive determinants of these inadmissible behaviors.

Nevertheless, it is worth noting that participants who cheated or were honest on our task, may not act similarly in a different context. We demonstrate that cheating on our task can be predicted by stable and reliable resting-state functional connectivity patterns and well-established personality characteristic related to impulsivity. Therefore, our findings suggest that these individual differences in (dis)honesty may indeed reflect stable moral traits. Further research is needed, however, to confirm that these individual differences in (dis)honesty and its manifestation in the resting connectome extends to other tasks, contexts and moral domains.

As mentioned in the introduction, adopting an out-of-sample prediction approach on independent samples instead of using in-sample correlation, as was done in most previous rsfMRI studies, is a more conservative approach of investigating brain-behavior relationships and therefore increases the likelihood of replication in future studies (Shen et al., 2017). Combining CPM with regularized regression, we further reduced the chance of overfitting as irrelevant predictors are shrunk to zero, which additionally improves the robustness of our models. Moreover, the predictive power of our models remains even after several subjects (~10%; the subjects that did not fill out the questionnaires) are added, further highlighting the robustness of our findings. In addition, we ruled out potential confounds induced by the specific samples used for training and testing by using a different composition of participants and still observing similar predictive accuracy. Finally, the same networks of brain regions were found to be predictive of honesty in the current study as in a task-based fMRI study (Speer, Smidts, Boksem, 2020) using the same paradigm, which further provides evidence for the reliability of our resting state results. Nevertheless, our results provide only a first insight into how functional connectivity is related to individual differences in (dis)honesty and replication with larger samples is needed to further validate the robustness of these findings.

In summary, our extended CPM model applied to a large and diverse sample revealed that stronger connectivity within the brain at rest between brain areas associated with self-referential thinking and reward-processing, predispose individuals to behave honestly even in absence of any temptation to cheat. We showed that connectome-based predictive modelling on resting-state data exceeds self-report measures in robustly predicting cheating behavior out-of-sample with high accuracy. This approach may prove highly useful for future studies investigating the neural manifestation of socially undesirable traits like dishonesty, as they provide direct access to the mechanisms underlying behavior that would likely remain concealed using more traditional methods.

Appendix Chapter 5

Appendix 5.1: Testing different cross validation schemes

To test the robustness of the lasso models we repeated this analysis with different metrics for predictive accuracy. Specifically, we repeated the analysis with absolute measures of predictive accuracy, namely the mean absolute error (MAE) and the mean squared error (MSE).

For the absolute measure (MAE & MSE) only the combined model achieved significant out-of-sample prediction. It has to be noted that the brain model was marginally significant (MSE: p = 0.06; MAE: p = 0.07) and with a slightly larger sample may have reached the level of significance. The questionnaire model did not achieve significant out of sample prediction for any of the measures (see Figure S1). It is important to note, that the selection of brain regions in the training set was done with the goal to maximize relative accuracy and not absolute accuracy, which explains why relative measures are superior to absolute measures.

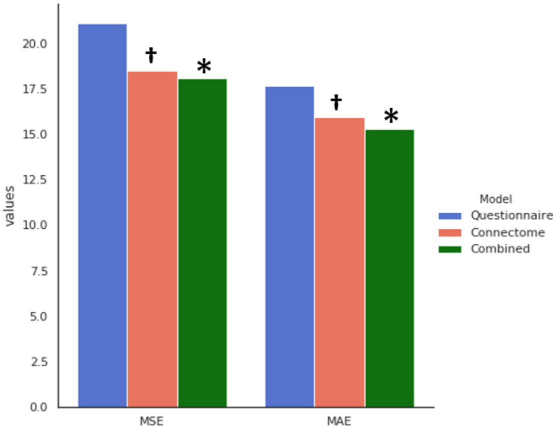


Figure S6. Comparing the predictive accuracy between models (Questionnaires, Neuro & Combined) using two absolute measures (mean squared error and mean absolute error). Only the combined model achieves

significant absolute prediction. *: significant at $p<0.05$ established using permutation tests (N=1000); † = $p < 0.1$.

Appendix 5.2: Personality measures

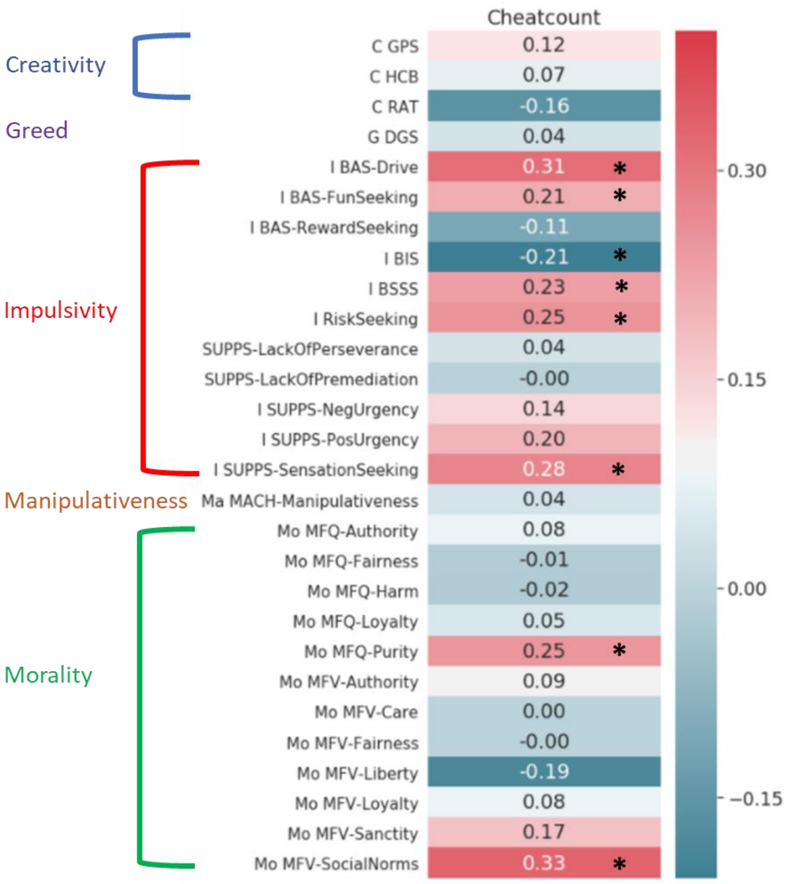


Figure S7. Correlations between all personality measures used and the cheatcount.

C = Creativity; G = Greed; I = Impulsivity; Ma = Manipulativeness; Mo = morality; GPS= Gough's Personality Scale; HCB = Hovecar's Creative Behavior; DGS = Dispositional Greed Scale; BAS = Behavioral Approach System; BIS = Behavioral Inhibition System; BSSS= Brief Sensation Seeking Scale; SUPPS = Short Urgency, Premediation (lack of), Perseverance (lack of), Sensation Seeking, Positive Urgency Impulsive Behavior Scale; MACH = MACH-IV test of Machiavellianism; MFQ = Moral Foundations Questionnaire; MFV = Moral Foundations Vignettes.* = indicates significant correlations (p<0.05 uncorrected).

Chapter 6

Decoding fairness motivations from multivariate brain activity patterns[5]

6.1 Abstract

A preference for fairness may originate from prosocial or strategic motivations: we may wish to improve others' well-being or avoid the repercussions of selfish behavior. Here, we used fMRI to identify neural patterns that dissociate these two motivations. Participants played both the Ultimatum and Dictator Game (UG/DG) as proposers. Because responders can reject the offer in the UG, but not the DG, offers and neural patterns between the games should differ for strategic players, but not prosocial players. Using multivariate pattern analysis we found that the decoding accuracy of neural patterns associated with UG and DG decisions correlated significantly with differences in offers between games in regions associated with theory of mind (ToM), such as the temporoparietal junction, and cognitive control, such as the dorsolateral prefrontal cortex and inferior frontal cortex. We conclude that individual differences in prosocial behavior may be driven by variations in the degree to which self-control and ToM processes are engaged during decision-making such that the extent to which these processes are engaged is indicative of either selfish or prosocial motivations.

6.2 Introduction

Our social life abounds with situations in which there is a conflict between selfish urges and the welfare of others. An example would be whether to donate a recently received windfall to charity or spend it on a new phone. Whereas highly prosocial individuals such as Martin Luther King or Mother Theresa may devote their life to improve the condition of the oppressed or the less fortunate, others tend to focus on maximizing their own gains and in

[5] This chapter has been published in SCAN (Speer & Boksem, 2019).

some cases even exploit others. These large differences in how individuals weigh their own benefits against another person's welfare are crucially important in understanding social decision making that enables cooperation on a societal level. Yet, until now the underlying psychological and neural mechanisms of these individual differences in prosociality have remained largely elusive.

A common framework to study the tradeoff between prosociality and selfishness is the Ultimatum Game (UG; Gueth et al., 1982). In the UG a proposer can decide how to divide a sum of money and a recipient can subsequently accept or reject this proposal. In case the recipient accepts the offer, the money is distributed as proposed, whereas nobody receives anything in case the responder rejects the offer. Previous research shows that the majority of proposers split the money about equally and offer on average 40% of the initial endowment (Henrich et al., 2005; Oosterbeek, Sloof, & Kuilen, 2004). In order to explain why proposers deviate from purely selfish behavior two competing mechanisms have been proposed. Proposers may decide to share because they care for the welfare of their opponent and may thus be driven by prosocial, or fairness concerns (Oosterbeek et al., 2004; Thaler, 1988). Alternatively, proposers may give high offers to reduce the chance of rejection and strategically maximize their financial gains, reflecting essentially selfish motives (Forsythe et al., 1994; Fehr & Schmidt, 1999).

To test which of these mechanisms dominates, the UG can be used in combination with the Dictator Game (DG), in which the responders cannot reject the offers. Hence, there is no punishment threat in place to prevent selfish behavior. As a consequence, the proposers' decisions in the DG are considered a straightforward reflection of their social preferences, as they are uncontaminated by strategic considerations. Critically, proposers motivated by prosocial fairness concerns should offer equally in both games as the importance of welfare of the opponent is unchanged across games, while selfish proposers are expected to strategically split their endowment equally in the UG to reduce the chance of rejection and give low offers in the DG.

Accumulative evidence from previous studies indicated that while the average offers in the DG are significantly lower than in the UG, they still remain well above what would be expected by standard economic theory (Forsythe et al., 1994; Kahneman et al., 1986). These findings suggest that some proposers indeed change the size of their offer quite drastically between games, while others offer similar amounts in both games. Thus, based on high vs low differences in offers between games, we can distinguish between those who are essentially selfish and only strategically increase their offer when punishment is possible, whom we characterize as selfish players, and those who do not, characterized as prosocial players.

Previous neuroimaging research on the proposers motivations in the UG and DG has focused particularly on social norm compliance, converging on the crucial involvement of the dorsolateral prefrontal cortex (dlPFC) and cognitive control in strategic fairness in the UG (Yamagishi et al, 2016; Strang et al., 2014; Spitzer et al., 2007; Steinbeis et al., 2012). The experimental evidence from these studies generally suggests that in order to act prosocially in the UG, proposers need to control selfish impulses, which is reflected in increased dlPFC activity.

Other studies, however, have challenged this central role of cognitive control in prosocial behavior. Evidence comes from two meta-analyses indicating that prosocial choices and cooperation across various different economic games, such as the trust game, public goods game and Dictator Game are more frequent when cognitive control is *reduced,* such as when one is under time pressure (Rand et al., 2015; Rand, 2016) and when primed to trust ones intuitions (Rand, Greene, & Nowak, 2012; Lotz, 2015) . Furthermore, it has been shown that when cognitive control is interfered with, for example by concurrently performing an N-Back-task, proposers made more generous offers in the DG (Schulz et al., 2014). Collectively, these results suggest that, at least under some conditions, self-control may not be necessary to act prosocially, and reduced self-control may even promote prosocial behavior.

In a largely separate literature, it is widely accepted that theory of mind (ToM) is a strong determinant of prosocial sharing (e.g. Batson, 1991;

Penner et al., 2005; Eisenberg & Miller, 1987; Pavey et al., 2012; Edele et al., 2013; Artinger et al., 2014). For example, when proposers were instructed to imagine being in the position of a responder, they offered significantly higher amounts of money (Hoffman, McCabe, & Smith, 2000).

Thus, there seems to be a disconnect between behavioral research on the proposers in the UG, emphasizing the role of ToM in prosocial behavior, and the neuroimaging literature on the proposers' motivations and behavior, focusing on the role of self-control in social norm compliance. These neuroimaging studies do not discuss the relevance of ToM and do not report activity in the neural network underlying ToM: a widely distributed network of brain regions including the temporo-parietal junction (TPJ), the precuneus, the medial prefrontal cortex (MPFC), the angular gyrus and the temporal lobes (Gallagher & Frith, 2003; Ciaramidaro et al., 2007; Schurz et al., 2014).

Here, we try to bridge this disconnect by specifically testing for differences in the ToM network in participants playing the UG. Specifically, we combined a mixed UG–DG paradigm with functional resonance imaging (fMRI) to identify neural patterns that dissociate selfish from more prosocial individuals, using multivariate pattern analysis (MVPA), which may be more sensitive to pick up more subtle differences in psychological processes such as empathy and ToM (Kriegeskorte et al., 2006; Norman et al., 2006).

6.3 Methods

6.3.1 Participants

The reported analyses are based on 31 participants (22 females; age 18-44 years; *M* = 24,2, *SD* = 6,2) from two separate studies. The reason for running two studies was driven by funding issues. We ran out of funding half-way through the data collection and once we obtained additional funding continued with the data collection. The first sample of participants consisted of a student sample (N=19, 13 females; age 18-44 years; *M* = 23,3, *SD* = 7,0) from now on referred to as Study 1. The second sample also consisted of a student sample, but from a different university, (N=12, 9 females; age 22 -31 years; *M* = 26,2, *SD* = 2,7) from now on referred to as Study 2. No significant differences in demographics were identified between samples. All participants were right-handed with normal or corrected to normal vision, and no record of neurological or psychiatric diseases. The studies were approved by the university Ethics Committees and were conducted according to the Declaration of Helsinki.

6.3.2 fMRI acquisition

For Study 1, the functional magnetic resonance images were obtained using a 3T Siemens Allegra MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in ascending interleaved order (3.5 mm slice thickness, 3.5 × 3.5 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 90°, FOV =224). Echo time (TE) was 30ms and repetition time (TR) was 2000ms. A T1-weighted image was acquired for anatomical reference (1.0 × 1.0 × 1.0 mm resolution, 192 sagittal slices, flip angle = 9°, TE = 2.6ms, TR = 2250ms).

For Study 2, the functional magnetic resonance images were collected using a 3T Siemens Verio MRI system. Functional scans were acquired by a T2*-weighted gradient-echo, echo-planar pulse sequence in descending interleaved order (3.0 mm slice thickness, 3.0 × 3.0 mm in-plane resolution, 64 × 64 voxels per slice, flip angle = 75°). TE was 30ns and TR was 2030ms. A T1-weighted image was acquired for anatomical reference (1.0 × 0.5 × 0.5 mm resolution, 192 sagittal slices, flip angle = 9°, TE = 2.26ms, TR = 1900ms).

6.3.3. Experimental task and procedures

Participants played a mixed Ultimatum/Dictator Game in the MRI scanner. On 24 trials, participants received €20 and had to decide how to split the endowment between themselves and an opponent. On each trial they were presented with a picture of the opponent before and during the decision process, in order to ensure that participants knew they were playing against a different human player on each trial. The pictures used in the study were obtained from the NimStim face database (Tottenham et al., 2009). We selected faces which were categorized as neutral and most representative of our participant population in terms of age and ethnicity. This was done to minimize the effect of the difference in faces on offers made in the games. Further choosing a representative sample of pictures was intended to increase the credibility of our cover story that participants played against previous participants. On half of the trials, the opponents were able to reject the offer, which would result in both the participant and opponent receiving nothing (UG). On the other half of the trials the opponents were passive recipients and could not reject the offer (DG). The critical difference between these conditions is that in the UG trials the participant can be punished for unfair offers whereas in the DG trials no punishment is possible. In the non-social control condition (24 trials), participants played against a computer algorithm, allegedly programmed to mimic human behavior. Again, half of the trials in the control condition were UG trials and the other half DG trials. Practice trials were implemented to in order to familiarize the participants with the task. In addition, participants were told that they were playing against participants who had previously participated in the study. As mentioned above, pictures of opponents were chosen with the aim of maximizing representativeness of the sample used in order to increase credibility of the story.

The trials started with a screen that presented a picture of the opponent and their power to reject the offer or not (UG or DG). Subsequently, the response options appeared, 0 to 14 € in steps of two, and participants could indicate their choice. Lastly, a wait screen appeared for eight seconds (see Figure 6.1).

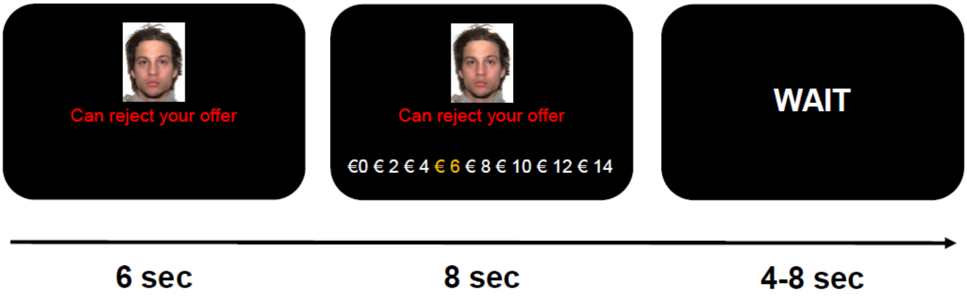


Figure 6.1. The Timing of the Ultimatum Game task. The experiment starts with a waiting period of 8 seconds, followed by a decision phase of 6 seconds

in which all the information relevant to the decision maker is present and ends with the response phase of 8s in which participants could indicate their choice. For the fMRI analysis the period 6s prior to the button press was used.

6.3.4 fMRI Analysis

6.3.4.1 Preprocessing

All fMRI data underwent the standard FSL (5.0) preprocessing pipeline. Anatomical scans were reoriented to the FSL standard orientation and skull-stripped. The functional data was motion corrected to the mean image using FSL's MCFLIRT and coregistered to the anatomical scan and normalized to the standard MNI brain using boundary-based registration (FSL FLIRT & FNIRT). Subsequently, Gaussian high-pass filtering with 100 seconds FWHM was applied.

To obtain neural activation patterns for multivariate analysis individual time series were modeled using a double γ hemodynamic response function in a single trial GLM design using FSL's FEAT. Specifically, one GLM fitted a hemodynamic response function (HRF) for each trial, following the Least-Squares all (LSA) approach (Mumford, Turner, Ashby, & Poldrack, 2012), using the six seconds prior to the keypress in each trial, resulting in 48 (12 trials * 4 conditions) parameter estimates of sustained activations for each participant. Specifically, there were 12 trials of UG and DG in the social condition (human opponent) and 12 trials of both games in the non-social condition (computer opponent). The resulting β-values were converted to t-values (Misaki, Kim, Bandettini, & Kriegeskorte, 2010), resulting in a whole-brain pattern of t-values for each trial. The duration of the epoch we used for our fMRI analysis was 6s and onset times were determined by counting back 6s from the point in time when the participant had indicated his choice. This window was used as it provides all the necessary information to make the decision and is free of brain activity related to motor responses. Average background signal and white matter signal were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. In order to test whether differences between games can be decoded from the brain we applied a classification analysis to the whole brain activity patterns estimated from our single-trial-GLM. Classification analyses were conducted with the PyMVPA toolbox (Hanke et al., 2009) and custom Python scripts, which will be made publicly available.

6.3.4.2 Relating classification accuracy to behavior in the theory of mind and cognitive control network

First, we investigated whether neural patterns in regions associated with ToM and cognitive control processes distinguish between prosocial and selfish players. In order to do so, we obtained a ToM-brain-activation mask and a cognitive control-brain activation mask by conducting a meta-analysis via Neurosynth (Wager et al., 2011, See Appendix 6.1). The masks were derived from a meta-analysis of previous studies reporting brain regions that are consistently active in articles that include the term "theory of mind" and "cognitive control" in the abstract (N=181, N=598 respectively; see Figure 6.2 & 6.3). Using this large scale automated meta-analysis increases confidence that neural activation in these regions indeed reflect engagement on the hypothesized processes, while reducing problems with reverse inference (Poldrack, 2006).

We extracted the activation patterns within the ToM and cognitive control mask in native space, corresponding to one decision phase each, which were labelled according to their experimental condition, either UG trial or DG. Only activation patterns for the social condition (human opponents) were selected for this analysis. Subsequently, for each activation pattern each voxel was standardized such that they have zero mean and unit variance. Lastly, we applied univariate feature selection, in which the 1000 voxels with the highest F-value resulting from an ANOVA on the contrasts of interest (UG vs. DG), were selected from within the masks. Subsequently, a multivariate pattern classification using a support vector machine (C=1) was applied to the selected voxels (Cox & Savoy, 2003; Mitchell, 2004). To account for possible effects of the size of the masks, 1000 voxels were selected for both masks. In order to avoid overfitting and inflated prediction accuracy (Vul, Harris, Winkielman, & Pashler, 2009) the three steps (feature scaling, feature selection & classification) were implemented using 6-fold cross validation.

To explore how differences in neural patterns between conditions (UG vs DG) across participants relate to differences in behavior we correlated the classification accuracies for each participant, derived from the above classification analysis, with the average difference in offers between games of each participant, while controlling for cohort effects. This partial correlation with a dummy variable for the cohort was used to account for possible variance introduced by difference between the two samples and scanners used. The same analysis was conducted for activation patterns for the non-social condition in order to test whether possible effects are specific to the social condition or generalize across all trials.

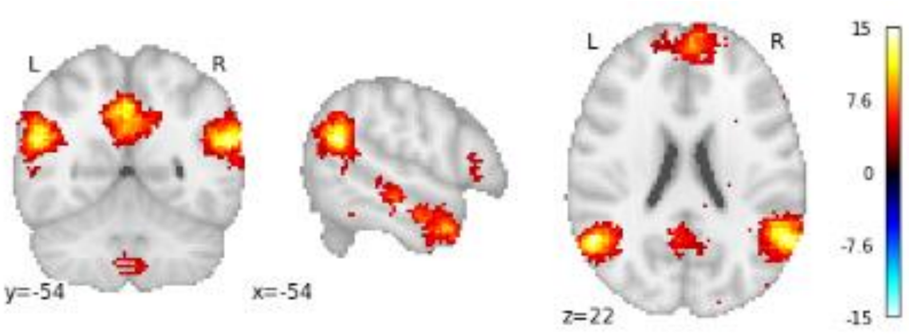


Figure 6.2. The theory of mind mask obtained from the meta-analysis on Neurosynth showing clusters in the precuneus left and right TPJ, temporal pole and the MPFC.

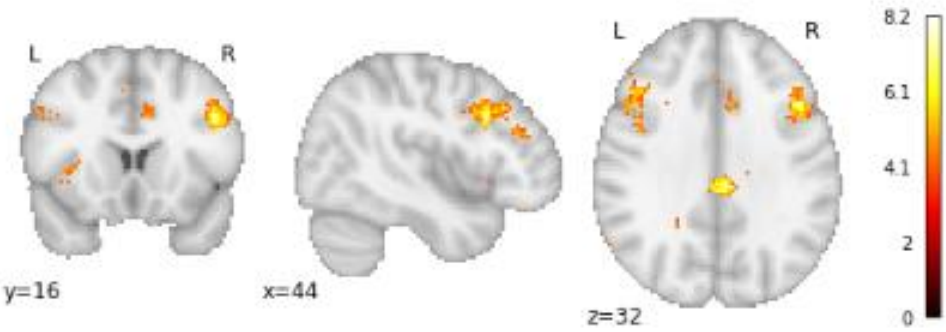


Figure 6.3. The cognitive control mask obtained from the meta-analysis on Neurosynth showing clusters in the left and right dlPFC and the PCC.

6.3.4.3 Relating classification accuracy to behavior in individual regions in the ToM and cognitive control network.

Next, we explored whether neural patterns in individual regions within the ToM and cognitive control networks differentiate between prosocial and selfish players using the same masks as above. Individual clusters within the masks were extracted and converted to each subject's native space (see Appendix 6.1). We then extracted the activation patterns within these ROIs. As before, only activation patterns for the social condition (human opponents) were selected for this analysis. Apart from the feature selection, which in this analysis was done by means of selecting individual ROIs, we applied the same analysis procedure as above.

6.3.4.4 Localizing other areas that discriminate between prosocial and players

To explore whether activity patterns in regions outside of the ToM and cognitive control networks correlated with individual differences in behavior, a classification searchlight approach was used. More specifically, we investigated where in the brain average UG and DG offers between subjects correlate with classification accuracy between UG and DG trials. To this end, we employed a spherical searchlight of 3-voxel radius. For this analysis the individual t-stat maps were smoothed (fwhm=8) to render the output of the searchlight more clustered and more interpretable. At each location, for all voxels included in the current sphere, the searchlight performed a classification of game type using a support vector classifier implemented with 6-fold cross validation. This was done for all subjects, resulting in 31 classification maps, where each voxel represents the classification accuracy of local UG vs. DG patterns. Subsequently, to test whether within-subject classification accuracy between UG and DG correlates with individual differences in behavior, we calculated at each voxel a Pearson's correlation between the classification accuracy and mean difference in offers between UG and DG trials while partialling out the group membership to the two cohorts (Study 1 & Study 2). We estimated statistical significance by permutation testing. Specifically, we shuffled the differences in behavior 5000 times and obtained null correlation maps and derived the empirical *p* value at each voxel from the voxel's own null distribution.

6.4 Results

6.4.1 Behavioral results

6.4.1.1 Social trials

On average, participants made higher offers in the UG than in the DG, but this difference varied substantially between subjects (UG-DG; M= 8.65, SD=3.27; see Figure 6.4): some participants offered much more in the UG, while other subjects made very similar offers in both games. The lowest mean offer observed on a UG trial for a participant was €8.5 which excludes the possibility that low differences in offers between games could be due to low offers in both games (see Appendix 6.2, Figure S1). From now on we refer to participants with a low difference in offers between games as prosocial players (as they give high offers in both games), whereas participants that exhibited high difference in offers between games as selfish players (as they only strategically offer high amounts when punishment is possible, but give low offers otherwise).

As this study contained data from two different samples using different scanners, we tested whether there were differences in behavior between the two samples using a two-sample t-test and found that the differences in offers between games was significantly higher in the second sample as compared to the first sample (*t(df=29)*=-4.41, *p*<0.001), which, in addition to the different scanning parameters used in the two studies, necessitates using partial correlation for our fMRI analyses.

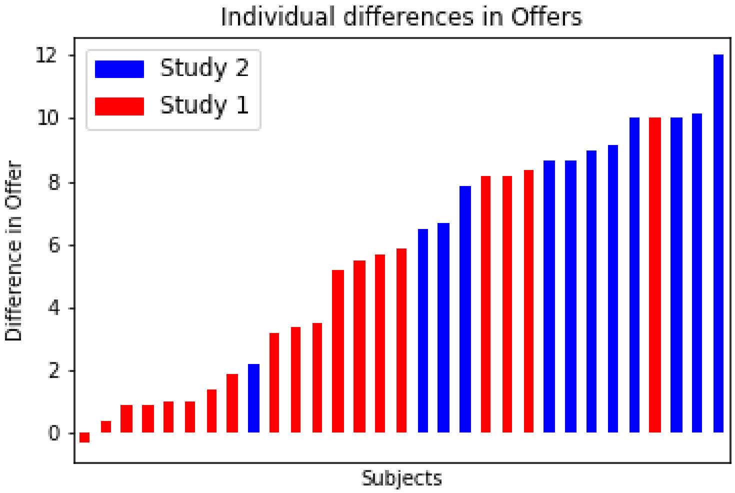


Figure 6.4. Differences in Offers (UG-DG) between games for Human trials across participants in both studies

6.4.1.2 Behavioral results for social vs. non-social trials

Here, we investigated whether participants give high offers out of *social* motivations (compliance with social norms or prosocial behavior) or whether they are driven by a more general response to any type of social or non-social punishment threat. If the underlying decision-making processes are indeed social, we would expect prosocial participants to exhibit larger differences in offers between the social and non-social condition as compared to the more strategic players. To test this, a multiple regression analysis was performed to predict offers in the DG based on the type of opponent (human vs. computer) as a dummy variable, the mean difference in offer between games as continuous regressor (indicating the extent to which participants are prosocial or selfish), and the interaction between these two variables. Only DG trials were used as differences between motivations (social vs. non-social & fair vs selfish) should be expressed particularly on these trials. The regression model was found to be significant ($F(2,58)=67.78$, $p<0.001$), with an R^2 of .882. Participants gave higher offers in the social condition than in the non-social condition (b = 3.725, SE = 0.66, t = 5.63, p = <.001) and DG offers decreased with higher difference in offers between games (more selfish participants offered less; *b* = -0.461, *SE* = 0.07, *t* = -6.585, *p* = <.001). Importantly, a significant interaction effect was found between the difference in offers (prosocial vs. selfish players) and type of opponent (human vs. computer), *b* = -0.358, *SE* = 0.099, *t* = -3.618, *p* = <.001: in the social condition participants made significantly higher DG offers than in the non-social condition and that this effect was stronger for prosocial players as compared to selfish players.

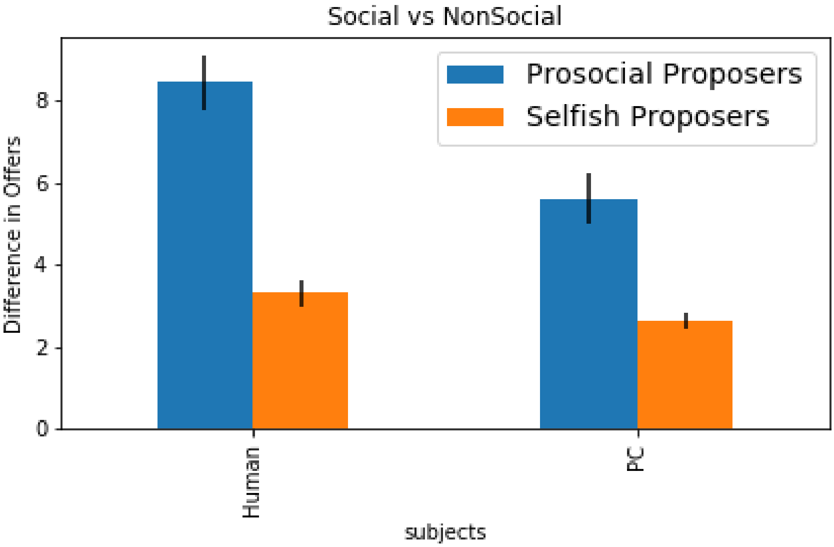


Figure 6.5. Bar graph illustrating the difference in offers in the DG between prosocial and selfish proposers (median split, for illustrative purposes only) on human and computer trials.

6.4.2 fMRI results

6.4.2.1 Classification accuracy of game type correlates with prosociality in the ToM & cognitive control networks

For the social trials we found that classification accuracy in the cognitive control network significantly correlated with the difference in offers between games (*r* = 0.48, *p* < 0.01). Further, classification accuracy on patterns from the ToM network also significantly correlated with difference in offers (*r* =

0.37, *p* < 0.05). This suggests that neural patterns in both the cognitive control network and the ToM network differentiate between prosocial and selfish players (see figure 6.6). For the non-social trials no significant correlation was found for the ToM network (*r* = 0.18, *p* = 0.34), or the cognitive control network (*r* = 0.30, *p* = 0.09).

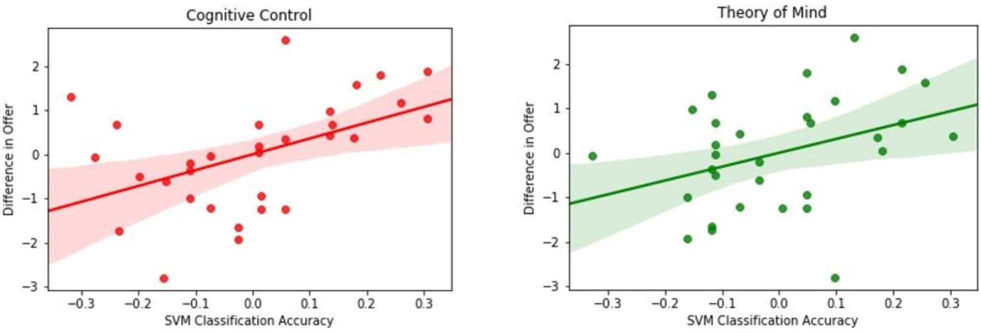


Figure 6.6. Correlation between classification accuracy and difference in offer (UG-DG) for neural patterns in the cognitive control network (red) and the ToM network (green).

6.4.2.2 classification accuracy correlates with prosociality in specific regions within the cognitive control and ToM network.

ToM network: When correlating the individual differences in offers between games with the classification accuracy from the support vector classifiers we found a strong positive partial correlation in the left temporoparietal junction (TPJ) (r=0.5, p<0.005, $p_{adjusted}$=0.03, FDR corrected for multiple tests at p=0.05). Thus, neural patterns in the left TPJ differ more strongly between games for selfish players than for prosocial players. We repeated this analysis for the non-social condition and found that there were no significant correlations between the classification accuracy and differences in offers between games in any of the ROIs. Hence, the effects observed are specific to the social condition, which suggests that the differences in neural patterns observed indeed reflect social processes.

Cognitive control network: Within the cognitive control network, the analysis on the social trials showed that difference in offers correlated significantly with classification accuracy in the left dorsolateral prefrontal cortex (dlPFC; $r = 0.44$, $p = 0.01$, $p_{adjusted} = 0.03$, FDR corrected at $p = 0.05$) , the right dlPFC ($r = 0.41$, $p = 0.02$, $p_{adjusted} = 0.04$, FDR corrected at $p = 0.05$), and the posterior cingulate cortex (PCC; $r = 0.46$, $p = 0.008$, $p_{adjusted} = 0.03$, FDR corrected at $p = 0.05$). For the non-social trials no significant correlations were found for any of the regions extracted from the cognitive control network. These findings indicate that neural patterns in the bilateral dlPFC and the PCC differentiate between prososcial and selfish players and that this effect is specific to the social condition. This suggests that prosocial, as opposed to selfish participants, do not need cognitive control to overcome selfish impulses and give high offers in the social condition as they empathize with their *human* opponents, resulting different patterns in the cognitive control regions. However, in the non-social condition ToM processes are not engaged so cognitive control processes are similarly engaged for prosocial and selfish players resulting in no difference in patterns.

As the size of the ROIs (number of voxels) varied within and between the two networks we also tested whether differences in the size of the ROIs between the two masks might have had an effect on the correlation between difference in offer and classification accuracy. We computed the correlation between these two measures and found no significant correlation ($r= -0.25$, $p = 0.21$). In addition, no significant correlation was found between classification accuracy and the size in voxels of the ROIs used ($r = -0.04$, $p = 0.84$).

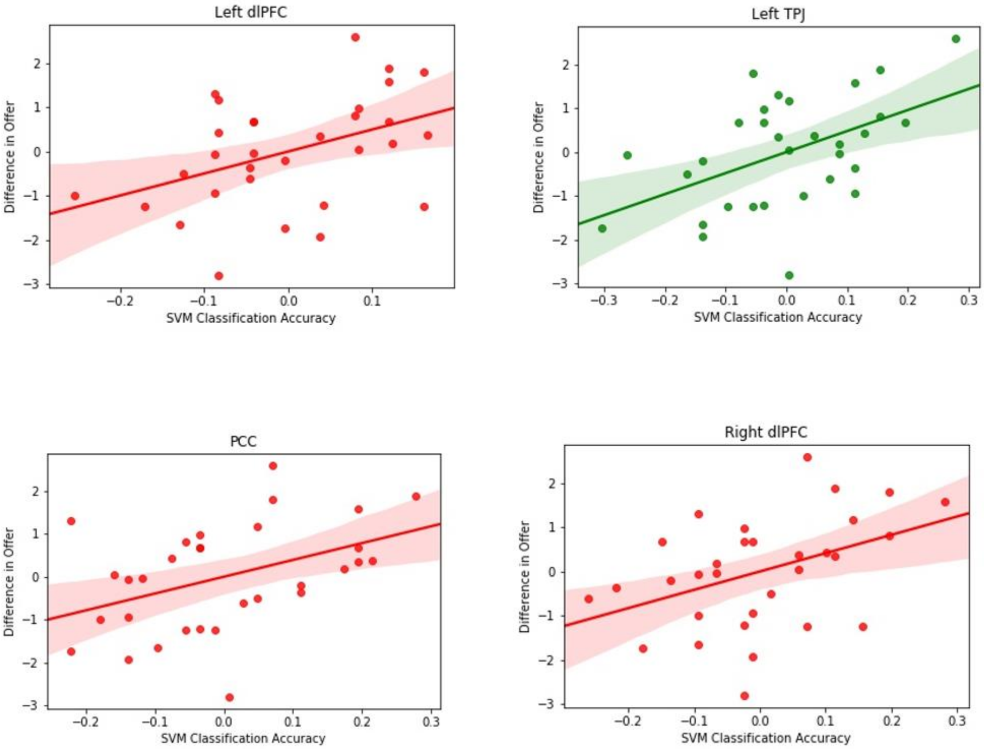


Figure 6.7. Correlation between classification of game type accuracy in the right dlPFC, left TPJ, PCC and left dlPFC and average difference in offer (UG-DG) between games across participants while controlling for difference in cohorts.

6.4.2.3 Behavioral-neural classification searchlight

The searchlight analysis conducted on trials from the social condition revealed that classification accuracy for decoding neural patterns associated with UG and DG trials correlated significantly with individual differences in offer size between games in the left TPJ, the right TPJ, the left middle temporal gyrus (MTG), the medial prefrontal cortex (MPFC), and the left inferior frontal gyrus (IFG). For selfish players, the support vector machines better dissociated game type in these regions (only clusters with size of >15 voxels are reported here, see figure 6.9; for all significant clusters see Appendix 6.3, Table S7). The fact that some of the regions (R TPJ & MPFC) were found to show significant correlation using the searchlight approach but not in the previous ROI analysis may be due to the higher specificity of the searchlight approach. Whereas in the ROI analysis uninformative voxels may have been included, the searchlight assesses differences in patterns in highly localized neighborhoods of a very small number of voxels. The effects observed here may be specific to only specific parts of these regions, which may also explain the small size of the clusters. In addition, a significant cluster in the right occipital cortex was found, which most likely is due to the different colors (red & blue) used to demarcate an UG or DG trial during the decision phase.

We repeated the analysis for the non-social condition and found no significant correlations in any other region (for significant clusters see Appendix 6.3, Table S8). Comparing the correlations between classification accuracy and behavior in the social and non-social condition in these regions directly after performing a Fisher-z transformation we found that the correlations are indeed significantly higher in the social condition in all reported areas.

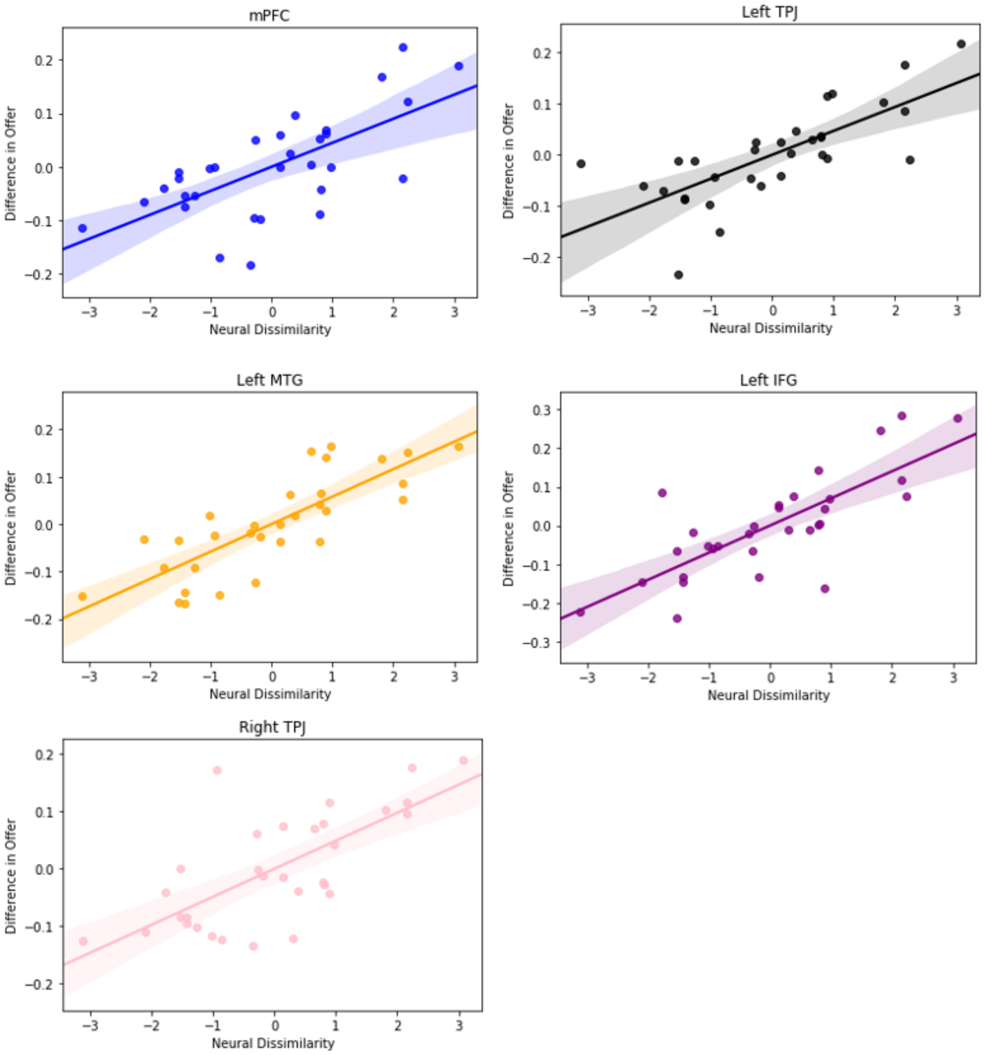


Figure 6.8. The correlations between classification accuracy and the average difference in offers (UG-DG) between games in the MPFC (blue), the left TPJ (black), the left MTG (yellow), the left IFG (purple), the right TPJ (pink).

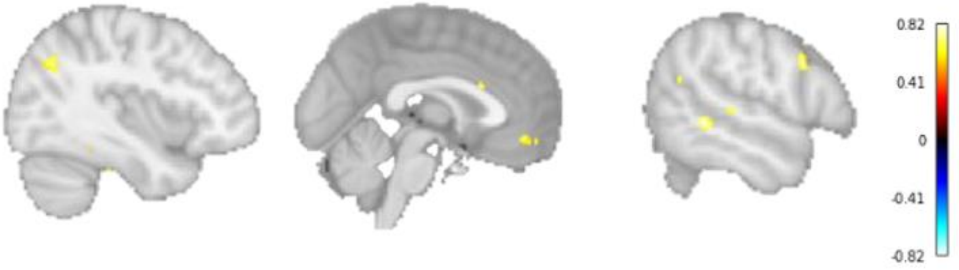


Figure 6.9. The classification searchlight analysis reveals clusters in the left TPJ (left), MPFC (middle) and right MTG and IFG (right) exhibiting significant correlation between difference in offers between UG and DG and classification accuracy on game type.

6.5 Discussion

In order to investigate whether prosocial versus selfish motivations can be decoded from neural activation patterns we conducted an experiment in which participants played the UG and DG against humans and computers while in an MRI scanner. Our behavioral results indicate that there are large individual differences in offers between the UG and DG, demonstrating that there is strong heterogeneity in motivations.

In the literature both ToM and the cognitive control processes have been proposed to underlie individual differences in prosociality in economic exchanges. Therefore, we wanted to test whether the relationship between differences in neural patterns and behavior can be attributed to either the cognitive control or the ToM network specifically or whether both networks contribute to the individual differences in prosociality. To this end, we identified brain regions associated with ToM and cognitive control using large-scale automated meta-analysis, to increase our confidence that the selected regions indeed represent the hypothesized psychological mechanisms, and thus reducing reverse inference problems (Poldrack, 2006).

Our classification analysis on the obtained ToM and cognitive control networks revealed that classification accuracy correlated significantly with difference in offers between games in both of these networks. These results indicate that neural patterns in regions in the ToM network (particularly in the TPJ) as well as in the cognitive control network (particularly in the dlPFC and PCC) appear to dissociate between prosocial and selfish players, suggesting that more selfish players engage these processes differently between games, while more prosocial players employ ToM and cognitive control similarly in both UG and DG. A whole brain classification searchlight confirmed a significant correlation, specifically in the social condition, between neural activation patterns and classification accuracy in the bilateral TPJ, but also in the MPFC and in the left MTG; all areas associated with ToM (for meta analysis see Schurz et al., 2014; Young et al., 2010).

Neuroimaging research has identified the ToM network as the basic system that facilitates social understanding (Saxe & Wexler, 2005; Schurz et al., 2014; Young et al., 2010). In this network, the TPJ has not only been consistently associated with reorienting of attention and perspective taking (Gallagher & Frith, 2003.; Krall et al., 2015) but also with appreciating differences between one's own and others perspectives in cases of conflict between those perspectives (Hétu et al., 2012; Overwalle & Baetens, 2009). Downregulating the TPJ using TMS impaired participants' ability to take into account another person's intention in moral judgements (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). Similarly, neuroimaging research has found the MTG to be preferentially activated for inferring others' beliefs (Zaitchik et al., 2010; Bruneau et al., 2012). A meta-analysis of 40 imaging studies also found the MPFC to be a core area in social mentalizing responsible for inferring others' dispositions in order to make accurate predicitions of their next moves (see review Van Overwalle & Baetens, 2009, Schurz et al., 2014). In addition, Krause and colleagues (2012) have found that for more empathic individuals deep rTMS stimulation of the MPFC disrupted affective ToM performance, confirming the prominence of the MPFC in mentalizing processes.

In the UG and DG the TPJ, MTG, and MPFC may thus be involved in orienting the proposer's attention to the responder and understanding her intentions and desires for both prosocial and strategic reasons. Our findings suggest that prosocial participants may have been more inclined to consider the welfare of their opponents and utilizing ToM processes equally in both games. In contrast, selfish players may have selectively and strategically engaged ToM to infer the expectations of their opponents in the UG to avoid financial punishment, whereas the opponents' desires were irrelevant in the DG rendering inferences about their expectations redundant.

However, our findings also emphasize the involvement of cognitive control processes in social decisions. We found higher classification accuracy in the dlPFC, PCC, and IFG when decoding game type for selfish players than for prosocial players. The dlPFC has been frequently implicated in cognitive control to overcome impulsive behavior (Dalwani et al., 2011; Weygandt et al., 2013), wheras the IFG and PCC have been consistently associated with the inhibition of predominant responses (Verbruggen & Logan, 2008; Sharp et al., 2010; Stokes et al., 2011; Wager et al., 2005).

This suggests that prosocial participants, concerned with the welfare of their opponents, intuitively gave fair offers in both the UG and DG and did not require cognitive control or response inhibition during the decision process. Selfish players, in contrast, needed to exert cognitive control in order to overcome the selfish motivation of giving a low offer in the UG, which was necessary to avoid punishment and to maximize financial gains. Further, to inhibit this predominant selfish response and make a more generous offer, response inhibition was crucial. However, in the DG, where no punishment for selfishness was expected, the selfish player could rely on his intuitions and give low offers to maximize his monetary reward, thus not requiring cognitive control and response inhibition processes. Collectively, our findings highlight the importance of both self-control processes, reflected by activity patterns in the dlPFC, the IFG and the PCC, and the crucial involvement of empathy and ToM, represented by neural patterns in the TPJ, MPFC and the MTG.

To validate that our findings are indeed of a *social* nature and are not driven by more general processes related to response inhibition, avoidance of punishment or reward sensitivity we ran several manipulation checks contrasting the social condition (human opponents) against our control condition in which participants were playing against a computer algorithm. We found that in the social condition participants gave significantly higher offers than in the control condition and this effect was significantly stronger for prosocial players as compared to players with more selfish motivations. This implies that particularly the prosocial participants care about the welfare of their *human* opponents, as for them the increase due to the social context is strongest. Furthermore, a supplementary univariate analysis showed that there was higher activation in the social condition than in the control condition in the MPFC, Precuneus and left TPJ, which have all have been associated with empathy and perspective taking, as mentioned above (see Appendix 6.4, Table S9 for list of significant clusters). In addition, when running our classification analysis on the ToM and cognitive control network, as well as on individual ROIs within those networks, no significant correlations were found between behavior and classification accuracy for the non-social condition. Collectively, these results provide evidence that participants indeed empathized with their *human* opponents and took their desires into consideration.

In the literature on individual differences in prosocial behavior during economic decision-making there has been a disconnect between behavioral research promoting the role of ToM and empathy and most of the neuroimaging research focusing on self-control in social norm-compliance. On the one hand, it has been shown that empathy is the strongest predictor of prosocial sharing (Edele et al., 2013). Further, prosocial participants have been found to hold more accurate beliefs about their opponents offers in the UG and DG, unless these beliefs become instrumental to maximizing financial gains, in which case the selfish players perform equally well (Artinger et al., 2014). On the other hand, neuroscientific evidence converges on the crucial involvement of cognitive control of the impulsive pursuit of self-interest (Strang et al., 2015; Spitzer et al., 2007, Steinbeis, Bernhardt & Singer, 2012). Further, Spitzer and colleagues (2007) showed that increased activity in the dlPFC, assumed to reflect cognitive control, was associated with higher offers in the UG.

Our findings bridge these two streams of research by highlighting the importance of both ToM and cognitive control processes. We propose that individual differences in the tendency to engage ToM may determine whether or not a person acts prosocially or selfishly. Based on this assumption a person who intuitively considers the welfare of the opponent will not require cognitive control to make fair offers regardless of whether a punishment threat is in place or not. In contrast, an individual who is less inclined to empathize may use ToM strategically to avoid punishment and may thus require cognitive control to overcome intuitive selfish impulses. Some individuals may have a natural tendency to empathize whereas others use empathy more strategically. In economic exchanges such as the UG and DG this tendency for empathy may determine whether players are prosocial or follow strategic motives.

Due to the fact that multivariate neural patterns do not provide information about the strength of activation but only indicate that information is *encoded differently* in these regions we cannot exclude the possibility that the effect goes in the opposite direction: prosocial proposers may have exerted cognitive control and response inhibition in both games and strategic players may have only done so in the UG and not in the DG, which would have also resulted in higher differences in patterns in the cognitive control network for selfish players. This interpretation would suggest that all participants are intuitively selfish and that the prosocial proposers overcome their selfish impulses in both games to make a fair offer. Although possible, this alternative explanation seems less compelling as it does not offer any explanation for why a prosocial participant would decide to exert cognitive control to make a high offer and not just selfishly make a low offer in the DG.

The same reasoning applies to our findings regarding the theory of mind network. Potentially, prosocial individual did not engage theory of mind processes in either of the games and selfish individuals only did so in the Ultimatum Game, resulting in the pattern of effects we observed. This interpretation would explain why selfish participants give high offers in the UG, but it does not offer an explanation as to why prosocial players decided to give high offers in any of the games. In sum, while unlikely, we cannot rule out the alternative explanations with complete confidence, our results clearly demonstrate that not only cognitive control processes but also theory of mind processes differentiate between selfish and strategic players.

To conclude, we were able to dissociate prosocial from selfish players based on multivariate neural patterns during decision making. Taking advantage of the higher sensitivity of multivariate techniques and meta-analytically defined networks associated with ToM and cognitive control, our study is the first to suggest that individual differences in prosociality are associated with differences in how cognitive control and ToM processes are engaged in the decision-making process. Our study bridges the disconnect between previous neuroimaging research focusing on the role of cognitive control and behavioral research promoting the importance of ToM in prosocial behavior. Highlighting the involvement of ToM processes in prosocial decisions, it thus contributes to a deeper understanding of the underlying neural mechanisms of human prosocial behavior. Our results suggest that in order to promote prosocial behavior in social and economic exchanges not only fostering cognitive control capacities but also increasing the propensity to empathize and engage ToM processes is required.

Appendix Chapter 6

Appendix 6.1 – Regions extracted for ROI analyses

Depicted here are the tables showing the regions extracted from Neurosynth.

Table S1. ToM mask link for download

Network	**Studies**	**Date of**	**Link to download**
ToM	181	28.02.2019	http://neurosynth.org/analyses/terms/theory%20mind/
Cognitive Control	598	03.02.2019	http://neurosynth.org/analyses/terms/cognitive%20control/

Table S2. Regions extracted from ToM mask

Region	**Cluster ID**	**x**	**y**	**z**	**Peak Stat**	**Cluster Size (mm3)**
TPJ L	1	-54	-54	22	15	10088
TPJ R	2	58	-54	20	15	12008
Precuneus	3	-2	-56	40	14	10816
Temporal Pole R	4	54	4	-32	11	13088
MPFC	5	4	58	24	11	15376
Temporal Pole L	6	-56	4	-28	11	13728
Cerebellum R	7	30	-82	-38	10	2328
MPFC	8	6	52	-16	9	4120
IFG R	9	58	28.0	8	8	1872

*Based on meta-analysis of 181 studies. The activation survives FDR correction ($P < 0.05$) for multiple comparisons. Locations coordinates are in stereotactic MNI space with 2x2x2 voxel size. The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. TPJ=

temporoparietal junction, MTG= middle temporal gyrus, MPFC= medial prefrontal cortex, IFG = Inferior frontal gyrus, R= right; L = left; R= right; L = left;

Table S3. Regions extracted from Cognitive Control mask

Region	Cluster ID	X	Y	z	Peak Stat	Cluster Size (mm3)
PCC	1	4	-26	32	8	848
dlPFC R	2	46	18	32	8	2032
dlPFC L	3	-44	22	24	7	1520
Insula R	4	30	22	2	7	1048
lPFC R	5	32	54	10	6	664

*Based on meta-analysis of 181 studies. The activation survives FDR correction ($P < 0.05$) for multiple comparisons. Locations coordinates are in stereotactic MNI space with 2x2x2 voxel size. The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. PCC= posterior cingulate cortex, dlPFC= dorsolateral prefrontal cortex, lPFC= lateral prefrontal cortex, R= right; L = left;R= right; L = left;

Appendix 6.2 - Tables for Behavioral Manipulation Checks

Table S4. Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.882	0.778	0.767	1.401

Table S5. ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	399.2	3	133.077	67.78	< .001
	Residual	113.9	58	1.963		
	Total	513.1	61			

Table S6. Coefficients of the Regression Model

Model		B	SE B	β	t	p
1	(Intercept)	6.727	0.467		14.395	< .001
	Diff_OfferAll_UG_DG_Human	-0.461	0.070	-0.576	-6.585	< .001
	opp	3.725	0.661	0.647	5.635	< .001
	int	-0.358	0.099	-0.472	-3.618	< .001

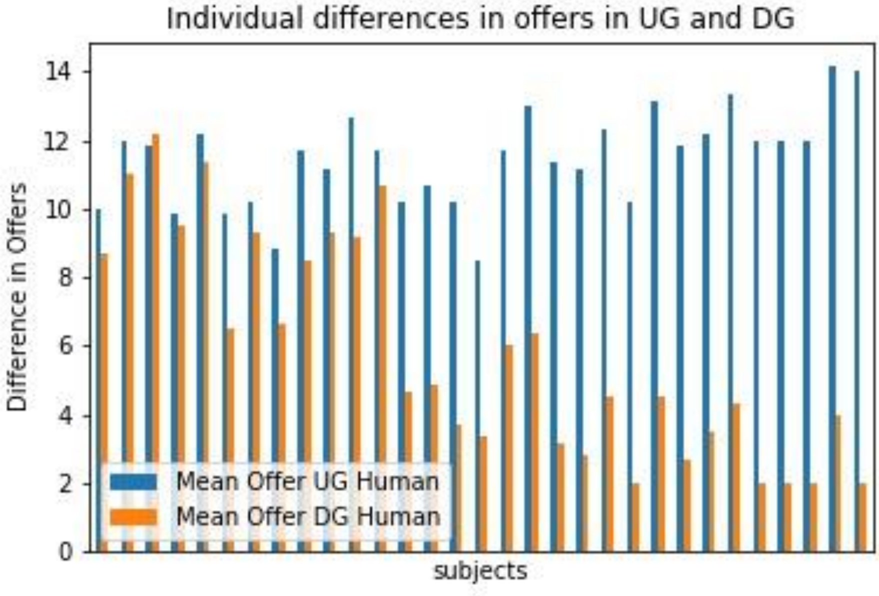


Figure S1. Individual differences in average offers in the UG and DG illustrating that smaller differences in offers between games indeed reflect high offers in both games.

Appendix 6.3 - Classification searchlight analyses

Depicted here are the significant clusters from the classification searchlight analysis.

Table S7. Cluster for behavior-neural classification searchlight analysis social condition.

Region	Cluster Index	Voxels	Max r	x	y	z
TPJ L	70	42	0,731	-38	-66	36
MTG L	69	42	0,819	-56	-44	-2
MPFC	68	33	0,695	0	44	-10
IFG L	67	30	0,743	-54	12	30
MTG L	65	20	0,67	-66	-30	0
TPJ R	63	16	0,718	56	-54	20
Occipital R	62	16	0,699	40	-90	12

*The activation survives FDR correction (P < 0.05) for multiple comparisons. Locations coordinates are in stereotactic MNI space with 2x2x2 voxel size. The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. TPJ= temporoparietal junction, MTG= middle temporal gyrus, MPFC= medial prefrontal cortex, IFG = Inferior frontal gyrus, R= right; L = left;

Table S8. Cluster for behavior-neural classification searchlight analysis non-social condition

Region	Cluster Index	Voxels	Max r	x	y	z
Occipital L	20	23	0,672	34	-94	20
SMG	19	17	0,726	-54	-40	38
Temporal sulcus	18	16	0,698	-48	-36	4
Cerebellum R	17	6	0,677	14	-76	-16
Cerebellum R	16	3	0,552	42	-30	-46

*The activation survives FDR correction (P < 0.05) for multiple comparisons. Locations coordinates are in stereotactic MNI space with 2x2x2 voxel size. The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. SMG= supramarginal gyrus, R= right; L = left;

Appendix 6.4 - Univariate fMRI analysis for manipulation checks

Preprocessing Univariate Analyses. We preprocessed the fMRI images using standard software (SPM12, Wellcome Department of Cognitive Neurology, London, UK). To correct for head motion, the functional images were realigned to the mean image. Functional images were slice-time corrected, coregistered to the anatomical image, spatially normalized to the Montreal Neurological Institute (MNI) template and lightly smoothed with a Gaussian kernel (3 × 3 × 3 mm full width at half maximum). Whole-brain activation analysis was conducted with SPM12.

For neural activation, we estimated for each participant a general linear model (GLM) using regressors for the decision phase of the UG trials, DG trials and for keypresses. The duration of the epoch for the decision phase was 6s and onset times were determined by counting back 6s from the point in time when the participant had indicated his choice. This decision phase was used as it provides all the necessary information to make the decision and is free of brain activity related to motor responses. Average background signal and white matter signal were entered as regressors of no interest. All regressors were convolved with the canonical hemodynamic response function. Linear contrasts were created between decision in UG trials versus decision in DG trials and were subjected to a random effects analysis to compute main effects (one-sample t test), and to regression analyses with behavioral data (mean difference in offers between games) as regressors. More specifically, we contrasted the activation in the decision phase in the UG against baseline, the activation in the DG against baseline and the UG against the DG in both directions and we also entered the difference in offers between games as covariate on the second level to investigate the correlation between behavior and neural activation in the contrasts of interest. Further, linear contrasts were added between decision in human trials (with human opponent) versus decision in computer (computer as opponent) and were subjected to a random effects analysis to compute main effects (one-sample t test). Based on the resulting beta images, second-level random-effects group contrast maps were then created in both directions (i.e. positive and negative correlation between activation and average difference in offers between games). The threshold applied to the statistical maps was a voxel-wise intensity of $p < 0.001$ (uncorrected) with a false discovery rate of $p < 0.05$ (FDR) correction at the cluster level where the whole brain served as the volume of interest. Clusters of activation resulting from the thresholding were characterized in terms of their peak voxels in the MNI coordinate space.

fMRI results for UG vs. DG trials. None of the univariate analyses contrasting the UG and DG trials resulted in any significant clusters. Further, adding difference in offers on the second level also did not reveal any significant effects.

fMRI results for social vs. non-social trials. As part of the Manipulation checks we also tested whether there were differences in neural activation between the social and nonsocial condition. In order to do so, we ran a univariate analyses in which we contrasted activity during decision phase of social trials (human opponent) with activity in the decision phase of the non-social (computer opponent) trials (see above for Methods). Here, we wanted to confirm whether participants indeed make their offers in based on *social* considerations or whether underlying mechanisms represent more general processes related to response inhibition and avoidance of punishment. Univariate analyses were used to allow inferences about the directionality of the effects. In other words, we wanted to test whether certain regions are more strongly activated in the social condition than in the non-social condition in order to establish that the effects found using multivariate methods indeed reflect social processes. The whole brain analysis revealed that the medial prefrontal cortex [MPFC, MNI: 0 56 -7; extent threshold pFDR <.0001 (punc. < .001, height threshold); z = 6.12; k = 226] the inferior frontal gyrus [IFG, MNI: 54 32 11; extent threshold pFDR <.0001 (punc. < .0001, height threshold); z = 5.62; k = 199] and the Precuneus [MNI: 0 -58 21; extent threshold pFWE <.0001 (punc. < .0001, height threshold); z = 5.04; k = 272] were significantly stronger activated in human trials than in computer trials. For all significant clusters, please see Table S9.

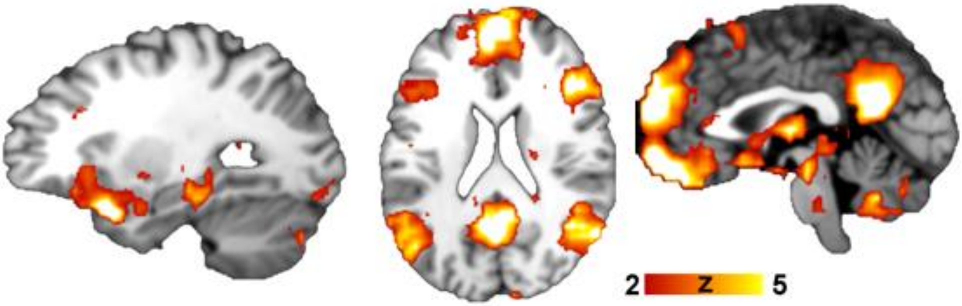


Figure S2. Showing the effect of the univariate analyses for the social vs non-social condition. The left image shows the activation in the IFG, the right image shows activation in the Precunues/PCC and MPFC and the middle image also shows the activations in precuneus MPFC and IFG.

Table S9. Showing clusters exhibiting a higher activation during social trials as compared to non-social trials

Region	pFWE	pFDR	Voxels	p(unc)	T	Z	x	y	z
Fusiform R	<0.001	<0.001	766	<0.001	10.32	6.69	42.00	-49.00	-20.86
Temp.Pole L	<0.001	<0.001	594	<0.001	9.29	6.33	-27.00	11.00	-20.86
vmPFC	<0.001	<0.001	226	<0.001	8.76	6.12	0.00	56.00	-6.82
IFG R	<0.001	<0.001	199	<0.001	7.58	5.62	54.00	32.00	10.73
MTG L	<0.001	<0.001	92	<0.001	7.49	5.58	-54.00	-16.00	-10.33
mPFC	<0.001	<0.001	303	<0.001	7.09	5.39	0.00	56.00	17.75
Caudate R	<0.001	<0.001	115	<0.001	6.54	5.12	9.00	8.00	3.71
Precuneus L	<0.001	<0.001	272	<0.001	6.40	5.04	0.00	-58.00	21.26

* Clusters of activation resulted from the thresholding at voxel level at p<0.001 and with false discovery rate of p<0.05 at the cluster level where the whole brain served as the volume of interest. Locations coordinates are in stereotactic MNI space with 2x2x2 voxel size. The source of anatomical labels: AAL Atlas tools cross referenced with Neurosynth. TPJ= temporoparietal junction, MTG= middle temporal gyrus, vmPFC= ventromedial prefrontal cortex, IFG = Inferior frontal gyrus, mPFC= medial prefrontal cortex, R= right; L = left;

Testing whether differences in behavior can be explained by different levels of engagement.

To test whether participants who change their offer more between UG and DG (selfish participants) are more engaged in the task than participants who give a similar response in the games (more prosocial participants), we checked whether reaction times correlate with difference in offers across participants. Presumably, participants who are less engaged spend less time thinking about their offer and would show shorter RTs. We found no significant correlation between prosociality and RTs (r=-0.32, p=0.074). The negative correlation even suggests that more prosocial players take a bit more time to respond, which would suggest that prosocial players rather than selfish players deliberate longer and might be more engaged, even if they do not change their offer between games. Further, if participants who were more selfish were simply more engaged in the task than the prosocial participants we would expect the more selfish participants to also exhibit greater difference in activation in the ToM network and the cognitive control network between the social and the non-social condition than the prosocial participants. In order to test this, we conducted univariate analyses in which we tested whether individual differences in offers between games (our measure of prosociality) correlates with differences in neural activation between social and non-social trials. The whole brain analysis (with difference in offer as covariate) revealed that there were no significant differences in neural activation between social and non-social trials between prosocial and selfish players in the cognitive control or ToM network. Significant differences were only found in the supplementary motor area. This further provides evidence that prosocial and selfish players do not seem to differ with regard to their levels of engagement as their ToM and cognitive control networks did not show differential activation during task performance

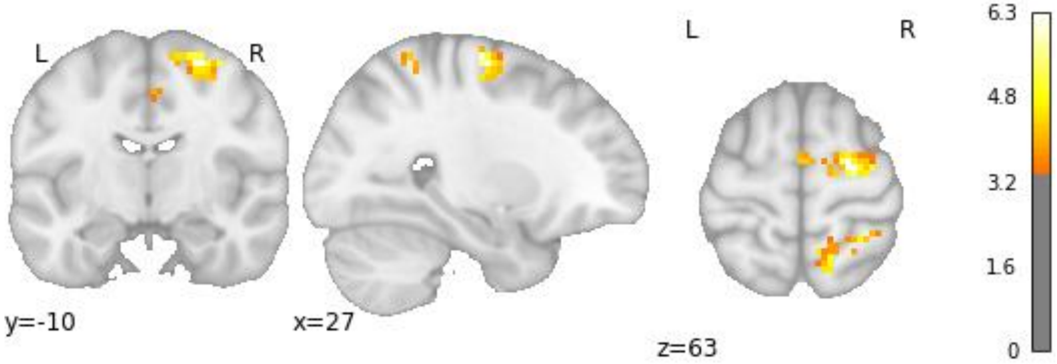


Figure S3. Correlation between difference in offer and the social vs non-social contrast. Maps are FWE corrected at 0.05 for multiple comparisons.

Chapter 7

Conclusion

In this dissertation, I investigated the neurocognitive mechanisms underlying moral decisions and how they differ across individuals. The empirical studies have revealed that there are large individual differences in (dis)honesty and fairness and that this heterogeneity is reflected in differences in the neural underpinnings of moral decisions. Specifically, my research has revealed that: 1) the role of cognitive control depends on a person's moral default (fMRI study in Chapter 2), and the same electrophysiological signature evoked by cognitive control demands in the Stroop task can be used to estimate (dis)honest choices in an independent cheating task, providing converging evidence that cognitive control is indeed contingent on an individual's inclination to cheat (EEG study in Chapter 3); 2) inhibiting the habitual response to cheat differs from inhibiting a default of behaving honestly (fMRI study in Chapter 4); 3) the brain at rest can be used to predict cheating long before the decision is made (rsfMRI study in Chapter 5); and 4) different fair versus selfish motivations can be decoded from distributed neural patterns (fMRI study in Chapter 6). In this final chapter, I will discuss the theoretical and societal contributions of the insights gained, point to the limitations, and outline future research directions.

7.1 Contributions to the social neuroscience of fairness and (dis)honesty

This dissertation provides three contributions to the literature on moral decision-making. Firstly, it provides insights into the neural mechanisms underlying individual differences in fairness and honesty. These insights also offer a reconciliation of the longstanding controversy on the role of cognitive control on moral decision-making. Secondly, we identify stable neural markers of (dis)honesty that can be used to predict an individual's propensity to cheat in the absence of a temptation to cheat. Thirdly, we developed a novel paradigm that pioneers in measuring cheating, repeatedly, voluntarily, spontaneously, and inconspicuously in the scanner or with EEG on a trial-by-trial basis. Below these three contributions are elaborated on.

7.1.1 Reconciling the debate between will and grace

There has been a long-standing debate among scientists and philosophers concerning the cognitive nature of (dis)honesty: Is honesty the natural tendency of people, termed the Grace hypothesis, or does it demand willpower in form of cognitive control to overrule a default inclination for dishonesty, named the Will hypothesis.

In *Chapters 2, 3, 4* and *6* we aimed at reconciling this long-standing debate, by identifying the neural mechanism underlying decisions to be honest and fair.

Synthesizing the findings across chapters, a common pattern emerges: Humans are distributed along a continuum, from individuals who are honest and fair in our experiments to people who behave dishonestly and selfishly (see Figure 1). Participants on one side of the continuum have a default inclination to be honest and fair which is associated with more self-referential thinking when given the opportunity to cheat and more theory of mind processes when interacting with someone else. This is corroborated by research showing that perspective-taking and empathy promote prosocial behavior (e.g. Eisenberg and Miller, 1987; Batson et al., 1991; Penner et al., 2005; Pavey et al., 2012; Edele et al., 2013; Artinger et al., 2014). It also dovetails with early neuroimaging research on moral judgements and dilemmas emphasizing the role of moral intuition (Moll et al., 2002, Greene et al., 2001).

Conversely, people on the other side of the continuum have a default tendency for dishonesty and selfishness, and their decisions appear to be motivated more strongly by (monetary) rewards. This observation is supported by research showing that greedy people find moral transgressions more permissible and act unethically more frequently as compared to less greedy individuals (Seuntjens et al, 2019). It also aligns with previous neuroimaging showing stronger neural activation in anticipation of reward has been linked to more (dis)honesty (Abe & Greene).

To strike and maintain a subjectively justifiable balance where one can occasionally benefit from cheating and selfishness yet still preserve a positive self-image, people on both ends of the continuum sometimes need to overcome their predominant impulses and default behavior (Mazar et al., 2008). A typically honest and fair person will need to override the default of being honest to occasionally benefit from cheating, whereas a cheater needs to suppress the predominant selfish response in order to occasionally be honest and maintain their self-concept. In the interaction with others, an honest and fair person automatically engages in theory of mind and splits equally. In contrast, a selfish person needs to exert cognitive control to inhibit the default inclination to keep everything to themselves, and instead engages theory of mind to realize that an approximately equal split is necessary to appease the opponent and avoid being punished. Thus, it appears that the effect of cognitive control depends on our moral default. For honest and fair people the Grace hypothesis applies: honesty and fairness result from the absence of temptation, and response inhibition is needed to cheat. In contrast, for cheaters the predictions of the Will hypothesis apply, and active resistance of temptation in the form of inhibition is needed to be honest/fair. Expanding on insights from cognitive psychology to the social/moral domain, our findings suggest that cognitive control appears to serve the purpose of overriding our default behavior. Notably, our research also demonstrates that while overriding our moral default to be honest (nonhabitual cheating) recruits the same brain region as resisting the default to cheat in favor of honest (nonhabitual honesty), the neural patterns within this region differ. This suggests that cognitive control with regard to moral decisions is context specific rather than domain general.

Across my empirical chapters I demonstrate that cognitive control processes can serve different purposes for different individuals in the context of (dis)honesty and fairness. This dissertation therefore contributes to the reconciliation of the controversy on the role of cognitive control in moral decision-making.

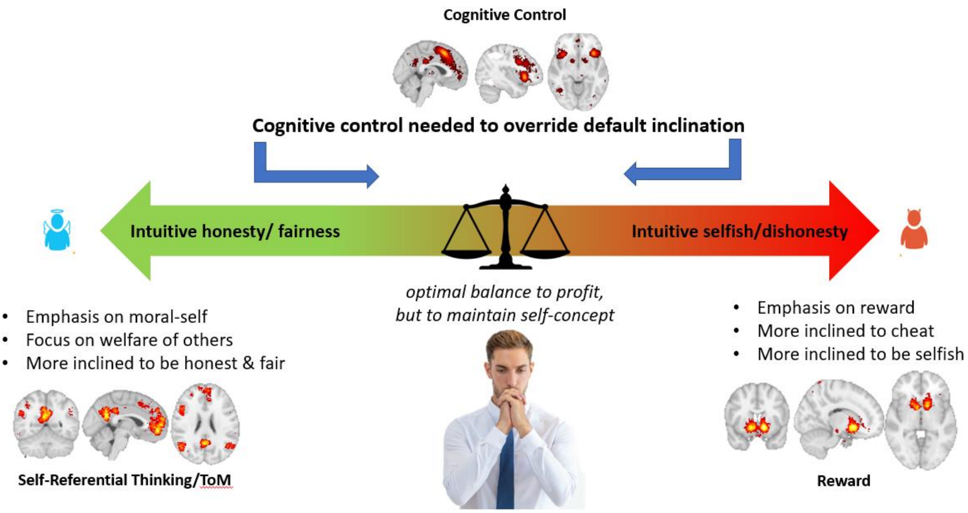

Cognitive Control
Cognitive control needed to override default inclination
Intuitive honesty/ fairness
Intuitive selfish/dishonesty
optimal balance to profit, but to maintain self-concept
Emphasis on moral-self
Focus on welfare of others
More inclined to be honest & fair
Emphasis on reward
More inclined to cheat
More inclined to be selfish
Self-Referential Thinking/ToM
Reward

Figure 1. People are distributed along a continuum from intuitively fair and honest to intuitively selfish and dishonest. On the left side of the continuum people have a stronger focus on the moral self-concept and on the welfare of others, while on the right-side participants focus more on reward. Cognitive control is needed to override the default in both cases to achieve a subjectively justifiable balance.

7.1.2 Identifying stable neural markers of (dis)honesty

Whereas *Chapters 2, 3 & 4* focus on investigating the neural mechanisms while (dis)honest decisions are made, *Chapter 5* contributes by identifying stable neural patterns that can be used to predict the propensity to cheat long before the decision is made. As has become clear throughout the dissertation, the conflict between the temptation to cheat for financial gains and maintaining a positive image of ourselves as being a 'good person' is omnipresent. It is obvious that some individuals are more successful than others in maintaining this balance. However, because the underlying mechanisms of socially undesirable traits and behaviors are difficult to measure, it remains poorly understood whether there are stable markers of these individual differences in (dis)honesty.

In Chapter 5 we explored whether the brain's functional connectome at rest could be used to reliably predict the propensity to cheat in the absence of any temptation to cheat. We combined resting state fMRI data from four independent samples with connectome-based predictive modelling (CPM) and the Spot-The-Difference task that allows measuring spontaneous and voluntary cheating behavior inconspicuously. This combination enabled us to not only investigate which task-independent functional patterns promote honesty but also to explore which functional connections contribute most to predicting (dis)honesty in unseen participants from an independent sample.

Our results indicate that resting-state functional connectivity between brain networks linked to self-referential thinking and reward processing, can reliably and accurately predict, in independent samples from different scanners and cities, whether a participant is honest, or a cheater. Participants with a higher propensity to cheat also scored higher on a range of well-established self-report impulsivity measures, which emphasizes the ecological validity of our cheating task. Notably, when comparing neural and questionnaire measures, the neural measures were found to be substantially more important at predicting (dis)honesty out-of-sample. The use of a regularized regression model, trained and tested on independent samples, enables more conservative and accurate inference of the brain-behavior relationship and thus increases the likelihood to generalize and successful replication. The high predictive accuracy, produced by our models, combined with the reliable and task-independent nature of resting state functional connectivity, provides an excellent avenue for the development of useful neuroimaging-based biomarkers of dishonesty or other socially undesirable behaviors. This combination of techniques may be particularly useful for studying socially undesirable behaviors since it offers unobstructed insights into their neurocognitive determinants which would be more easily concealed in self-report measures.

Regarding the development of neuroimaging-based biomarkers, the current study also adds practical utility as resting-state scans are inexpensive as compared to task-based fMRI studies, as they are much shorter on average. It has to be noted, however, that while this chapter contributes by providing a significant first step in developing a biomarker for (dis)honesty, it is my opinion that the technique presented here is not yet suitable for use as lie detector in courts or a screening tool in companies. This is because accuracy is currently too low to rule out false positives, which could have detrimental effects on applicants' careers or a defendant in court.

7.1.3 Measuring cheating on a trial-by-trial basis: The Spot-The-Difference task

As introduced in Chapter 1, methodological challenges have so far prevented to measure cheating without identifiable victim on a trial-by-trial basis in the scanner. While some tasks, such as the die-roll task and the coin-flip task have been adapted to allow measuring (dis)honesty in the MRI scanner, (dis)honesty can only be inferred on the aggregate level. As a result, until now it was not possible to investigate what happens in the brain of generally honest participants when they decide to override their moral default to cheat or, conversely, when a cheater decides to be honest.

The third contribution of this dissertation was to fill this gap and develop a task that allows to investigate how cognitive control processes resolve the conflict between reward and self-concept on a trial-by-trial basis.

Using a paradigm originally proposed by Gai & Puntoni (2021), I created and validated a large set of stimuli and added several task features such as an increasing reward structure to encourage cheating and different levels of difficulty in combination with point-and-click trials to reduce suspicion (as explained in the empirical chapters). This resulted in a task, in which participants were able to cheat repeatedly, deliberately, and voluntarily inside the MRI scanner or with EEG without suspicion of the real purpose of the experiment. The important advantage of the Spot-The-Difference task is that we can directly measure on which trials the participants cheated, enabling us to study within subject variation in moral decisions and its neural underpinnings. Stated differently, this allows us to not only observe the neural underpinnings of (dis)honesty in general, but also explore the neural processes that underlie overriding one's moral default. That is, we can explore the neural patterns associated with an honest decision made by cheaters or the neural correlates of an honest person’s nonhabitual dishonesty. Notably, due to our sophisticated cover story, that we are investigating the neural mechanisms underlying visual processing, participants are under the impression that the experimenter does not know that they are cheating, which is crucial considering the social desirability bias and the reactance effect introduced in Chapter 1.

To conclude, the Spot-The-Difference paradigm is therefore a task that measures (dis)honest behavior inconspicuously on a trial-by-trial basis enabling us to study individual differences in neurocognitive processes underlying cheating behavior. This task may prove useful to further investigating the neural correlates of (dis)honesty and may also be helpful in assessing the influence of interventions aimed at reducing unethical behavior and how the effect of these interventions differ across individuals.

7.2 Practical contributions to society

There are immense economic costs caused by dishonest and selfish behavior, such as tax evasion, insurance fraud, music piracy or business scandals, so finding effective ways to reduce dishonest behavior are of great relevance to policy makers. Also, during the COVID 19 pandemic, in particular in the hectic early months of the pandemic, dishonesty in the form of companies selling low quality face masks or committing fraud on governmental subsidies were highly prevalent, which highlights the timeliness and relevance of our findings. Dishonesty also pervaded daily behaviors such as people not reporting corona symptoms when entering restaurants or theatres, or fraudulently reporting corona symptoms to get a ‘free’ corona test thus increasing the waiting lines and costs and even negatively affecting the testing of urgent corona cases. By understanding how different neurocognitive processes determine honesty or dishonesty in different individuals, these insights can prove instrumental in the development of more efficient strategies to reduce dishonesty and strengthen trust in society.

An important contribution of this dissertation is that different individuals need different interventions as opposed to a one-size-fits-all solution. For example, an intervention such as having people think hard about the consequences of their actions only works for a subset of individuals (the cheaters and selfish individuals). That is, thinking hard about the consequences may increase their cognitive control and thus they may indeed decide against their natural inclination to cheat and be more honest. For others (the more fair and honestly inclined), thinking about the opportunity to cheat may actually increase dishonesty: their default is to be honest, but if they start considering the option to cheat, they may decide that it occasionally makes sense to do so as it can be profitable.

Taken together, the empirical studies suggest that there are three processes that influence (dis)honest and (un)fair behavior, namely cognitive control, self-referential thinking and reward. Therefore, interventions targeted at reducing (dis)honesty should focus on these three processes.

First of all, our research clearly demonstrates that increasing cognitive control capacities in people with an inclination to cheat is an effective strategy to make them more honest. Previous research has shown that an effective way to increase cognitive control is by means of precommitment (Prelec, 1989; Schelling, 1992; Strotz, 1956; Thaler & Shefrin, 1981; Wertenbroch, 1998). Precommitment can be understood as a voluntary restriction of access to temptation. For instance, people who want to diet, but recognize that a chocolate mousse will be hard to resist, can avoid temptation by going to a restaurant with a less alluring menu. Notably, it has been found that precommitment is particularly useful for impulsive individuals. More impulsive individuals are more aware of their self-control problems and also show the highest preference for precommitment (Ariely & Wertenbroch, 2002; Soutschek & Tobler, 2020). Neuroimaging research supports these findings by showing that precommitment increases functional connectivity in the cognitive control network and this effect is strongest for impulsive individuals (Crockett et al., 2013). Therefore, precommitment could serve as a powerful tool to design interventions to reduce (dis)honesty. In *Chapter 5* we find that more impulsive individuals cheat more. So, offering options to precommit may effectively reduce the temptation to cheat and enable cheaters to remain honest.

In a science context, for example, academic institutions may use productivity (i.e., the number of studies published & the quality of the journals) as the most important metric for evaluating the quality of scholars. This creates a lot of pressure for scholars to publish significant effects which may seduce them to engage in p-hacking or fabrication of data to advance their academic careers. Precommitment in the form of pre-registration, where the design and analysis plan of study is detailed beforehand, may be an effective way of reducing such unethical behaviors.

In a business context, dishonesty is highly prevalent and costly (as introduced in *Chapter 1*). It is not uncommon that employees are exposed to a moral conflict between selfishly pursuing their own (financial) interests and being honest in favor of their client. Brokerage companies may best exemplify these moral conflicts. Even though brokers ought to act in their client's best interest, their own reward is contingent on commissions which provides ample opportunities to prioritize personal reward over the client's interest. Brokers may buy or sell at a point of time that results in larger commissions for themselves, advise to buy stocks or funds that are in line with their own financial interest but not their client's benefit, or abuse knowledge and information about a large impending order (Davis, 2004; McDonald, 2002; Mazar & Ariely, 2006). Precommitment may preempt these opportunities by forcing brokers to share all information about their own commissions related to the options presented to the customers or giving brokers the option to switch to a fixed salary rather than a commission based one.

It is also well known that consumers can behave unethically (Bagozzi 1995; Vitell 2003). One highly costly example is wardrobing, where clothing is returned after being used. Precommitment could be used to prevent wardrobing. Companies may reduce the temptation to return used items, by giving costumers a discount if they decide to limit themselves to a shorter return time, constraining the temptation to engage in wardrobing or even to precommit to keeping a purchased item.

Our findings also suggest that in order to promote fair and honest behavior in social interactions, not only fostering cognitive control capacities but also increasing the propensity to empathize and consider one's moral self-image is required. Policy makers should consider designing interventions that elicit the feelings and thoughts of others and bring to mind one's own moral values. Given that empathy is malleable and highly sensitive to situational forces (Ickes, Gesn, & Graham, 2000; Klein & Hodges, 2001; Thomas & Maio, 2008) several interventions have been developed in this regard. Most empathy interventions aim at fostering people's capacity to empathize by targeting their ability to experience empathy or expressing empathy to interaction partners. Specifically, experience-based interventions are based on tasks that promote tuning in to others' internal states. Examples of experience-based interventions are attending mindfulness and compassion-based meditation classes for several weeks (Condon et al., 2013, Jazaieri et al., 2015), role playing personal interactions (Jacobs, 1977), simulating experiences of a target group (Soble et al, 2011; Clore & Jeffery, 1972).

In contrast, expression-based interventions focus on developing skills to communicate empathy more effectively to a target. Examples of expression-based interventions are computer-based training in recognizing emotions in faces and voices (Golan & Baron-Cohen, 2006), or empathy-relational skills training focused on underlying neurobiology of empathy (Riess et al., 2012). Recently it has been found, that using virtual reality as a tool to teach empathy and perspective taking may further boost the effectiveness of these experience and expression-based interventions (Herrera et al., 2018). In addition, a social-psychological intervention focuses on fostering empathy by changing the views of the self, the perceived social norm or shifting people's construal of empathy-evoking situations (Weisz & Zaki, 2017). These interventions seek to improve empathic effort by teaching people that empathy is malleable, can be developed and can help to feel good about oneself thus increasing their willingness to connect to others.

Taken together, these tools can be easily implemented by managers to enhance empathy and consequently increase fairness and honesty amongst employees. For instance, brokerage companies may benefit from implementing experience-based interventions where brokers learn to take the perspective of their clients, by simulating an exchange with another broker while they are taking the role of a client.

As mentioned above, the research in this dissertation suggests that making one's own moral self-concept more salient may also promote honesty and fairness. This is in line with the previous studies, suggesting that activating and emphasizing one's moral standards reduces moral transgressions (Mazar et al., 2008, Omir et al, 2018). Specifically, research has suggested that reading the ten commandments (Mazar et al., 2008), signing an ethics statement (Shu et al., 2012), subconscious ethical priming (Welsh and Ordonez, 2014), listening to religious music (Nichols et al., 2020) or observing extreme (religious) rituals (Mitkidis et al., 2017) may increase honesty and fairness. These moral reminders and priming are expected to reduce unethical behavior by activating moral standards that individuals already possess rather than altering moral values, thus not interfering with an individual's autonomy. While these, sometimes subtle, moral reminders and primes may not deter calculating and callous corporate criminals who do not abide by generally accepted ethical norms, such reminders may implicitly influence typical employees who unconsciously neglect moral values when making decisions in the financially motivated context of business. It has to be noted, however, that some of this research has failed to replicate (Kristal et al., 2020; van Elk et al., 2015; Verschuere et al., 2018). Thus, future research is required to test whether this offers a promising avenue for designing interventions to reduce dishonesty and selfishness.

Finally, interventions targeted at reward are theoretically most straightforward: To reduce cheating and unfairness, one has to make the incentive to cheat less attractive (and simultaneously increasing the perceived probability of getting caught and punished). In practice this is often complicated as (financial) rewards are also necessary to motivate employees, scholars, or students. As mentioned above, in the example of the brokers, it may foster more fairness and honesty to switch from a commission based to a fixed salary, but this may also reduce the motivation of the broker. A potential way to increase honesty and fairness may be to explicitly incentivize these behaviors. In the academic context, scholars could be rewarded for pre-registrations, for example, by changing journal policies to give preference to articles that have precommitted on their design and analysis and making their data and code available. In addition, hiring committees could explicitly state that scholars who engage in these open science practices are favored.

7.3 Limitations and avenues for future research

In each chapter, I already pointed out limitations and suggestions for further research. Here, I want to point out two main limitations. A first limitation of the current work is that participants who cheated or were selfish on our task may not act similarly in different contexts. Participants' (dis)honesty or (un)fairness, as measured in our tasks, may thus not reflect a stable moral trait that generalizes to different moral contexts. Our resting state fMRI, however, seems to support the notion that the observed individual differences in honesty may generalize, as it indicates that (dis)honesty, as observed in our task, is represented in stable functional connections at rest, within the self-referential thinking network and between the self-referential thinking and the reward network. Resting-state functional connectivity has been found to be robust over time, which suggests that the observed correlation with cheating behavior may generalize as well. In addition, individual differences in cheating were observed to be strongly correlated with stable personality characteristics related to impulsivity, which has been linked to other forms of dishonesty such as academic cheating (Anderman et al., 2009). While this provides some initial evidence for generalizable moral inclinations, much more research is needed to resolve this concern. It is also important to note that the rsfMRI study was also conducted only using the Spot-The-Difference task, which reduces the evidence for generalizability. An important next step will be to investigate to what extent cheating in the Spot-The-Difference task correlates with dishonesty in other tasks such as the die-roll-task, coin-flip task or sender receiver game introduced in *Chapter 1*. In addition, future research could combine different behavioral paradigms measuring (dis)honesty and fairness with neuroimaging to investigate how moral decisions and their neural correlates differ across different contexts of honesty and fairness. As a next step, it would be relevant to study the underlying neural mechanisms in the domain of honesty/fairness in comparison to the other 4 moral domains: Harm/Care, Ingroup/Loyalty, Authority/Respect and Purity/Sanctity (Graham et al., 2011).

A second main limitation pertains to the correlational nature of neuroimaging methods. Although neuroimaging can provide insights into which neural processes are associated with (dis)honest and (un)fair decisions it cannot determine whether there is a causal relationship. To make inferences about the causality of the interaction effect of cognitive control on honesty and fairness, transcranial direct current stimulation (tCDS) or transcranial magnetic stimulation (TMS) could be used to directly manipulate activity in the cognitive control network in honest participants and cheaters to test whether the interaction effect can be causally induced. These techniques rely on a noninvasive form of brain stimulation in which a changing magnetic field is used to cause electric current at a specific area of the brain through electromagnetic induction. Relatedly, the causal relationship between moral default, cognitive control and honesty/fairness could be investigated by means of behavioral manipulations to induce cognitive load or evoke acute stress test (Wood et al., 1984), which have been found to reduce cognitive control capacity (Shiv & Fedorikhin, 1999; Schwabe & Wolf, 2009; Yu, 2016). Since stress has been found to impair cognitive control capacity and enhance intuitive response tendencies, we would expect that it enhances a person's moral default. With regard to the effect of stress on (dis)honesty, preliminary data from my study combining the Spot-The-Difference task with a task inducing acute stress, found that stress indeed increases cheating in cheaters, whereas it makes honest participants even more honest. Thus, stress seems to indeed amplify a person's moral default.

A promising avenue for further neuroimaging research may be to investigate to what extent the moral default is determined by nature as opposed to nurture. Stated differently, are we born a 'cheater', or is our moral default mainly a result of our socialization? Our rsfMRI study on adults suggests that there are stable neural patterns that are indicative of a person's propensity for (dis)honesty. This raises the question as to when these moral propensities arise. Future research may attempt to answer this question by collecting rsfMRI data from children and adolescents throughout their youth in combination with several tasks measuring (dis)honesty and fairness and personality measures.

7.4 Concluding remark

The endeavor to gain insights into the neural underpinnings of moral-decision making is still in its infancy. While respecting ethical and technical challenges and obstacles, the author is convinced that social neuroscience with a focus on morality will continue to produce discoveries that will enlighten us about human social life and cooperation and also may provide highly useful insights for policy makers. This dissertation constitutes one step in this long journey ahead.

References

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M., & Fujii, T. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex*, 16(2), 192. 199

Abe, N., & Greene, J. D. (2014). Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *Journal of Neuroscience*, 34(32), 10564-10572.

Abe, N. (2009). The neurobiology of deception: Evidence from neuroimaging and loss-of function studies. *Current Opinion in Neurology*, 22(6), 594-600.

Abe, N. (2020). Overriding a moral default for honesty or dishonesty. *Proceedings of the National Academy of Sciences*, *117*(36), 21844-21846.

Abraham, A., Dohmatob, E., Thirion, B., Samaras, D., & Varoquaux, G. (2014a). Region segmentation for sparse decompositions: better brain parcellations from rest fMRI. *arXiv preprint arXiv:1412.3925*.

Accenture (2003), “One-Fourth of Americans Say it’s Acceptable to Defraud Insurance Compa-nies,” February 12, (accessed December 1, 2006), [available at http://www.accenture.com/xd/xd.asp?it=enweb&xd=_dyn%5Cdynamic pressrelease_5 77. xml].

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3-4), 323-338.

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. *Journal of Cognitive Neuroscience*, *21*(6), 1179-1192.

Amemori, K. I., Amemori, S., Gibson, D. J., & Graybiel, A. M. (2020). Striatal Beta Oscillation and Neuronal Activity in the Primate Caudate Nucleus Differentially Represent Valence and Arousal Under Approach-Avoidance Conflict. *Frontiers in Neuroscience*, *14*, 89.

Anderman, E. M., Cupp, P. K., & Lane, D. (2009). Impulsivity and academic cheating. *The Journal of Experimental Education*, *78*(1), 135-150.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464-477.

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience*, *6*(2), 115-116.

Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, *13*(3), 219-224.

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In: *Advances in experimental social psychology* (Vol. 4, pp. 1-34). Academic Press.

Artinger, F., Exadaktylos, F., Koppel, H., Sääksvuori, L. (2014). In others' shoes: Do individual differences in empathy and theory of mind shape social preferences? *PLoS One*, 9(4), e92844.

Ashburner, J. and Friston, K.J.. (2005). Unified segmentation. *NeuroImage*, 26:839–851

Aurich, N. K., Alves Filho, J.O., Marques da Silva, A.M., Franco, A.R.. Evaluating the Reliability of Different Preprocessing Steps to Estimate Graph Theoretical Measures in Resting State fMRI data. *Frontiers in Neuroscience,* 2015;9:48.

Ballard, K., & Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. *NeuroImage, 45*(1), 143–150.

Baddeley, A. D., & Della Sala, S. (1996). Working memory and executive control. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1397-1404.

Bagozzi, Richard P. (1995), Reflections on Relationship Marketing in Consumer Markets, *Journal of the Academy of Marketing Science*, 23 (4), 272–77.

Bahnemann, M., Dziobek, I., Prehn, K., Wolf, I., & Heekeren, H. R. (2009). Sociotopy in the temporoparietal cortex: common versus distinct processes. *Social Cognitive and Affective Neuroscience*, *5*(1), 48-58.

Ballard, K., & Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. *NeuroImage*, *45*(1), 143-150.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7), 462.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, 412–427.

Batson, C.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*, 61(3), 413.

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

Barnes, C. M., Schaubroeck, J., Huth, M., & Ghumman, S. (2011). Lack of sleep and unethical conduct. *Organizational Behavior and Human Decision Processes*, *115*(2), 169-180.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., ...& Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron, 85*(2), 390-401.

Baumeister, Roy F. (1998), “The Self,” in Handbook of Social Psychology, Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, eds. New York: McGraw-Hill, 680-740.

Baumeister, R. F. (2010). *The self*. Oxford University Press.

Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., ...& Silvia, P. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, *115*(5), 1087-1092.

Becker, G. S. (1968). Crime and punishment: An economic approach. In: The economic dimensions of crime (pp. 13-68). *Palgrave Macmillan, London*.

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

Bem, D. J. (1972). Self-perception theory. In: *Advances in experimental social psychology* (Vol. 6, pp. 1-62). Academic Press.

Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. *Current Opinion in Psychology, 6,* 195-198.

Berthoz, S., Grèzes, J., Armony, J. L., Passingham, R. E., & Dolan, R. J. (2006). Affective response to one's own moral violations. *NeuroImage*, *31*(2), 945-950.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, 16.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K. R. (2007). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal processing magazine*, *25*(1), 41-56.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84-88.

Brehm, J. W. (1989). Psychological reactance: Theory and applications. *ACR North American Advances*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

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*, 50(2), 219–31.

Bucciol, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, *32*(1), 73-78.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860-1862.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*; 14(5):365–76.

Caballero-Gaudes, C. & Reynolds, R. C. Methods for cleaning the BOLD fMRI signal. *NeuroImage* 154, 128–149 (2017).

Cai, H., Zhu, J., & Yu, Y. (2020). Robust prediction of individual personality from brain functional connectome. *Social cognitive and affective neuroscience*, 15(3), 359-369.

Campbell, E. Q. (1964), “The Internalization of Moral Norms,” *Sociometry*, 27 (4), 391-412.

Cao H, Plichta MM, Schäfer A, Haddad L, Grimm O, Schneider M, Esslinger C, Kirsch P, et al. (2014) Test–retest reliability of fmri- based graph theoretical properties during working memory, emotion processing, and resting state. *NeuroImage* 84:888–900.

Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Economics Letters*, *158*, 54-57.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics*, *79*, 93-99.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., ... & Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience, 8,* 20.

Carp J. On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neurosci*ence. 2012;6(149).

Carter, C. S., & Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 367-379.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, *49*(2), 220-238.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414-421.

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex*, 19(7), 1557-1566.

Christie, R. & Geis, F. (1970). Studies in Machiavellianism. NY: Academic Press.

Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., & Walter, H. (2007). The intentional network: how the brain reads varieties of intentions. *Neuropsychologia*, *45*(13), 3105-3113.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178-1198.

Clore, G. L., & Jeffery, K. M. (1972). Emotional role playing, attitude change, and attraction toward a disabled person. *Journal of Personality and Social Psychology*, *23*(1), 105– 111. doi:10.1037/ h0032867

Cochran, J. K., Wood, P. B., Sellers, C. S., Wilkerson, W., & Chamlin, M. B. (1998). Academic dishonesty and low self-control: An empirical test of a general theory of crime. *Deviant Behavior*, *19*(3), 227-255.

Cohen, M. X., & Cavanagh, J. F. (2011). Single-trial regression elucidates the role of prefrontal theta oscillations in response conflict. *Frontiers in Psychology*, *2*, 30.

Cohen, M. X., Ridderinkhof, K. R., Haupt, S., Elger, C. E., & Fell, J. (2008). Medial frontal cortex and response conflict: evidence from human intracranial EEG and medial frontal cortex lesion. *Brain Research*, *1238*, 127-142.

Condon, P., Desbordes, G., Miller, W. B., & DeSteno, D. (2013). Meditation increases compassionate responses to suffering. *Psychological Science*, 24(10), 2125– 2127.

Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2013). Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *9*(8), 1175-1184.

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

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)“brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261-270.

Crockett, M. J., Braams, B. R., Clark, L., Tobler, P. N., Robbins, T. W., & Kalenscher, T. (2013). Restricting temptations: neural mechanisms of precommitment. *Neuron*, 79(2), 391-401.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*, *11*(1), 126.

Cyders, M. A., Littlefield, A. K., Coffey, S., & Karyadi, K. A. (2014). Examination of a short English version of the UPPS-P Impulsive Behavior Scale. *Addictive Behaviors*, *39*(9), 1372-1376.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., ... & Alzheimer's Disease Neuroimaging Initiative. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, *192*, 115-134.

Dalwani, M., Sakai, J.T., Mikulich-Gilbertson, S.K., et al. (2011). Reduced cortical gray matter volume inmale adolescents with substance and conduct problems. *Drug and Alcohol Dependence*, 118(2–3), 295–305.

Davis, Ann (2004), “Open Secrets; Head of the Line: Client Comes First? On Wall Street, It Isn’t Always So; Investing Own Money, Firms Can Misuse Knowledge of a Big Impending Order; Mischief in the ‘Back Books,’” *The Wall Street Journal*, (December 16), A1.

Dhar, R., & Wertenbroch, K. (2012). Self-signaling and the costs and benefits of temptation in consumer choice. *Journal of Marketing Research*, *49*(1), 15-25.

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135-168.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment*, *Research & Evaluation*, 14(20), 1-11

Dockès, J., Poldrack, R., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., ... & Varoquaux, G. (2020). NeuroQuery: comprehensive meta-analysis of human brain mapping. *arXiv preprint arXiv:2002.09261*.

Donoho, D.L., Maleki, A., Rahman, I., Shahram, M., Stodden, V. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering.* 2009;11(1):8–18.

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.

Edgeworth, F. (1881). *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: C. K. Paul.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. Dryden Press.

Eisenberg, N., Miller, P.A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, 101(1), 91–119.

Engelmann, J. B., Damaraju, E., Padmala, S., & Pessoa, L. (2009). Combined effects of attention and motivation on visual task performance: Transient and sustained motivational effects. *Frontiers in Human Neuroscience*, *3*, 4.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149.

Esteban, O., Zosso, D., Daducci, A., Bach-Cuadra, M., Ledesma-Carbayo, M. J., Thiran, J. P., & Santos, A. (2016). Surface-driven registration method for the structure-informed segmentation of diffusion MR images. *NeuroImage*, *139*, 450-461.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A. & Oya, H. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111-116.

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). NeuroImage Searchlight analysis : Promise , pitfalls, and potential. *NeuroImage*, *78*, 261–269.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–68.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., ... & Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11), 1664-1671.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525-547.

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347–69.

Gabrieli, J.D., Ghosh, S.S. & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, 85, 11–26

Gai. P. J., & Puntoni, S. (2017). *Does language shape dishonesty? Evidence from a spot-the-difference task.* Poster presented at the Society for Judgment and Decision Making Annual Conference, Vancouver, Canada.

Gai, P. J., & Puntoni, S. (2021). Language and Consumer Dishonesty: A Self-Diagnosticity Theory. *Journal of Consumer Research* 48 (2), 333–351.

Gallagher, H.L., Frith, C.D. (2003). Functional imaging of ‘theory of mind’. *Trends in Cognitive Sciences*, 7(2), 77–83.

Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping*, 28(12), 1287-1301.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, *19*(12), 1727.

Gash, J. (1992). *The Great California Game: A Lovejoy Mystery.* Penguin books.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1.

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191-203.

Gino, F., & Ariely, D. (2012). The dark side of creativity: original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*(3), 445.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384-394.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review, 108*(2), 419-53.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 5-13.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience, 7*, 267.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*(30), 12506-12511.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517-523.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4.

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing*.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L 1-penalized estimation. *Statistics and Computing*, *24*(2), 137-154.

Guilford, J. P. (1967). The nature of human intelligence. *Cambridge University Press.*

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(7), 4259-4264.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157-1182.

Güth,W., Schmittberger, R., Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367–88.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., Pollmann, S. (2009). PyMVPA: A Unifying Approach to the Analysis of Neuroscientific Data. *Frontiers in Neuroinformatics*, *3*(February), 3. https://doi.org/10.3389/neuro.11.003.2009

Hanslmayr, S., Pastötter, B., Bäuml, K. H., Gruber, S., Wimber, M., & Klimesch, W. (2008). The electrophysiological dynamics of interference during the Stroop task. *Journal of Cognitive Neuroscience*, *20*(2), 215-225.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2), 73-78.

Hétu, S., Taschereau-Dumouchel, V., Jackson, P.L. (2012). Stimulating the brain to study social interactions and empathy. Brain Stimulation, 5(2), 95–102.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65-70.

Henrich, J., Boyd, R., Bowles, S., et al. (2005). “Economic man” in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.

Herrera, F., Bailenson, J., Weisz, E., Ogle, E., & Zaki, J. (2018). Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one*, *13*(10), e0204494.

Herwig, U., Baumgartner, T., Kaffenberger, T., Brühl, A., Kottlow, M., Schreiter-Gasser, U., ... & Rufer, M. (2007). Modulation of anticipatory emotion and perception processing by cognitive control. *Neuroimage*, *37*(2), 652-662

Hilbig, B. E., & Thielmann, I. (2017). Does everyone have a price? On the role of payoff magnitude for ethical decision making. Cognition, 163, 15–25.

Hoffman, E., McCabe, K., Smith, V.L. (2000). The impact of exchange context on the activation of equity in ultima- tum games. *Experimental Economics*, 3(1), 5–9.

Hocevar, D. (1979). The Development of the Creative Behavior Inventory (CBI).

Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P., & Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, *32*(3), 401-414.

Ickes, W., Gesn, P. R., & Graham, T. (2000). Gender differences in empathic accuracy: Differential ability or differential motivation? *Personal Relationships*, 7, 95– 109.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2(8):e124.

Jacobs, D. (1977*). Role playing— Role reversal as a method for training adolescents in empathy*. Retrieved on 9/ 715 from http:// www.ruor.uottawa.ca/ handle/ 10393/ 8194

Jazaieri, H., McGonigal, K., Jinpa, T., Doty, J. R., Gross, J. J., & Goldin, P. R. (2014). A randomized controlled trial of compassion cultivation training: Effects on mindfulness, affect, and emotion regulation. *Motivation and Emotion*, 38(1), 23–35. http:// doi.org/ 10.1007/ s11031- 013- 9368- z

Klein, K. J. K., & Hodges, S. D. (2001). Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin*, 27(6), 720– 730.

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417-429.

Kahneman, D., Knetsch, J. L., Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, 59, 5285–5300.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, 303(5660), 1023-1026.

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends inCognitive Sciences*, *18*(4), 163-166.

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.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, *21*(16), RC159-RC159.

Koles, Z. J., Lazar, M. S., & Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain topography*, *2*(4), 275-284.

Kovacevic, S., Azma, S., Irimia, A., Sherfey, J., Halgren, E., & Marinkovic, K. (2012). Theta oscillations are sensitive to both early and late conflict processing stages: effects of alcohol intoxication. *PLoS One*, *7*(8), e43957.

Kozel, F. A., Laken, S. J., Johnson, K. A., Boren, B., Mapes, K. S., Morgan, P. S., & George, M. S. (2009). Replication of functional MRI detection of deception. *The open forensic science journal*, 2, 6

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science*, *14*(5), 778-796.

Krall, S.C., Rottschy, C., Oberwelland, E., et al. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function*, 220(2), 587–604.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–3868.

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, 117(13), 7103-7107.

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., ... & Childress, A. R. (2002). Brain activity during simulated deception: an event-related functional magnetic resonance study. *Neuroimage*, 15(3), 727-732.

Li N, Ma N, Liu Y, He X-S, Sun D-L, Fu X-M, Zhang X, Han S, et al. (2013) Resting-state functional connectivity predicts impulsivity in economic decision-making. *Journal of Neuroscience*. 33:4886–4895.

Lisofsky, N., Kazzer, P., Heekeren, H. R., & Prehn, K. (2014). Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. *Neuropsychologia*, *61*, 113-122.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539-550.

Loeber, R., Menting, B., Lynam, D. R., Moffitt, T. E., Stouthamer-Loeber, M., Stallings, R., ... & Pardini, D. (2012). Findings from the Pittsburgh Youth Study: Cognitive impulsivity and intelligence as predictors of the age–crime curve. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(11), 1136-1149.

Lotz, S. (2015). Spontaneous giving under structural inequality: Intuition promotes cooperation in asymmetric social dilemmas. *PLoS ONE*, *10*(7), 1–9. https://doi.org/10.1371/journal.pone.0131562

Luck, S. J. (2005). An introduction to the event-related potential technique. Cambridge, MA: MIT press

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288(5472), 1835-1838.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163-203.

Maréchal, M. A., Cohn, A., Ugazio, G., & Ruff, C. C. (2017). Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, *114*(17), 4360-4364.

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.

Matsumoto, K., & Tanaka, K. (2004). Conflict and cognitive control. *Science*, *303*(5660), 969-970.

Mazar, N., & Ariely, D. (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy & Marketing*, 25(1), 117-126.

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503-507.

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology*, *45*(3), 594-597.

Mednick, S. A. (1968). The remote associates test. *The Journal of Creative Behavior*, *2*(3),213-214.

Meffert, H., Blanken, L., Blair, K. S., White, S. F., & Blair, J. R. (2013). The influence of valence and decision difficulty on self-referential processing. *Frontiers in Human Neuroscience*, *7*, 46.

Mensch, A., Mairal, J., Thirion, B., Varoquaux, G., 2018. Stochastic subsampling for factorizing huge matrices. *IEEE Trans. Signal Process*. 66, 113–128.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). NeuroImage Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118.

Mitkidis, P., Ayal, S., Shalvi, S., Heimann, K., Levy, G., Kyselo, M., ... & Roepstorff, A. (2017). The effects of extreme rituals on moral behavior: The performers-observers gap hypothesis. *Journal of Economic Psychology*, *59*, 1-7.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine learning, 57*(1-2), 145-175.

Moll, J., de Oliveira-Souza, R., Bramati, I. E., & Grafman, J. (2002). Functional networks in emotional moral and nonmoral social judgments. *Neuroimage*, 16(3), 696-703.

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.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636-2643.

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., ... & Yeo, B. T. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience*, *20*(3), 299-303.

Nigbur, R., Cohen, M. X., Ridderinkhof, K. R., & Stürmer, B. (2012). Theta dynamics reveal domain-specific control over stimulus and response conflict. *Journal of Cognitive Neuroscience*, *24*(5), 1264-1274.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553.

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.

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457.

Nostro, A. D., Müller, V. I., Varikuti, D. P., Pläschke, R. N., Hoffstaedter, F., Langner, R., ... & Eickhoff, S. B. (2018). Predicting personality from network-based resting-state functional connectivity. *Brain Structure and Function*, 223(6), 2699-2719.

Ogawa, S., Lee, T. M., Stepnoski, R., Chen, W., Zhu, X. H., & Ugurbil, K. (2000). An approach to probe some neural systems interaction by functional MRI at neural time scale down to milliseconds. *Proceedings of the National Academy of Sciences*, *97*(20), 11026-11031.

Oosterbeek, H., Sloof, R., Van De Kuilen, G. (2004). Cultural differ- ences in ultimatum game experiments: evidence from a meta- analysis. *Experimental Economics*, 7(2), 171–88.

Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yücel, M., & Lorenzetti, V. (2018). The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human brain mapping*, 39(8), 3398-3418.

Overwalle, F. Van, & Baetens, K. (2009). NeuroImage Understanding others ' actions and goals by mirror and mentalizing systems : A meta-analysis. *NeuroImage*, *48*(3), 564–584.

Pavey, L., Greitemeyer, T., Sparks, P. (2012). "I help because I want to, not because you tell me to": empathy increases autonomously motivated helping. *Personality and Social Psychology Bulletin*, 38(5), 681–9.

Penner, L.A., Dovidio, J.F., Piliavin, J.A., Schroeder, D.A. (2005). Prosocial behavior: multilevel perspectives. *Annual Review of Psychology*, 56, 365–92.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59-63.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press.

Prelec, D. (1989). Decreasing impatience: Definition and consequences. Unpublished manuscript, Massachusetts Institute of Technology, Cambridge, MA.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, *1*(4), 515-526.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427.

Rand, D. G. (2016). Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychological Science*, *27*(9), 1192–1206.

Rand, D. G., Kraft-Todd, G., & Gruber, J. (2015). The collective benefits of feeling good and letting go: Positive emotion and (dis) inhibition interact to predict cooperative behavior. *PLoS ONE*, *10*(1), 1–12. https://doi.org/10.1371/journal.pone.0117426

Riess, H., Kelley, J. M., Bailey, R. W., Dunn, E. J., & Phillips, M. (2012). Empathy training for resident physicians: A randomized controlled trial of a neuroscience- informed curriculum. *Journal of General Internal Medicine*, *27*(10), 1280– 1286.

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*(2), 752-763.

Rosenberg MD, et al. (2016) A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci* 19:165–171.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, 97-109

Qin, P., & Northoff, G. (2011). How is our self related to midline regions and the default-mode network? *Neuroimage*, *57*(3), 1221-1233.

Saxe, R., Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, 43, 1391–9.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological science*, *17*(8), 692-699.

Schelling, T. (1992). Self-command: A new discipline. In J. Elster & G.F. Loewenstein (Eds.), *Choice over time* (pp. 167–176). New York: Russell Sage Foundation.

Schulz, J. F., Fischbacher, U., Thöni, C., & Utikal, V. (2014). Affect and fairness : Dictator games under cognitive load. *Journal of Economic Psychology, 41*, 77–87.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of func- tional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 42, 9–34.

Seuntjens, T. G., Zeelenberg, M., Van de Ven, N., & Breugelmans, S. M. (2015). Dispositional greed. *Journal of Personality and Social Psychology*, *108*(6), 917.

Seuntjens, T. G., Zeelenberg, M., van de Ven, N., & Breugelmans, S. M. (2019). Greedy bastards: Testing the relationship between wanting more and unethical behavior. *Personality and Individual Differences*, 138, 147-156.

Scheres, A., & Sanfey, A. G. (2006). Individual differences in decision making: drive and reward responsiveness affect strategic bargaining in economic games. *Behavioral and Brain Functions, 2*(1), 1-8.

Schwabe, L., Wolf, O.T., 2009. Stress prompts habit behavior in humans. *Journal of Neuroscience*. 29, 7191–7198.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science, 23*(10), 1264-1270.

Shen X, et al. (2017) Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols.* 12:506–518.

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, *26*(3), 278- 292.

Sharp, D.J., Bonnelle, V., De Boissezon, X., et al. (2010). Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences*, 107(13), 6106–11.

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109*(38), 15197-15200.

Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, *13*(2), 121-134.

Simon, J. R., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics*, *6*(1), 99-105.

Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*. 2014;143:534–47

Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*. 2011;22(11):1359–66.

Small, D. M., Gitelman, D., Simmons, K., Bloise, S. M., Parrish, T., & Mesulam, M.-M. (2005). Monetary Incentives Enhance Processing in Brain Regions Mediating Top-down Control of Attention. *Cerebral Cortex*, *15*(12), 1855–1865.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83-98.

Soble, J. R., Spanierman, L. B., & Liao, H.- Y. (2011). Effects of a brief video intervention on white university students' racial attitudes. *Journal of Counseling Psychology*, *58*(1), 151– 157.

Soutschek, A., & Tobler, P. N. (2020). Know your weaknesses: Sophisticated impulsiveness motivates voluntary self-restrictions. Journal of experimental psychology: learning, memory, and cognition.

Speer, S. P. H., & Boksem, M. A. S. (2019). Decoding fairness motivations from multivariatebrain activity patterns. *Social Cognitive and Affective Neuroscience*, *14*(11), 1197-1207.

Speer, S. P.H., Smidts, A., & Boksem, M. A. S. (2020). Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proceedings of the National Academy of Sciences*, *117*(32), 19080-19091.

Speer, S.P.H., Smidts, A., & Boksem, M. A. S. (2021). Different neural mechanisms underlie nonhabitual honesty and nonhabitual cheating. *Frontiers in Neuroscience*, *15*, 83.

Speer, S. P., Smidts, A., & Boksem, M. A. S (2021). Cognitive control promotes either honesty or dishonesty, depending on one's moral default. *Journal of Neuroscience (in press).*

Spence, S. A., Farrow, T. F., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. (2001). Behavioral and functional anatomical correlates of deception in humans. *Neuroreport*, 12(13), 2849-2853.

Speights, David and Mark Hilinski (2005), "Return Fraud and Abuse: How to Protect Profits," *Retailing Issues Letter*, 17 (1), 1-6.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, 56(1), 185–96.

Steinbeis, N., Bernhardt, B.C., Singer, T. (2012). Article impulse control and underlying functions of the left DLPFC medi- ate age-related and age-independent individual differences in strategic social behavior. *Neuron*, 73(5), 1040–51.

Stokes, P.R.A., Rhodes, R.A., Grasby, P.M., Mehta, M.A. (2011). The effects of the COMT val108/158met polymorphism on BOLD activation during working memory, planning and response inhibition: a role for the posterior cingulate cortex? *Neuropsychopharmacology*, 36(4), 763–71.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. T. (2014). Be nice if you have to—the neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, *10*(6), 790-796.

Strotz, R.H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, 23, 165–180.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428.

Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., & Yamagishi, T. (2010). Theory of mind enhances preference for fairness. *Journal of Experimental Child Psychology*, *105*(1-2), 130-137.

Tang, D., Hu, L., & Chen, A. (2013). The neural oscillations of conflict adaptation in the human frontal region. *Biological Psychology*, *93*(3), 364-372.

Tang, T. L. P., & Chen, Y. J. (2008). Intelligence vs. wisdom: The love of money, Machiavellianism, and unethical behavior across college major and gender. *Journal of Business Ethics*, *82*(1), 1-26.

Thaler, R.H. (1988). Anomalies: the ultimatum game. *Journal of Economic Perspectives*, 2(4), 195–206.

Thomas, G., & Maio, G. R. (2008). Man, I feel like a woman: When and how gender-role motivation helps mindreading. *Journal of Personality and Social Psychology*, 95(5), 1165– 1179.

Thirion, B., Varoquaux, G., 2016a. Dictionary learning for massive matrix factorization. *Proceedings of the International Conference on Machine Learning*, pp. 1737–1746. Doi: arXiv:1605.00937v2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., ... & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. *PloS One*, *11*(3), e0152472.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Tustison, N. J. et al. N4ITK: Improved N3 Bias Correction. IEEE Transactions on Medical Imaging 29, 1310–1320 (2010).

Van Buuren, M., Gladwin, T. E., Zandbelt, B. B., Kahn, R. S., & Vink, M. (2010). Reduced functional coupling in the default-mode network during self-referential processing. *Human Brain Mapping*, *31*(8), 1117-1127.

Van Elk, M., Matzke, D., Gronau, Q., Guang, M., Vandekerckhove, J., & Wagenmakers, E. J.(2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6,* 1365.

Vanhaudenhuyse, A., Demertzi, A., Schabus, M., Noirhomme, Q., Bredart, S., Boly, M., ... & Laureys, S. (2011). Two distinct neuronal networks mediate the awareness of environment and of self. *Journal of Cognitive Neuroscience*, *23*(3), 570-578.

Vanni, S., Tanskanen, T., Seppä, M., Uutela, K., & Hari, R. (2001). Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proceedings of the National Academy of Sciences*, *98*(5), 2776–2780.

Van Overwalle, F. Van, Baetens, K. (2009). NeuroImage under- standing others' actions and goals by mirror and mentalizing systems: a meta-analysis. *NeuroImage*, 48(3), 564–84.

Van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: a functional MRI study. *Neuroimage*, *27*(3), 497-504.

Van't Veer, A., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making*, *9*(3), 199-206.

Varoquaux, G., Craddock, R.C., 2013. Learning and comparing functional connectomesacross subjects. *NeuroImage* 80, 405.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166-179.

Verbruggen, F., Logan, G.D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, 12(11), 418–24.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition*, *7*(3), 462-469.

Vitell, Scott J. (2003), "Consumer Ethics Research: Review, Synthesis and Suggestions for the Future," *Journal of Business Ethics*, 43 (1–2), 33–47.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274-290.

Wang, C., Ulbert, I., Schomer, D. L., Marinkovic, K., & Halgren, E. (2005). Responses of human anterior cingulate cortex microdomains to error detection, conflict monitoring, stimulus-response mapping, familiarity, and orienting. *Journal of Neuroscience*, 25(3), 604-613.

Wager, T.D., Sylvester, C.Y.C., Lacey, S.C., Nee, D.E., Franklin, M., Jonides, J. (2005). Common and unique components of response inhibition revealed by fMRI. *NeuroImage*, 27(2), 323–40.

Wager, T. D., Nichols, T. E., Van Essen, D. C., Poldrack, R. A., & Yarkoni, T. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670.

Weisz, E., & Zaki, J. (2017). Empathy building interventions: A review of existing work and suggestions for future directions. *The Oxford handbook of compassion science*, 205-217.

Welsh, D. T., & Ordonez, L. D. (2014). The dark side of consecutive high performance goals: Linking goal set- ting, depletion, and unethical behavior. *Organizational Behavior and Human Decision Processes*, 123, 79–89.

Wertenbroch, K. (1998). Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing Science*, 17, 317–337.

Weygandt, M., Mai, K., Dommes, E., et al. (2015). Impulse control in the dorsolateral prefrontal cortex counteracts post-diet weight regain in obesity. *NeuroImage*, 109, 318–27.

Womelsdorf, T., Johnston, K., Vinck, M., & Everling, S. (2010). Theta-activity in anterior cingulate cortex predicts task rules and their adjustments following errors. *Proceedings of the National Academy of Sciences*, *107*(11), 5248-5253.

Wood, D. L., Sheps, S. G., Elveback, L. R., & Schirger, A. (1984). Cold pressor test as a predictor of hypertension. *Hypertension*, *6*(3), 301-306

Yamagishi, T., Takagishi, H., Fermin, A.D.S.R., Kanai, R., Li, Y., Matsumoto,Y. (2016). Cortical thickness ofthe dorsolateral pre- frontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences*, 113(20), 5582–7.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large scale automated synthesis of human functional neuroimaging data. *NatureMethods, 8*(8), 665–670.

Young, L., Dodell-Feder, D., Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, 48, 2658–64.

Yu, R. (2016). Stress potentiates decision biases: A stress induced deliberation-to-intuition (SIDI) model. *Neurobiology of stress*, *3*, 83-95.

Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., Dickerson, B.C. (2010). Mental state attribution and the temporoparietal junction: an fMRI study comparing belief, emotion, and perception. *Neuropsychologia*, 48(9), 2528–36.

Zhang, H., & Mo, L. (2016). Mentalizing and information propagation through social network: evidence from a resting-state-fMRI study. *Frontiers in psychology*, *7*, 1716.

Zimmerman, G. M. (2010). Impulsivity, offending, and the neighborhood: Investigating the person–context nexus. *Journal of Quantitative Criminology*, *26*(3), 301-332.

Zuo X-N, Xing X-X (2014) Test-retest reliabilities of resting-state fmri measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews,* 45:100–118.

Zysset, S., Müller, K., Lohmann, G., & von Cramon, D. Y. (2001). Color-word matching Stroop task: separating interference and response conflict. *Neuroimage*, *13*(1), 29-36.

Zysset, S., Schroeter, M. L., Neumann, J., & von Cramon, D. Y. (2007). Stroop interference, hemodynamic response and aging: an event-related fMRI study. *Neurobiology of aging*, *28*(6), 937-946.

Summary

Dishonest and selfish behavior is highly prevalent, which is manifested in financial fraud, scientific misconduct, plagiarism and during the corona crisis in form of emergency aid fraud. However, until now, the neurocognitive processes that underlie moral decisions have remained largely elusive. The main research theme of this dissertation is to study the underlying neurocognitive mechanisms that determine the outcome of moral decisions particularly as they pertain to honesty and fairness. This dissertation aims at exploring how these neurocognitive mechanisms are expressed differently across individuals resulting in heterogeneity in moral decision making and behavior.

In this dissertation, I employ several different neuroimaging techniques, namely, resting-state and task-based functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG). in combination with different behavioral paradigms aimed at evoking voluntary and spontaneous (un)fair and (dis)honest decisions. The advantage of this approach is that one obtains direct access to the processes and motivations for socially undesirable behaviors, such as selfishness and dishonesty, which are easily concealed in self-report measures. By means of triangulating moral decisions with different neuroimaging techniques and behavioral paradigms, I am able to provide convergent evidence, thus increasing confidence in replicability.

In the first empirical chapter we investigate the neural mechanisms underlying (dis)honesty and how these mechanisms may differ across individuals. A novel behavioral paradigm was used which allows for inconspicuously measuring spontaneous cheating on a trial-by-trial basis in the MRI scanner. While replicating previous findings that greed promotes dishonesty, the analysis revealed that self-referential thinking processes drive honest behavior. Notably, I found that the cognitive control network allowed cheaters to be honest, while it enabled cheating for honest participants. Therefore, these findings suggest that cognitive control is not required to be honest or dishonest per se but that it is contingent on a person’s moral default.

The goal of Chapter 3 was to test the robustness and replicability of the findings in Chapter 2 by using a different neuroimaging technique, EEG instead of fMRI, on a different sample. In addition, the Stroop task was used, which is a well-established task to measure cognitive control in form of response inhibition, as an external localizer task to identify the neural signature of cognitive control and prevent reverse inference problems. The analyses revealed the neural signature of cognitive control over midfrontal channels in the theta range. Notably, it was observed that midfrontal theta power is linked to higher probability for cheaters to be honest and conversely a higher likelihood for generally honest participants to cheat. I therefore replicate the observation that the effect of cognitive is contingent on a person's moral default.

Chapter 4 contributes to the literature by investigating whether cognitive control processes to enable nonhabitual honesty (being honest as a cheater) differ from cognitive control processes that allow nonhabitual cheating (cheating by a generally honest person). One of the core regions of the cognitive network is the inferior frontal gyrus (IFG), which has been consistently associated with the inhibition of predominant responses (Wager et al., 2005; Verbruggen and Logan, 2008; Sharp et al., 2010; Stokes et al., 2011).

The analysis showed that, while the level of average activation across all voxels in the IFG is the same for honest participants and cheaters, when overriding their moral default, the information encoded in the distributed pattern across voxels differs. This suggests that even though individuals have idiosyncratic default responses in morally ambiguous situations, the underlying moral norm may nonetheless be universal across individuals. That is, that while people differ in their inclination for honesty or (dis)honesty (fairness or selfishness), the norm that cheating is bad may be encoded similarly in the brain across individuals.

In the fifth chapter we focused more closely on individual differences in moral behavior and tested whether there are stable dispositions towards honesty or dishonesty that are encoded in the brain. To address this question, we combined connectome-based predictive modelling on resting state functional connectivity patterns with our cheating task, to investigate whether these task-independent neural patterns can be used to predict (dis)honesty at a later point in time (several weeks). Importantly, we found that connectivity, between reward, self-referential thinking and cognitive control networks, can reliably and accurately predict, in an independent sample, whether a participant is honest, or a cheater long before the decision is made.

Lastly, in Chapter 6, we use multivariate pattern analysis (MVPA) to decode prosocial versus selfish motivations in the Ultimatum and Dictator Game. Our findings indicate that differences in fairness may be driven by variations in the degree to which self-control and theory of mind processes are engaged during decision-making. These results highlight the importance of theory of mind and empathy in social decision making and suggest that fostering the propensity to empathize may help promote prosocial behavior.

In sum, this dissertation provides three main contributions. First, it provides reconciliation for a long-standing debate in the literature on the role of cognitive control on (dis)honesty, by showing that the role of cognitive control depends on a person's moral default. Second, this dissertation contributes by identifying stable neural markers that can be used to predict (dis)honesty in the absence of the temptation to cheat. Lastly, the dissertation provides a behavioral paradigm that can be used to inconspicuously measure voluntary, spontaneous and repeated cheating on a trial-by-trial basis in the scanner or while recording EEG.

The endeavor to gain insights into the neurocognitive processes of moral decisions and how they differ across individuals is still in its infancy. I am convinced that moral psychology and social neuroscience will continue to provide meaningful insights in how we decide on moral matters and that these insights will prove helpful in designing interventions that target the processes underlying (dis)honest and (un)fair decisions, such as interventions based on precommitment, strengthening empathy and the elicitation and incentivization of moral values.

Dutch Summary

Oneerlijk en zelfzuchtig gedrag komt veel voor. Dat uit zich in belastingontduiking, wetenschappelijk wangedrag, plagiaat, en nu tijdens de coronacrisis, in de vorm van fraude bij overheidssteun. Maar de neurocognitieve processen die aan de basis liggen van morele beslissingen zijn grotendeels ongrijpbaar gebleven. Het thema van dit proefschrift is het bestuderen van de onderliggende neurocognitieve mechanismen die morele beslissingen bepalen, vooral wat betreft eerlijkheid en billijkheid. Het doel van dit proefschrift is verkennen hoe deze neurocognitieve mechanismen verschillen van persoon tot persoon, wat leidt tot heterogeniteit in morele beslissingen en gedrag.

Ik gebruik verschillende beeldvormingstechnieken, in het bijzonder functionele magnetische resonantie imaging (fMRI) en elektro-encefalografie (EEG), in combinatie met verschillende gedragstaken die gericht zijn op het uitlokken van vrijwillige en spontane (on)billijke en (on)eerlijke beslissingen. Het voordeel van deze aanpak is dat er zo rechtstreekse toegang verkregen wordt tot de processen en beweegredenen voor sociaal ongewenst gedrag, zoals egoïsme en oneerlijkheid, die gemakkelijk verborgen kunnen worden als we mensen ernaar vragen. Door een driehoek te vormen tussen morele beslissingen, verschillende beeldvormingstechnieken en gedragsparadigma's, kan ik convergent bewijs leveren, wat het vertrouwen in repliceerbaarheid vergroot.

In het eerste empirische hoofdstuk onderzoek ik de neurale mechanismen die aan de basis van (on)eerlijkheid liggen, en hoe deze mechanismen kunnen verschillen van persoon tot persoon. Ik heb een nieuwe gedragstaak ontwikkeld waarmee spontaan bedrog onopvallend gemeten kan worden in de MRI-scanner. De analyse repliceerde eerdere bevindingen dat hebzucht oneerlijkheid bevordert en dat introspectie en zelfreflectie aan de basis liggen van eerlijk gedrag. Een nieuwe bevinding was dat bedriegers via het cognitieve controlenetwerk eerlijk konden zijn, terwijl eerlijke deelnemers hiermee in staat waren te bedriegen. Deze bevindingen suggereren dus dat cognitieve controle op zich geen vereiste is om eerlijk of oneerlijk te zijn, maar dat dit afhangt van iemands morele standaard.

Het doel van hoofdstuk 3 was om de robuustheid en repliceerbaarheid van de bevindingen in hoofdstuk 2 te testen via een andere beeldvormingstechniek, EEG in plaats van fMRI. Daarnaast is de Stroop-taak gebruikt. Dat is een veel gebruikte taak om cognitieve controle te meten. Deze taak is gebruikt als een externe lokalisatie-taak om het neurale substraat van cognitieve controle te identificeren. Hiermee kunnen problemen door omgekeerde inferentie verkleind worden. Uit de EEG-analyses bleek dat de neurale signatuur van cognitieve controle over midfrontale kanalen zich in het thèta-bereik bevond. Midfrontale thèta-power is gelinkt aan een hogere waarschijnlijkheid dat bedriegers een keer eerlijk zijn en omgekeerd ook aan een hogere waarschijnlijkheid dat mensen die in het algemeen eerlijk zijn, een keer bedriegen. Ik repliceer hiermee dus de bevinding dat het effect van cognitieve controle afhangt van iemands morele standaard.

Hoofdstuk 4 draagt bij aan de literatuur door te onderzoeken of cognitieve controleprocessen die niet-habituele eerlijkheid mogelijk maken (een bedrieger die soms eerlijk is) verschillen van cognitieve controleprocessen waarmee niet-habitueel bedriegen mogelijk wordt (bedrog door een over het algemeen eerlijk persoon). Een van de kerngebieden van het cognitieve controle netwerk is de *gyrus frontalis inferior*, die consistent geassocieerd wordt met de remming van predominante reacties.

Uit de analyse bleek dat, hoewel het niveau van gemiddelde activatie in alle voxels in deze gyrus hetzelfde is voor eerlijke deelnemers en bedriegers, het over voxels gedistribueerde neurale patroon verschilt wanneer de morele standaard overschreden wordt. Dit suggereert dat de onderliggende morele norm toch universeel kan zijn voor iedereen, alhoewel mensen idiosyncratische standaardreacties hebben in moreel twijfelachtige situaties. Dat betekent dat mensen wel verschillen in hun neiging tot eerlijkheid of oneerlijkheid (billijkheid of egoïsme), maar dat de norm dat bedriegen slecht is, op dezelfde manier gecodeerd lijkt te zijn in de hersenen van ieder mens.

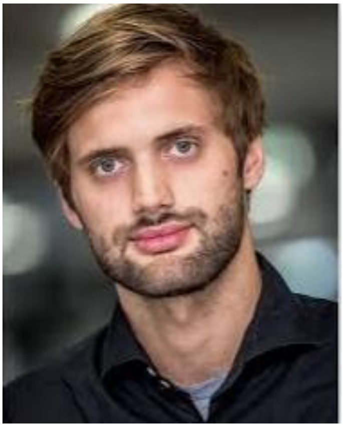
In het vijfde hoofdstuk zijn we dieper ingegaan op individuele verschillen in moreel gedrag en hebben we getest of er in de hersenen een stabiele aanleg voor eerlijkheid of oneerlijkheid gecodeerd is. Om een antwoord te vinden op deze vraag hebben we functionele connectiviteitspatronen in rust gemeten en deze gecombineerd met onze bedrogtaak. Op deze manier kunnen we onderzoeken of deze taak-onafhankelijke neurale patronen gebruikt kunnen worden om (on)eerlijkheid op een later tijdstip te voorspellen. We kwamen tot de belangrijke bevinding dat connectiviteit tussen beloning, introspectie en cognitieve controlenetwerken, op een betrouwbare en nauwkeurige manier kan voorspellen of een deelnemer over het algemeen eerlijk is dan wel een bedrieger.

Tenslotte hebben we in hoofdstuk 6 multivariate patroonanalyse (MVPA) gebruikt om prosociale versus zelfzuchtige beweegredenen te decoderen in het ultimatum- en dictatorspel. Onze bevindingen wijzen erop dat variaties in de mate waarin zelfbeheersings- en ‘theory-of-mind’-processen betrokken zijn bij de keuze ten grondslag liggen aan verschillen in billijkheid. Deze resultaten benadrukken het belang van ‘theory-of-mind’ en empathie bij het nemen van sociale beslissingen en suggereren dat een neiging tot empathie prosociaal gedrag kan bevorderen.

In het kort levert dit proefschrift drie belangrijke bijdragen. Op de eerste plaats geeft het uitsluitsel in een langdurig debat in de literatuur over de rol van cognitieve controle bij (on)eerlijkheid, want het toont aan dat cognitieve controle afhangt van iemands morele standaard. Ten tweede levert dit proefschrift een bijdrage omdat het stabiele neurale markers identificeert waarmee (on)eerlijkheid voorspeld kan worden in afwezigheid van de verleiding tot bedriegen. Ten slotte biedt het proefschrift een gedragsparadigma dat gebruikt kan worden om op een onopvallende manier vrijwillig, spontaan en herhaaldelijk bedrog te meten via trial-by-trial in de scanner of tijdens opname van een EEG.

Het onderzoek om inzicht te krijgen in de neurocognitieve processen van morele beslissingen en hoe die verschillen van persoon tot persoon, staat nog in de kinderschoenen. Ik ben ervan overtuigd dat morele psychologie en sociale neurowetenschap betekenisvolle inzichten zullen blijven leveren in hoe we beslissingen nemen over morele zaken, en dat deze inzichten nuttig zullen blijken bij het ontwerpen van interventies gericht op de onderliggende processen van (on)eerlijke en (on)billijke beslissingen, zoals interventies op basis van voorafgaand engagement, het versterken van empathie en het ontlokken en stimuleren van morele waarden.

About the Author



Sebastian Speer was born in Bad Friedrichshall, Germany. He received his Bachelor's degree in Psychology from the University of Groningen (cum laude). Subsequently, he obtained his Master's degree in Cognitive Neuroscience from the University of Amsterdam (cum laude). In 2016, he started his PhD research in Neuroeconomics at the Erasmus Research Institute of Management, specializing in social neuroscience with a focus on morality.

In his research, he investigates the psychological and neural underpinnings of decisions to either cheat and deceive, or to remain fair and honest. To do so he combines neuroimaging with behavioral experiments. His work has been published in top interdisciplinary journals such as the *Proceedings of the National Academy of Sciences of the United States of America* as well as in several neuroscience journal such as the *Journal of Neuroscience*, *Social Cognitive and affective Neuroscience* and *Frontiers of Neuroscience*. He also presented his work at major international conferences, including the Annual Conference of the Society for Neuroeconomics and the Interdisciplinary Symposium for Decision Neuroscience.

In September 2021, he started working as a postdoctoral researcher at the Netherlands Institute for Neuroscience.

Portfolio

Publications

Speer, S. P. H., Smidts., A, & Boksem, M. A. S. (2021). Cognitive control promotes either honesty or dishonesty, depending on one's moral default. *Journal of Neuroscience.* Advance online publication. Doi: https://doi.org/10.1523/JNEUROSCI.0666-21.2021

Speer, S., Smidts, A., & Boksem, M. A. S. (2021). Different neural mechanisms underlie nonhabitual honesty and nonhabitual cheating. *Frontiers in Neuroscience*, *15,* 83.

Speer, S. P. H., Smidts, A., & Boksem, M. A. S. (2020). Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proceedings of the National Academy of Sciences*, *117*(32), 19080-19091.

Voigt, K., Murawski, C., Speer, S. P. H., & Bode, S. (2020). Effective brain connectivity at rest is associated with choice-induced preference formation. *Human Brain Mapping*, 3077-3088.

Speer, S. P., & Boksem, M. A. (2019). Decoding fairness motivations from multivariate brain activity patterns. *Social Cognitive and Affective Neuroscience*, *14*(11), 1197-1207.

Voigt, K., Murawski, C., Speer, S., & Bode, S. (2019). Hard decisions shape the neural coding of preferences. *Journal of Neuroscience*, *39*(4), 718-726.

Working papers

Speer, S., Smidts, A., Boksem, M.A.S. Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest. *Preparing for resubmission to Neuroimage.*

Speer, S., Martinovici, A., Smidts, A., Boksem, M.A.S. The effect of stress on intuitive moral decisions. *Data analysis in progress.*

Speer, S., Prochazkova, L., van Elk, M. Investigating the neurocognitive effects of a psychedelic retreat on prosocial attitudes. *In preparation.*

Speer, S., van Dolder, D., van Assem, M., Turmunkh, U., Smidts, A., Boksem, M.A.S. Split or Steal: Predicting cooperation in prisoner dilemmas based on the neural response to naturalistic viewing of contestants' interactions. *In preparation.*

Conference Presentations

Decoding proposers' motivations in the ultimatum game

Society for Social and Affective Neuroscience, *Miami 2019,* Poster Presentation

Annual Conference of the Society for Neuroeconomics, *Philadelphia 2018,* Oral Presentation

European Society for Cogntive and Affective Neuroscience, *Leiden 2018,* Oral Presentation

Association for Neuropsychoeconomics • Zürich 2018 • Poster Presentation

Cognitive control processes override our moral default

Interdisciplinary Symposium on Decision Neuroscience, *Duke 2019,* Oral Presentation

Consumer Neuroscience Satellite Symposium, *Dublin 2019,* Oral Presentation

Society for Neuroeconomics, *Dublin 2019,* Poster Presentaiton

Society for Social and Affective Neuroscience , *Postponed to 2021* , Poster Presentation

Association for Neuropsychoeconomics, *Online 2020,* Oral Presentation

Predicting (dis)honesty from the brain at rest

Society for Neuroeconomics, *Online 2020*, Poster Presentation

Society for Social Neuroscience, *Online 2021*, Poster Presentation

European Society for Cogntive and Affective, *Online 2021*, Poster Presentation

The effect of stress on intuitive moral decisions

Interdisciplinary Symposium on Decision Neuroscience, *Online 2021*, Oral Presentation

Teaching

Neuroeconomics Minor (2017-2020)

Neuromarketing MSc Elective (2017-2018)

Bachelor Thesis Supervision (20-18-2020)

Workshop on Multivariate Pattern Analysis (2019)

Internship Coordinator Marketing Department (2020-2021)

Selected Coursework

Summer School for Advanced Modelling of Behavior, *Barcelona CRM*

Neuroeconomics Summer School 2019, *NYU Shanghai*

Summer School on the Neural Framework of Moral Cognition, *Siena, Italy*

Current Topics in Marketing Science, *Erasmus University Rotterdam*

Behavioral Decision Theory, *Erasmus University Rotterdam*

Machine Learning, *Erasmus University Rotterdam*

Programming in Matlab, *Erasmus University Rotterdam*

Statistical Methods, *Erasmus University Rotterdam*

Deep Learning, *Stanford University (Online)*

Machine Learning, *Stanford University (Online)*

Bayesian Statistics, *UC, Santa Cruz (Online)*

Reinforcement Learning, *University of Alberta (Online)*

The ERIM PhD Series

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam (EUR).

Dissertations in the last four years

Ahmadi, S., *A motivational perspective to decision-making and behavior in organizations*, Promotors: Prof. J.J.P. Jansen & Dr T.J.M. Mom, EPS-2019-477-S&E, https://repub.eur.nl/pub/116727

Akemu, O., *Corporate Responses to Social Issues: Essays in Social Entrepreneurship and Corporate Social Responsibility,*
Promotors: Prof. G.M. Whiteman & Dr S.P. Kennedy, EPS-2017-392-ORG, https://repub.eur.nl/pub/95768

Albuquerque de Sousa, J.A., *International stock markets: Essays on the determinants and consequences of financial market development,*
Promotors:
Prof. M.A. van Dijk & Prof. P.A.G. van Bergeijk, EPS-2019-465-F&A, https://repub.eur.nl/pub/115988

Alserda, G.A.G., *Choices in Pension Management,* Promotors:
Prof. S.G. van der Lecq & Dr O.W. Steenbeek, EPS-2017-432-F&A, https://repub.eur.nl/pub/103496

Anantavrasilp, S., *‘Essays on Ownership Structures, Corporate Finance Policies and Financial Reporting Decisions’*, Promotors: Prof. A. de Jong &
Prof. P.G.J. Roosenboom, EPS-2021-516-F&E, https://repub.eur.nl/pub/134947

Arampatzi, E., *Subjective Well-Being in Times of Crises: Evidence on the Wider Impact of Economic Crises and Turmoil on Subjective Well-Being,* Promotors: Prof. H.R. Commandeur, Prof. F. van Oort & Dr. M.J. Burger, EPS-2018-459-S&E, https://repub.eur.nl/pub/111830

Arslan, A.M., *Operational Strategies for On-demand Delivery Services,* Promotors: Prof. R.A. Zuidwijk & Dr N.A. H. Agatz, EPS-2019-481-LIS, https://repub.eur.nl/pub/126463

Aydin, Z. *‘Mobile Consumers and Applications: Essays on Mobile Marketing’*, Promotors: Prof. G.H. van Bruggen & Dr B. Ataman, EPS-2021-519-MKT, https://repub.eur.nl/pub/135352

Azadeh, K., *’Robotized Warehouses: Design and Performance Analysis’*, Promotors: Prof. dr. ir M.B.M. de Koster & Prof. D. Roy, EPS-2021-515-LIS, https://repub.eur.nl/pub/135208

Avci, E., *Surveillance of Complex Auction Markets: a Market Policy Analytics Approach*, Promotors: Prof. W. Ketter, Prof. H.W.G.M. van Heck &
Prof. D.W. Bunn, EPS-2018-426-LIS, https://repub.eur.nl/pub/106286

Balen, T.H. van, *Challenges of Early Stage Entrepreneurs: the Roles of Vision Communication and Team Membership Change,*
Promotors: Prof. J.C.M. van den Ende & Dr M. Tarakci, EPS-2019-468-LIS, https://repub.eur.nl/pub/115654

Bansraj, S.C., *The Principles of Private Equity: Ownership and Acquisitions*, Promotors: Prof. J.T.J Smit & Dr V. Volosovych, EPS-2020-507-F&A, https://repub.eur.nl/pub/132329

Bavato, D., *With New Eyes: The recognition of novelty and novel ideas*, Promotors: Prof. D.A. Stam & Dr. S. Tasselli, EPS-2020-500-LIS, https://repub.eur.nl/pub/134264

Bernoster, I., *Essays at the Intersection of Psychology, Biology, and Entrepreneurship*, Promotors: Prof. A.R. Thurik, Prof. I.H.A. Franken & Prof. P.J.F Groenen, EPS-2018-463-S&E, https://repub.eur.nl/pub/113907

Blagoeva, R.R., *The Hard Power Of Soft Power: A behavioral strategy perspective on how power, reputation, and status affect firms*, Promotors: Prof. J.J.P. Jansen & Prof. T.J.M. Mom, EPS-2020-495-S&E, https://repub.eur.nl/pub/127681

Bouman, P., *Passengers, Crowding and Complexity: Models for Passenger Oriented Public Transport*, Prof. L.G. Kroon, Prof. A. Schöbel & Prof. P.H.M. Vervest, EPS-2017-420-LIS, https://repub.eur.nl/pub/100767

Breugem, T., *'Crew Planning at Netherlands Railways: Improving Fairness, Attractiveness, and Efficiency'*, Promotors: Prof. D. Huisman & Dr T.A.B. Dollevoet, EPS-2020-494-LIS, https://repub.eur.nl/pub/124016

Bunderen, L. van, *Tug-of-War: Why and when teams get embroiled in power struggles*, Promotors: Prof. D.L. van Knippenberg & Dr. L. Greer, EPS-2018-446-ORG, https://repub.eur.nl/pub/105346

Burg, G.J.J. van den, *Algorithms for Multiclass Classification and Regularized Regression*, Promotors: Prof. P.J.F. Groenen & Dr. A. Alfons, EPS-2018-442-MKT, https://repub.eur.nl/pub/103929

Chammas, G., *Portfolio concentration*, Promotor: Prof. J. Spronk, EPS-2017-410-F&E, https://repub.eur.nl/pub/94975

Chan, H.Y., *'Decoding the consumer's brain: Neural representations of consumer experience'*, Promotors: Prof. A. Smidts & Dr M.A.S. Boksem, EPS-2019-493-MKT, https://repub.eur.nl/pub/124931

Couwenberg, L*., Context dependent valuation: A neuroscientific perspective on consumer decision-making,* Promotors: Prof. A. Smit, Prof. A.G. Sanfrey &
Dr M.A.S. Boksem, EPS-2020-505-MKT, https://repub.eur.nl/pub/129601

Dalmeijer, K., *Time Window Assignment in Distribution Networks,* Promotors: Prof A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-486-LIS, https://repub.eur.nl/pub/120773

Dennerlein, T. *Empowering Leadership and Employees' Achievement Motivations: the Role of Self-Efficacy and Goal Orientations in the Empowering Leadership Process,* Promotors: Prof. D.L. van Knippenberg & Dr J. Dietz,
EPS-2017-414-ORG, https://repub.eur.nl/pub/98438

Dolgova, E., *On Getting Along and Getting Ahead: How Personality Affects Social Network Dynamics,* Promotors: Prof. P.P.M.A.R Heugens & Prof. M.C. Schippers, EPS-2019-455-S&E, https://repub.eur.nl/pub/119150

Duijzer, L.E., *Mathematical Optimization in Vaccine Allocation,* Promotors: Prof. R. Dekker & Dr W.L. van Jaarsveld, EPS-2017-430-LIS, https://repub.eur.nl/pub/101487

Fasaei, H., *Changing the Narrative: The Behavioral Effects of Social Evaluations on the Decision Making of Organizations*, Promotors: Prof. J.J.P. Jansen,
Prof. T.J.M. Mom & Dr. M.P. Tempelaar, EPS-2020-492-S&E, https://repub.eur.nl/pub/129598

Eijlers, E., *Emotional Experience and Advertising Effectiveness: on the use of EEG in marketing*, Prof. A. Smidts & Prof. M.A.S. Boksem, EPS-2019-487-MKT, https://repub.eur.nl/pub/124053

El Nayal, O.S.A.N., *Firms and the State: An Examination of Corporate Political Activity and the Business-Government Interface*, Promotor: Prof. J. van Oosterhout & Dr. M. van Essen, EPS-2018-469-S&E, https://repub.eur.nl/pub/114683

Feng, Y., *The Effectiveness of Corporate Governance Mechanisms and Leadership Structure: Impacts on strategic change and firm performance,* Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda & Dr J.S. Sidhu,
EPS-2017-389-S&E, https://repub.eur.nl/pub/98470

Frick, T.W., *The Implications of Advertising Personalization for Firms, Consumer, and Ad Platfroms*, Promotors: Prof. T. Li & Prof. H.W.G.M. van Heck, EPS-2018-452-LIS, https://repub.eur.nl/pub/110314

Fytraki, A.T., *Behavioral Effects in Consumer Evaluations of Recommendation Systems*, Promotors: Prof. B.G.C. Dellaert & Prof. T. Li, EPS-2018-427-MKT, https://repub.eur.nl/pub/110457

Gai, J., *Contextualized Consumers: Theories and Evidence on Consumer Ethics, Product Recommendations, and Self-Control*, Promotors: Prof. S. Puntoni &
Prof. S.T.L. Sweldens, EPS-2020-498-MKT, https://repub.eur.nl/pub/127680

Ghazizadeh, P. *Empirical Studies on the Role of Financial Information in Asset and Capital Markets*, Promotors: Prof. A. de Jong & Prof. E. Peek, EPS-2019-470-F&A, https://repub.eur.nl/pub/114023

Giurge, L., *A Test of Time; A temporal and dynamic approach to power and ethics*, Promotors: Prof. M.H. van Dijke & Prof. D. De Cremer, EPS-2017-412-ORG, https://repub.eur.nl/pub/98451

Gobena, L., *Towards Integrating Antecedents of Voluntary Tax Compliance,*
Promotors: Prof. M.H. van Dijke & Dr P. Verboon, EPS-2017-436-ORG, https://repub.eur.nl/pub/103276

Groot, W.A., *Assessing Asset Pricing Anomalies*, Promotors: Prof. M.J.C.M. Verbeek & Prof. J.H. van Binsbergen, EPS-2017-437-F&A, https://repub.eur.nl/pub/103490

Hanselaar, R.M., *Raising Capital: On pricing, liquidity and incentives,* Promotors: Prof. M.A. van Dijk & Prof. P.G.J. Roosenboom, EPS-2018-429-F&A, https://repub.eur.nl/pub/113274

Harms, J. A., *Essays on the Behavioral Economics of Social Preferences and Bounded Rationality*, Prof. H.R. Commandeur & Dr K.E.H. Maas, EPS-2018-457-S&E, https://repub.eur.nl/pub/108831

Hendriks, G., *Multinational Enterprises and Limits to International Growth: Links between Domestic and Foreign Activities in a Firm's Portfolio,*
Promotors: Prof. P.P.M.A.R. Heugens & Dr. A.H.L Slangen, EPS-2019-464-S&E, https://repub.eur.nl/pub/114981

Hengelaar, G.A., *The Proactive Incumbent: Holy grail or hidden gem? Investigating whether the Dutch electricity sector can overcome the incumbent's curse and lead the sustainability transition*, Promotors: Prof. R.J. M. van Tulder & Dr K. Dittrich, EPS-2018-438-ORG, https://repub.eur.nl/pub/102953

Jacobs, B.J.D., *Marketing Analytics for High-Dimensional Assortments,* Promotors: Prof. A.C.D. Donkers & Prof. D. Fok, EPS-2017-445-MKT, https://repub.eur.nl/pub/103497

Jia, F., *The Value of Happiness in Entrepreneurship,* Promotors: Prof. D.L. van Knippenberg & Dr Y. Zhang, EPS-2019-479-ORG, https://repub.eur.nl/pub/115990

Kahlen, M. T., *Virtual Power Plants of Electric Vehicles in Sustainable Smart Electricity Markets,* Promotors: Prof. W. Ketter & Prof. A. Gupta, EPS-2017-431-LIS, https://repub.eur.nl/pub/100844

Kampen, S. van, *The Cross-sectional and Time-series Dynamics of Corporate Finance: Empirical evidence from financially constrained firms*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2018-440-F&A, https://repub.eur.nl/pub/105245

Karali, E., *Investigating Routines and Dynamic Capabilities for Change and Innovation,* Promotors: Prof. H.W. Volberda, Prof. H.R. Commandeur &
Dr J.S. Sidhu, EPS-2018-454-S&E, https://repub.eur.nl/pub/106274

Keko. E, *Essays on Innovation Generation in Incumbent Firms,* Promotors: Prof. S. Stremersch & Dr N.M.A. Camacho, EPS-2017-419-MKT, https://repub.eur.nl/pub/100841

Kerkkamp, R.B.O., *Optimisation Models for Supply Chain Coordination under Information Asymmetry,* Promotors: Prof. A.P.M. Wagelmans & Dr. W. van den Heuvel, EPS-2018-462-LIS, https://repub.eur.nl/pub/109770

Khattab, J., *Make Minorities Great Again: a contribution to workplace equity by identifying and addressing constraints and privileges,* Promotors: Prof. D.L. van Knippenberg & Dr A. Nederveen Pieterse, EPS-2017-421-ORG, https://repub.eur.nl/pub/99311

Kim, T. Y., *Data-driven Warehouse Management in Global Supply Chains*, Promotors: Prof. R. Dekker & Dr C. Heij, EPS-2018-449-LIS, https://repub.eur.nl/pub/109103

Klitsie, E.J., *Strategic Renewal in Institutional Contexts: The paradox of embedded agency*, Promotors: Prof. H.W. Volberda & Dr. S. Ansari, EPS-2018-444-S&E, https://repub.eur.nl/pub/106275

Koolen, D., *Market Risks and Strategies in Power Systems Integrating Renewable Energy*, Promotors: Prof. W. Ketter & Prof. R. Huisman, EPS-2019-467-LIS, https://repub.eur.nl/pub/115655

Kong, L. *Essays on Financial Coordination*, Promotors: Prof. M.J.C.M. Verbeek, Dr. D.G.J. Bongaerts & Dr. M.A. van Achter. EPS-2019-433-F&A, https://repub.eur.nl/pub/114516

Korman, B., *'Leader-Subordinate Relations: The Good, the Bad and the Paradoxical'*, Promotors: S.R. Giessner & Prof. C. Tröster, EPS-2021-511-ORG, https://repub.eur.nl/pub/135365

Kyosev, G.S., *Essays on Factor Investing*, Promotors: Prof. M.J.C.M. Verbeek &
Dr J.J. Huij, EPS-2019-474-F&A, https://repub.eur.nl/pub/116463

Lamballais Tessensohn, T., *Optimizing the Performance of Robotic Mobile Fulfillment Systems*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Dr D. Roy, EPS-2019-411-LIS, https://repub.eur.nl/pub/116477

Leung, W.L., *How Technology Shapes Consumption: Implications for Identity and Judgement*, Promotors: Prof. S. Puntoni & Dr G Paolacci, EPS-2019-485-MKT, https://repub.eur.nl/pub/117432

Li, Wei., *'Competition in the Retail Market of Consumer Packaged Goods'*, Promotors: Prof. D.Fok & Prof. Ph.H.B.F. Franses, EPS-2021-503-MKT, https://repub.eur.nl/pub/134873

Li, X. *Dynamic Decision Making under Supply Chain Competition*, Promotors: Prof. M.B.M de Koster, Prof. R. Dekker & Prof. R. Zuidwijk, EPS-2018-466-LIS, https://repub.eur.nl/pub/114028

Liu, N., *Behavioral Biases in Interpersonal Contexts*, Supervisors: Prof. A. Baillon & Prof. H. Bleichrodt, EPS-2017-408-MKT, https://repub.eur.nl/pub/95487

Maas, A.J.J., *Organizations and their external context: Impressions across time and space*, Promotors: Prof. P.P.M.A.R Heugens & Prof. T.H. Reus, EPS-2019-478-S&E, https://repub.eur.nl/pub/116480

Maira, E., *Consumers and Producers*, Promotors: Prof. S. Puntoni & Prof. C. Fuchs, EPS-2018-439-MKT, https://repub.eur.nl/pub/104387

Manouchehrabadi, B., *Information, Communication and Organizational Behavior*, Promotors: Prof. G.W.J. Hendrikse & Dr O.H. Swank, EPS-2020-502-ORG,
https://repub.eur.nl/pub/132185

Matawlie, N., *Through Mind and Behavior to Financial Decisions*, Promotors: Prof. J.T.J. Smit & Prof. P. Verwijmeren, EPS-2020-501-F&A, https://repub.eur.nl/pub/134265

Mirzaei, M., *'Advanced Storage and Retrieval Policies in Automated Warehouses'*, Promotors: Prof. M.B.M. de Koster & Dr N. Zaerpour, EPS-2020-490-LIS, https://repub.eur.nl/pub/125975

Nair, K.P., *Strengthening Corporate Leadership Research: The relevance of biological explanations*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R Heugens, EPS-2019-480-S&E, https://repub.eur.nl/pub/120023

Nullmeier, F.M.E., *Effective contracting of uncertain performance outcomes: Allocating responsibility for performance outcomes to align goals across supply chain actors*, Promotors: Prof. J.Y.F.Wynstra & Prof. E.M. van Raaij,
EPS-2019-484-LIS, https://repub.eur.nl/pub/118723

Okbay, A., *Essays on Genetics and the Social Sciences*,
Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger & Prof. P.J.F. Groenen, EPS-2017-413-S&E, https://repub.eur.nl/pub/95489

Peng, X., *Innovation, Member Sorting, and Evaluation of Agricultural Cooperatives*, Promotor: Prof. G.W.J. Hendriks, EPS-2017-409-ORG, https://repub.eur.nl/pub/94976

Petruchenya, A., *Essays on Cooperatives: Emergence, Retained Earnings, and Market Shares*, Promotors: Prof. G.W.J. Hendriks & Dr Y. Zhang, EPS-2018-447-ORG, https://repub.eur.nl/pub/105243

Plessis, C. du, *Influencers: The Role of Social Influence in Marketing*, Promotors: Prof. S. Puntoni & Prof. S.T.L.R. Sweldens, EPS-2017-425-MKT, https://repub.eur.nl/pub/103265

Pocock, M., *Status Inequalities in Business Exchange Relations in Luxury Markets*, Promotors: Prof. C.B.M. van Riel & Dr G.A.J.M. Berens, EPS-2017-346-ORG, https://repub.eur.nl/pub/98647

Polinder, G.J., *New Models and Applications for Railway Timetabling*, Prof. D. Huisman & Dr. M.E. Schmidt, EPS-2020-514-LIS, https://repub.eur.nl/pub/134600

Pozharliev, R., *Social Neuromarketing: The role of social context in measuring advertising effectiveness*, Promotors: Prof. W.J.M.I. Verbeke & Prof. J.W. van Strien, EPS-2017-402-MKT, https://repub.eur.nl/pub/95528

Qian, Z., *Time-Varying Integration and Portfolio Choices in the European Capital Markets,* Promotors: Prof. W.F.C. Verschoor, Prof. R.C.J. Zwinkels & Prof. M.A. Pieterse-Bloem, EPS-2020-488-F&A, https://repub.eur.nl/pub/124984

Reh, S.G., *A Temporal Perspective on Social Comparisons in Organizations,* Promotors: Prof. S.R. Giessner, Prof. N. van Quaquebeke & Dr. C. Troster, EPS-2018-471-ORG, https://repub.eur.nl/pub/114522

Riessen, B. van, *Optimal Transportation Plans and Portfolios for Synchromodal Container Networks*, Promotors: Prof. R. Dekker & Prof. R.R. Negenborn, EPS-2018-448-LIS, https://repub.eur.nl/pub/105248

Romochkina, I.V., *When Interests Collide: Understanding and modeling interests alignment using fair pricing in the context of interorganizational information systems*, Promotors: Prof. R.A. Zuidwijk & Prof. P.J. van Baalen, EPS-2020-451-LIS, https://repub.eur.nl/pub/127244

Schie, R. J. G. van, *Planning for Retirement: Save More or Retire Later?* Promotors: Prof. B. G. C. Dellaert & Prof. A.C.D. Donkers, EOS-2017-415-MKT, https://repub.eur.nl/pub/100846

Schneidmüller, T., *Engaging with Emerging Technologies: Socio-cognitive foundations of incumbent response*, Promotors: Prof. H. Volberda & Dr S.M. Ansari, EPS-2020-509-S&E, https://repub.eur.nl/pub/131124

Schouten, K.I.M. *Semantics-driven Aspect-based Sentiment Analysis,* Promotors: Prof. F.M.G. de Jong, Prof. R. Dekker & Dr. F. Frasincar, EPS-2018-453-LIS, https://repub.eur.nl/pub/112161

Sihag, V., *The Effectiveness of Organizational Controls: A meta-analytic review and an investigation in NPD outsourcing*, Promotors: Prof. J.C.M. van den Ende & Dr S.A. Rijsdijk, EPS-2019-476-LIS, https://repub.eur.nl/pub/115931

Slob, E., *'Integrating Genetics into Economics'*, Promotors: Prof. A.R. Thurik,
Prof. P.J.F. Groenen & Dr C.A. Rietveld, EPS-2021-517-S&E, https://repub.eur.nl/pub/135159

Smolka, K.M., *Essays on Entrepreneurial Cognition, Institution Building and Industry Emergence*, Promotors: P.P.M.A.R. Heugens, & Prof. J.P. Cornelissen,
Eps-2019-483-S&E, https://repub.eur.nl/pub/118760

Straeter, L.M., *Interpersonal Consumer Decision Making*,
Promotors: Prof. S.M.J. van Osselaer & Dr I.E. de Hooge, EPS-2017-423-MKT, https://repub.eur.nl/pub/100819

Stuppy, A., *Essays on Product Quality*, Promotors: Prof. S.M.J. van Osselaer &
Dr N.L. Mead. EPS-2018-461-MKT, https://repub.eur.nl/pub/111375

Subaşi, B., *Demographic Dissimilarity, Information Access and Individual Performance*, Promotors: Prof. D.L. van Knippenberg & Dr W.P. van Ginkel, EPS-2017-422-ORG, https://repub.eur.nl/pub/103495

Suurmond, R., *In Pursuit of Supplier Knowledge: Leveraging capabilities and dividing responsibilities in product and service contexts*, Promotors: Prof. J.Y.F Wynstra & Prof. J. Dul. EPS-2018-475-LIS, https://repub.eur.nl/pub/115138

Toxopeus, H.S. *Financing sustainable innovation: From a principal-agent to a collective action perspective*, Promotors: Prof. H.R. Commandeur & Dr. K.E.H. Maas. EPS-2019-458-S&E, https://repub.eur.nl/pub/114018

Turturea, R., *Overcoming Resource Constraints: The Role of Creative Resourcing and Equity Crowdfunding in Financing Entrepreneurial Ventures,*
Promotors: Prof. P.P.M.A.R Heugens, Prof. J.J.P. Jansen & Dr. I. Verheuil, EPS-2019-472-S&E, https://repub.eur.nl/pub/112859

Valboni, R.,*’Building Organizational (Dis-)Abilities: The impact of learning on the performance of mergers and acquisitions’*,Promotors: Prof. T.H. Reus &
Dr A.H.L. Slangen, EPS-2020-407-S&E, https://repub.eur.nl/pub/125226

Vandic, D., *Intelligent Information Systems for Web Product Search,* Promotors: Prof. U. Kaymak & Dr Frasincar, EPS-2017-405-LIS, https://repub.eur.nl/pub/95490

Verbeek, R.W.M., *Essays on Empirical Asset Pricing,* Promotors: Prof. M.A. van Dijk & Dr M. Szymanowska, EPS-2017-441-F&A, https://repub.eur.nl/pub/102977

Visser, T.R. Vehicle Routing and Time Slot Management in Online Retailing, Promotors: Prof. A.P.M. Wagelmans & Dr R. Spliet, EPS-2019-482-LIS, https://repub.eur.nl/pub/120772

Vlaming, R. de.,*Linear Mixed Models in Statistical Genetics,* Prof. A.R. Thurik,
Prof. P.J.F. Groenen & Prof. Ph.D. Koellinger, EPS-2017-416-S&E, https://repub.eur.nl/pub/100428

Vongswasdi, P., *Accelerating Leadership Development: An evidence-based perspective*, Promotors: Prof. D. van Dierendonck & Dr H.L. Leroy, EPS-2020-512-ORG, https://repub.eur.nl/pub/134079

Vries, H. de, *Evidence-Based Optimization in Humanitarian Logistics*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2017-435-LIS, https://repub.eur.nl/pub/102771

Wang, R., *Corporate Environmentalism in China*, Promotors: Prof. P.P.M.A.R Heugens & Dr F. Wijen, EPS-2017-417-S&E, https://repub.eur.nl/pub/99987

Wang, R., *Those Who Move Stock Prices*, Promotors: Prof. P. Verwijmeren & Prof. S. van Bekkum, EPS-2019-491-F&A, https://repub.eur.nl/pub/129057

Wasesa, M., *Agent-based inter-organizational systems in advanced logistics operations*, Promotors: Prof. H.W.G.M van Heck, Prof. R.A. Zuidwijk &
Dr A. W. Stam, EPS-2017-LIS-424, https://repub.eur.nl/pub/100527

Wessels, C., *Flexible Working Practices: How Employees Can Reap the Benefits for Engagement and Performance*, Promotors: Prof. H.W.G.M. van Heck,
Prof. P.J. van Baalen & Prof. M.C. Schippers, EPS-2017-418-LIS, https://repub.eur.nl/

Wiegmann, P.M., *Setting the Stage for Innovation: Balancing Diverse Interests through Standardisation*, Promotors: Prof. H.J. de Vries & Prof. K. Blind,
EPS-2019-473-LIS, https://repub.eur.nl/pub/114519

Wijaya, H.R., *Praise the Lord!: Infusing Values and Emotions into Neo-Institutional Theory*, Promotors: Prof. P.P.M.A.R. Heugens &
Prof. J.P. Cornelissen, EPS-2019-450-S&E, https://repub.eur.nl/pub/115973

Williams, A.N., *Make Our Planet Great Again: A Systems Perspective of Corporate Sustainability,* Promotors: Prof. G.M. Whiteman & Dr. S. Kennedy, EPS-2018-456-ORG, https://repub.eur.nl/pub/111032

Witte, C.T., *Bloody Business: Multinational investment in an increasingly conflict-afflicted world*, Promotors: Prof. H.P.G. Pennings, Prof. H.R. Commandeur &
Dr M.J. Burger, EPS-2018-443-S&E, https://repub.eur.nl/pub/104027

Wu, J.*, A Configural Approach to Understanding Voice Behavior in Teams,* Promotors: Prof. D.L. van Knippenberg & Prof. S.R. Giessner,
EPS-2020-510-ORG, https://repub.eur.nl/pub/132184

Ye, Q.C., *Multi-objective Optimization Methods for Allocation and Prediction*, Promotors: Prof. R. Dekker & Dr Y. Zhang, EPS-2019-460-LIS, https://repub.eur.nl/pub/116462

Yuan, Y., *The Emergence of Team Creativity: a social network perspective,* Promotors: Prof. D. L. van Knippenberg & Dr D. A. Stam, EPS-2017-434-ORG, https://repub.eur.nl/pub/100847

Zhang, Q., *Financing and Regulatory Frictions in Mergers and Acquisitions,*
Promotors: Prof. P.G.J. Roosenboom & Prof. A. de Jong, EPS-2018-428-F&A, https://repub.eur.nl/pub/103871