The dual-process approach to human sociality: A review

Valerio Capraro¹

Middlesex University London

Correspondence:

v.capraro@mdx.ac.uk

Abstract

Which social decisions are intuitive? Which are deliberative? The dual-process approach to human sociality has emerged in the last decades as a vibrant and exciting area of research. Here, I review the existing literature on the cognitive basis of cooperation, altruism, honesty, positive and negative reciprocity, and (act) utilitarianism. I conclude by introducing a game-theoretical framework that organizes the majority of the empirical regularities. This framework extends Rand and colleagues' Social Heuristics Hypothesis to any one-shot game G. The core assumption of this "General Social Heuristics Hypothesis" is that, under intuition, people play a Nash equilibrium of the "real analogue of G", Greal, while under deliberation people play a Nash equilibrium of G. Greal differs from G along two dimensions: G is one-shot, while Greal might be iterated; the payoffs of Greal might be different from the payoffs of G, although they are ordered in (almost) the same way.

Keywords: dual-process, pro-sociality, cooperation, altruism, honesty, reciprocity, moral judgments.

The dual-process approach to human sociality: A review

We, humans, are unique in the animal kingdom for our capacity to live in large societies made of thousands, if not millions, of unrelated individuals. Bees, ants, and the mole naked rat, for example, live in large societies, but individuals in the same society tend to share a substantial degree of biological relatedness. Other non-human primates live in groups of unrelated individuals, but these groups tend to be rather small. We are the only animals that are able to organize themselves in large societies, to the point that several scholars have argued that this uniquely-human capacity has been crucial for our evolutionary success as an animal species (Boyd & Richerson, 2005; Fehr & Fischbacher, 2003; Hill, 2002; Kaplan, Hill, Lancaster & Hurtado, 2000; Tomasello, Carpenter, Call, Behne & Moll, 2005). One obvious consequence of living in large societies is that most of our decisions have consequences beyond our own: they affect other people or the society as a whole. Since the well-being of our societies highly depend on these social decisions, understanding how people make this type of decisions is a topic of great interest, that has attracted the attention of social scientists for decades. In the last twenty years, in particular, there has been a proliferation of studies viewing at human sociality through a dual-process lens. Dual-process theories contend that people's choices result from the interplay between two cognitive systems, one that is fast and intuitive, and one that is slow and deliberative. By applying this dual process lens to human sociality, scholars have started investigating questions such as: Which social choices are intuitive? Which are deliberative? These questions turned out to be extremely fruitful and led to a series of exciting empirical results. These experimental findings have been paralleled by the development of theoretical frameworks. At the same time, however, a number of important problems remain unsolved, making this among the most thrilling and fascinating fields of research across behavioural

sciences. I thus believe that the time is mature to present a review of the state of the art. The goal of this survey is fourfold. First, I would like to introduce new researchers to a field that is growing exponentially and offers a number of open questions. Having this in mind, I will try to keep this review self-contained. Second, I would like to conduct an exhaustive literature review, which I believe can be helpful for scholars that are already in the field. Third, I would like to present a novel theoretical framework to organize the empirical evidence. This framework is inspired to the "Social Heuristics Hypothesis", that has been proposed by Rand et al. (2014) to explain their result that intuition promotes cooperation in one-shot social dilemmas. Here, I make their descriptive theory mathematically formal, which allows me to extend it to every one-shot game. I show that this "General Social Heuristics Hypothesis" organizes the majority of the empirical results, although it does have some limitations. Fourth, I would like to describe a number of open problems that I think are fundamental to further advance our understanding of the cognitive basis of human sociality.

Preliminaries

Social decisions

I start by defining the type of decisions that are the focus of this review: social decisions. Broadly speaking, one might define *social* whatever decision affects people other than the decision maker. Clearly, this is a very general class of decisions and it is unlikely that, without further specifications, one could draw general conclusions regarding the cognitive foundations of all social decisions. In fact, social decisions can differ from one another along several dimensions. One important dimension regards the consequences they bring about. What kind of consequences do these decisions have on other people? Good consequences or bad consequences? Right consequences or wrong consequences? Another important dimension along

which social decisions may differ is whether those people affected by the decisions have, in turn, the possibility to affect the decision maker. If so, does this happen after, simultaneously, or before knowing the decision maker's choice?

For these reasons, it is useful to restrict the attention to particular subclasses of social decisions. Previous research has mainly focused on cooperation, altruism, honesty, positive and negative reciprocity, and (act) utilitarianism. Following this work, I will also focus primarily on these types of social decisions. Exploring the cognitive basis of other types of decisions is certainly an important direction for future work. I will come back to this in the final section, where I will discuss the few studies that have investigated the cognitive basis of equity-efficiency, ingroup favouritism, and aggression. In what follows, I list only the types of decisions that are the primary focus of this review. (I selected these decisions because, for each of them, there have been at least *ten* experiments exploring their cognitive basis. I refer the reader to the final section for a short review of the other social decisions.) For each of these decision types, I give a short definition. More details about the way these decisions are measured are postponed to the corresponding sections.

Cooperation

Two or more decision-makers have to simultaneously decide whether to pay a cost to increase the payoff of their group.

Altruism

A decision-maker has to decide whether to pay a cost to increase the payoff of another person or group of people.

Honesty

A decision-maker has to decide whether to lie or to tell the truth. The decision-maker's and another person's payoffs depend on whether the decision-maker lies or tells the truth. The other person might be abstract (i.e., the experimenter) or concrete (i.e., another participant). Furthermore, the consequences of the lie could be positive or negative, for either or both the decision-maker and the other person.

Reciprocity

After observing (or being the recipient of) an action, a decision-maker has to decide whether to pay a cost to punish or to reward the actor. Several types of reciprocity exist, depending on the details of the context. "Direct reciprocity" usually refers to situations in which the decision-maker was the recipient of the action (Trivers, 1971); "indirect reciprocity" usually refers to situations in which the decision-maker only observed the action, without being affected by it (Nowak & Sigmund, 2005). Additionally, one can distinguish between "positive reciprocity" and "negative reciprocity", depending on whether the decision-maker has observed (or was the recipient of) a good or a bad action.

Deontology vs. act utilitarianism

A person has to decide whether to act so as to maximize the greater good, or to follow certain rules and norms, even if these lead to suboptimal outcomes.

The dual-process approach

Dual process theories refer to a set of frameworks sharing the core idea that people's choices result from the interplay between two cognitive systems, one that is fast and intuitive, and one that is slow and deliberative.

This idea has been formalized in several different ways (Fodor, 1983, 2001; Scheider & Schiffrin, 1977; Epstein, 1994; Epstein & Pacini, 1999; Chaiken, 1980; Chen & Chaiken, 1999; Reber, 1993; Evans & Over, 1996; Evans, 1989, 2006; Sloman, 1996; Smith & DeCoster, 2000; Hammond, 1996; Stanovich, 1999, 2004; Nisbett et al, 2001; Wilson, 2002; Lieberman, 2003; Toates, 2006; Strack & Deustch, 2004; Kahneman & Frederick, 2005; Evans, 2008; Glöckner & Betsch, 2008; Glöckner & Witteman, 2010; Kahneman, 2011; Stanovich, 2011; Evans & Stanovich, 2013; De Neys & Pennycook, in press), leading to a class of models that differ from each other in many details, not only in the names given to the two cognitive systems, but also in the specific functions that these systems are assumed to perform, to the point that these theories do not exactly map onto each other (Evans, 2008; Evans & Stanovich, 2013).

In this review, I adopt the point of view of Evans and Stanovich (2013), who provided convincing evidence of the existence of a dual-processing distinction which, theoretically, can be explained by assuming the existence of "rapid, autonomous processes, Type 1, that are assumed to yield default responses unless intervened on by distinctive higher order reasoning processes (Type 2)", which rely heavily on hypothetical thinking and working memory.

It is important to note that there is an ongoing discussion regarding the exact moment in which typically Type 1 and typically Type 2 responses are produced. Classic dual process theories tend to assume that Type 1 responses are produced immediately after the stimulus, whereas Type 2 responses comes at a later time. This assumption has been challenged by a recent line of research suggesting that people can process basic logical principles intuitively (Handley et al, 2011; De Neys, 2012; Banks & Hope, 2014; Pennycook, Fugelsang & Koehler, 2015; Ball, Thomson & Stupple, 2017). This observation is leading towards a revision of dual process models. In "dual process 2.0" models (see De Neys and Pennycook, in press, for a review)

"heuristic", intuitive, responses and "logical", intuitive responses are both generated automatically. Then, if they are in conflict (i.e., if they have similar strength), deliberation overrides the "heuristic" intuitive response in favour of the "logical" intuitive response; by contrast, if they are not in conflict, deliberation favours whichever intuitive response has a greater strength. It is important to emphasize, however, that according to this 2.0 model, it can never be the case that Type 2 overrides the "logical" intuitive response in favour of the "heuristic" intuitive response. Therefore, dual process 2.0 models differ from their 1.0 antecedents on the moment in which typically Type 2 responses are produced, but not on which choice is preferred by Type 2.

In this review, I will not take a personal stand on which specific dual-process theory is correct or on the specific functions of the two types of processes. I will rather follow a pragmatic approach by using the so-called "typical correlates". These are measures that can be manipulated in the lab which *typically* correlate with the two types of process. Importantly, these measures do *not* define the types. For example, to say that Type 1 processes tend to be fast, whereas Type 2 processes tend to be slow, does not mean that Type 1 processes will always be faster than Type 2 processes. It does not even mean that Type 1 responses will be generated earlier than Type 2 responses. It simply means that, by experimentally manipulating peoples' response time, for example by including a time pressure that forces people to provide their answer within a certain window of time, the experimenter would be more likely to observe an increase in the frequency of Type 1 driven responses, compared to the control condition, with no time pressure; symmetrically, by including a time delay that forces people to provide their answer after a certain amount of time, the experimenter would be more likely to observe an increase in the frequency of Type 2 driven responses, compared to the control condition, with no time delay. This

assumption does not require that Type 2 responses are generated later than Type 1 responses. In fact, this assumption is satisfied even in dual process 2.0 models, where "heuristic" and "logical" intuitive responses are assumed to be generated at the same time; but then, in a later time, Type 2 can override the heuristic intuition in favour of the logical intuition, but never the converse.

Coming to the "typical correlates", I will focus on those that have been used in most previous research applying the dual process approach to human sociality. In this line of research, Type 1 responses are assumed to be typically fast, autonomous, and effortless, while Type 2 responses are assumed to be typically slow, controlled, and effortful. All these correlates appear in the list of correlates reported by Evans and Stanovich (2013), Table 1. Besides these correlates, I will consider another correlate that has been used by many experimentalists because easy to manipulate in the laboratory: reliance on emotions. Note that also this correlate appear in Evans and Stanovich (2013), Table 1, but it appears in the broader category of the correlates that have been attributed to Type 1, although not commonly.

An important property of these correlates is that they can be easily manipulated in the laboratory. The most popular manipulation techniques are time constraints, cognitive load, conceptual primes, ego-depletion, neurostimulation, and the two-response paradigm.²

Time constraints

Since Type 1 is fast and Type 2 is slow (Evans & Stanovich, 2013), one obvious technique to promote the use of Type 1 over Type 2 is to ask participants to respond within a given window of time ("time pressure" condition). Conversely, to promote the use of Type 2 over Type 1, one can ask participants to stop and think for some time over the decision problem before making a decision ("time delay" condition).

Although useful, time constraints have been criticized along several dimensions. The first criticism originates from the observation that time constraints must be set in the decision screen and not in the instruction screen, because otherwise they would interfere with task comprehension. However, this methodological necessity implies that will-be-under-time-pressure participants can actually stay as long as they want in the instruction screen, where they are free to start using Type 2. Therefore, it is unclear whether the decision that then they make under time pressure can be considered their Type 1 decision, or even correlated to it. The second limitation of studies using time pressure is that the length of the constraint, which is typically relatively long (5-10 seconds). If, on the one hand, relatively long time windows are needed for practical reasons (i.e., it requires a few seconds to inform participants that they have to make a decision), they, on the other hand, imply that it remains unclear whether decisions observed after 10 seconds can be considered Type 1 decisions, or even correlated to them, especially in light of the evidence that first decisions are made almost instantaneously, and sometimes even before people become aware of them (Libet, 2009; Soon, Brass, Heinze & Haynes, 2008). The third limitation regards the compliance rate in the time pressure condition. While it is possible to force participants to respond after a certain amount of time (for example, by making the "submit" button in the decision screen visible only after 10 seconds), it is not possible to force participants to respond within a given amount of time. This generates a methodological issue. If the experimenter does not allow participants to submit their choice after a certain amount of time, there will be a selection bias; if the experimenter does allow participants to submit at any time, then there will be a proportion of participants who will not respect the time constraint. This is particularly problematic because participants are typically heterogenous in their response time, therefore participants who fail to respond within the time constraint may systematically differ

from the remaining ones in some (potentially unobservable) variable. This raises the question of how these participants should be treated in the analysis. In the seminal work by Rand et al. (2012), these participants were eliminated from the analysis. However, critics have observed that this would create potentially dangerous selection bias (Bouwmeester et al, 2017; Tinghög et al, 2013). More recently, experimenters have started including all participants in the analysis (Capraro, 2017; Rand, Newman & Wurzbacher, 2015; Tinghög et al, 2015). A third, more radical, solution is to have participants practice in a series of games, before the actual decision, so that they get used to respond quickly, which helps to minimize the proportion of participants who fail to comply with the time pressure (Everett et al., 2017). Another solution that has been recently proposed is to incentivize participants to reply quickly by making them lose money for each second they spend on the decision screen (Alós-Ferrer & Garagnani, 2018).

Cognitive load

Since Type 2 relies heavily on working memory (Evans & Stanovich, 2013), the cognitive system with a limited capacity that is responsible for storing short-term memory (Miyake & Shah, 1999), one can inhibit the use of Type 2 by letting participants use their working memory to solve a parallel, unrelated task, that uses short-term memory (Gilbert & Hixon, 1991; Gilbert, Tafarodi & Malone, 1993), as, for example, memorizing a sequence of digits (Swann, Hixon, Stein-Seroussi & Gilbert, 1990; Gilbert, Giesler & Morris, 1995; Trope & Alfieri, 1997; Shiv & Fedorikhin, 1999; Shiv & Nowlis, 2004), or hearing a series of letters while having to press a button whenever they hear a character that resounded two letters before (Gevins & Cutillo, 1993; Schulz, Fischbacher, Thöni & Utikal, 2014). Since these tasks need participants to store a temporary information in their working memory, they will be less likely to use their working memory when making the primary, parallel, decision. For example, Gilbert &

Hixon (1991) found that that cognitively loaded participants are more likely to rely on stereotypes; Shiv & Fedorikhin (1999) found that cognitively loaded participants are less likely to exert less self-control when choosing between cake and fruit salad; Hinson, Jameson and Whitney (2003) found that cognitively loaded participants have greater discounting of future rewards.

Studies based on cognitive load have limitations too. First, it is possible that the task being used to impair participants' access to working memory interacts with the primary decision problem. Second, analogously to time constraint studies, typically a number of participants fail to correctly perform the parallel task aimed at taxing their working memory; how should these participants be treated? Similar to time constraint studies, to avoid selection bias, experimenters typically include all participants in the analysis. Third, cognitive load manipulations can be used only to *undermine* the used Type 2, and never to *promote* it.

Conceptual priming

A technique that can be used to promote both Type 1 and Type 2 takes the name of conceptual priming. The idea is simple: before making a choice, participants are *primed* to rely either on Type 1 or on Type 2. The priming can be done in several different ways.

Some scholars explicitly asked participants to use their intuition or deliberation while making their choice (Liu & Hao, 2011), other scholars asked participant to write a piece of text in which they describe a time in their life in which following their gut reactions (or thinking carefully) led them to good decision making (Cappelen, Sørensen & Tungodden, 2013; Shenhav, Rand & Greene, 2012; Rand, Greene & Nowak, 2012), yet others asked participants to make a decision by "relying on emotion" vs. "relying on reason" (Capraro, Everett & Earp, 2019; Levine, Barasch, Rand, Berman & Small, 2018).³

The main limitation of these explicit conceptual primes is that they mention the words "intuition", "deliberation", and similar. This may create demand effects such that participants make a decision according to what they believe the intuitive or the deliberative choice should be (Rand, 2016; Kvarven et al., 2019). A potential solution to this limitation would be to use subtler primes, as, for example, those used by Small, Loewenstein and Sloviv (2007). However, to the best of my knowledge, this kind of primes have never been implemented on social decisions of the type considered in the current review.

Another priming technique consists in having participants read the instructions and make their decision in a foreign language. Indeed, the native language is intuitive and automatic, while foreign languages tend to be processed more slowly. This suggests that foreign language might affect the cognitive process being used to make the decision. According to this idea, foreign language has been shown to reduce heuristics and biases (Costa et al., 2014a; Keysar et al., 2012; Costa, Vives & Corey, 2017; Hakayawa, Costa, Foucart & Keysar, 2016) and to reduce emotionally-driven responses in moral dilemmas (Cipolletti, McFarlane & Weissglass, 2016; Corey et al., 2017; Costa et al., 2014b, Geipel, Hadjichristidis & Surian, 2015; Hayakawa, Tannenbaum, Costa, Corey & Keysar, 2017). This line of research thus suggests that information processed in a second language reduces intuitive thinking.

Priming through a foreign language has two limitations. One is that several works have shown that information processed in a second language does not increase analytic thinking. For example, foreign language does not improve performance on CRT (Costa et al., 2014a), conjunction fallacy or base rate neglect problems (Vives et al., 2018), and detection of semantic illusions (Geipel et al., 2015; but see Hadjichristidis, Geipel, & Surian, 2017). Therefore, if anything, foreign language can be a useful technique to impair Type 1 thinking, but not to

promote Type 2 responses. The second limitation is potentially more critical. Reasoning in a second language is typically cognitively tiring and this might contribute to depletion of cognitive resources that work in the opposite direction of the prime (Volk, Köhler & Pudelko, 2014).

Ego depletion

Ego depletion is an experimental manipulation based on the theoretical assumption that Type 2 requires the exert of self-control to override immediate, autonomous, Type 1-driven reactions. Although there is an ongoing debate about whether self-control is a limited resource (Baumeister & Heatherton, 1996; Baumeister, Heatherton & Tice, 1994; Baumeister & Tierney, 2011; Muraven & Baumeister, 2000) or not (Inzlicht & Schmeichel, 2012; Inzlicht, Schmeichel & Macrae, 2014), several studies have provided evidence that exerting self-control at Time 1 reduces the ability to exert self-control at Time 2. Following this line of work, social scientists have investigated how participants make social decisions at Time 2 after having completed, at Time 1, a task meant to undermine their self-control.

Self-control can be impaired in several ways, for example using the Stroop task (Stroop, 1935), or the e-hunting task (Moller, Deci & Ryan, 2006). Other empirical techniques that are sometimes used to decrease self-control are sleep deprivation,⁴ hunger,⁵ alcohol intoxication,⁶ and mortality salience.⁷

Studies using ego depletion have their shortcomings too. First, while ego-depletion tasks can be used to impair Type 2 processing, there is no symmetric equivalent to impair Type 1 processing. A potentially more critical limitation is related to the ongoing debate about whether the ego-depletion effect actually exists. On the one hand, recent meta-analyses have shown that, if one corrects for publication bias, the ego-depletion effect may not be different from zero (Carter, Kofler, Forster & McCullough, 2015; Carter & McCullough, 2014; Hagger et al, 2016).

On the other hand, also these meta-analyses have been criticized along several dimensions, one of which is that they include also studies without a manipulation check. Therefore, it is possible that the overall lack of an effect is ultimately driven by the inclusion of studies that were unable to manipulate self-control. Consequently, whether ego-depletion is a real or an illusory effect is still under debate (Friese, Loschelder, Gieseler, Frankenbach & Inzlicht, 2019).

Neurostimulation

Neurostimulation is based on neuroscience research suggesting that a particular area of the brain, the right lateral prefrontal cortex (rLPFC), and, more specifically, its subarea, the right dorsolateral prefrontal cortex (rDLPFC), are implicated in working memory, planning, abstract reasoning, and self-restraint (Hare et al, 2009; Hutcherson et al, 2012; Barbey et al, 2013; Zhu et al, 2014). Crucially, this area, being relatively superficial, can be activated or disactivated using non-invasive electric or magnetic stimulation tools. These tools thus provide a useful way to explore the causal link between cognitive process and social decisions. There are two such tools. The tDCS (Transcranial Direct Current Stimulation) stimulates the target region of the brain by applying a small electrical direct current directly to the scalp; the TMS (Transcranial Magnetic Stimulation) uses magnetic fields to stimulate nerve cells in the target region of the brain.

The main limitation of these techniques relies in the difficulty of targeting the exact region of interest. For example, de Berker, Bikson and Bestmann (2013) criticize tDCS because the anode is typically placed over the region of interest, and this implicitly assumes that the current spreads uniformly from the anode to the target area. However, this is generally not true, but dramatically depends on the topography of the cortical surface, which, in some cases, can even reverse the polarity of the current (Rahman et al., 2013). Similarly, TMS has been criticized because of the unclearness and fundamental unpredictability of the propagation of its effects to

other areas of the brain. This is particularly problematic because "if a group of neurons are involved in a given task, introducing a TMS pulse is highly unlikely to selectively stimulate the same coordinated pattern of neural activity as performance of that task", suggesting that TMS may interfere with task performance (Walsh & Cowey, 2000).

The two-response paradigm

The two-response paradigm, introduced by Thompson, Turner and Pennycook (2011), is based on the postulate that the intuitive response comes to people's mind earlier than the deliberative responses. Then, if there is a conflict between the intuitive response and the deliberative one, the role of Type 2 is to override the intuitive response in favour of the deliberative one. Therefore, one can separate the two types of response using a two-response method whereby participants are asked to make two consecutive decisions, one under time pressure or cognitive load, and one under no constraint (or under time delay). By analysing if and how participants switch between choices, one can determine which choice is deliberate and which is intuitive.

The main conceptual limitation of this method is that it introduces an experimental demand for consistency, such that participants may be reluctant to change their initial, intuitive choice. Classic work indeed shows that people have strong preferences for maintaining their previous decision (Samuelson & Zeckhauser, 1988). One potential solution for this concern is to introduce a control treatment in which subjects make only the deliberative choice (Bago, Bonnefon & De Neys, 2019).

Other preliminary notions

The cognitive reflection test

The cognitive reflection test (CRT), originally developed by Frederick (2005), is a set of questions, each of which is characterized by the property that an immediate but wrong answer typically pops up to the one's mind; to find the correct answer, one needs to override this intuitive but wrong response. The classic CRT is made of three questions. A typical item reads as follows: "A bat and a ball cost \$1.10 in total. If the bat costs \$1 more than the ball, how much does the ball cost?" The typical reader would be tempted to respond that the ball costs \$0.10. However, a moment of reflection shows that, since the bat costs \$1 more than the ball, then, if the ball costed \$0.10, then the bat would cost \$1.10, which would imply that the cost of the bat and the ball together would be \$1.20, and not \$1.10 as assumed. For this reason, the score in the CRT is usually taken as a measure of deliberation. In support of this interpretation, Pennycook, Cheyne, Kohler and Fugelsang (2016) have shown that CRT scores correlate with Need for Cognition. However, interestingly, it is unclear whether the number of intuitive answers can be taken as a measure of intuition, because this number does not seem to correlate with Faith on Intuition⁹ (Pennycook et al. 2016). Another criticism to the CRT is that it requires numerical abilities, and this ultimately generates gender differences in CRT scores. In order to avoid this

issue, Thomson and Oppenheimer (2016) have proposed a new version of the CRT in which the use of numerical abilities is reduced. A typical question in this variant of the CRT is the following: "Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?" Here the intuitive answer is June; but the correct answer is, obviously, Emily.

The social value orientation

The Social Value Orientation (SVO) of an individual corresponds to the weight a person poses on the payoff of other individuals in relation to her own (Messick and McClintock, 1968). The basic idea is to represent the utility function of one individual as a combination between the material payoff of that individual and the material payoff of another individual:

$$U(p_s,p_o)=w_sp_s+w_op_o,$$

where p_s is the material payoff of the self, p_o is the payoff of the other, w_s is the weight that one individual places on her own payoff, and w_o is the weight that the same individual places on the other's payoff. An individual with $w_s = 1$ and $w_o = 0$ is perfectly individualist (or self-regarding), because she cares only about her own payoff; an individual with $w_s = 0$ and $w_o = 1$ is perfectly altruistic because she cares only about the other's payoff; an individual with $w_s = -1$ and $w_o = 0$ is perfectly masochist, because she only cares about damaging herself; an individual with $w_s = 0$ and $w_o = -1$ is perfectly sadist, because she only cares about damaging the other person; an individual with $w_s = w_o$ is perfectly cooperative, because she only cares about the total welfare (i.e., the sum of her own payoff and the payoff of the other person).

Clearly one can normalize w_s and w_o in such a way that $w_s^2 + w_o^2 = I$. This leads to a simple and useful geometric representation of the SVO, as the angle θ such that $w_s = cos(\theta)$ and $w_o = sin(\theta)$ (Griesinger and Livingston, 1973). In this way, one can represent all possible SVOs

as angles of a ring in the own-payoff/other-payoff plane, which in turn leads to what is known as "the ring measure" of the SVO (Liebrand, 1984), which allows to compute the SVO of a participant through a series of 24 mini-dictator games. The payoffs corresponding to the dictator and the recipient correspond to equally spaced points on the ring. By adding up all the resulting vectors, one can compute a single vector, whose angle represents the SVO of the individual. Individuals with a vector falling between 22.5° and 112.5° are classified as pro-socials; those with a vector falling between 292.5° (or -67.5°) and 22.5° are classified as pro-selves. The SVO has been shown to predict cooperation in social dilemmas (Balliet, Parks & Joireman, 2009) and charitable giving (Van Lange, Bekkers, Schuyt & Van Vugt, 2007).

Another measure of the SVO, that is receiving considerable attention, has been introduced more recently by Murphy, Ackermann and Handgraaf (2011). The starting point of this measure is the circle in the self-payoff/other-payoff plane with centre in (50,50) and radius 50. Then one selects, on this circle, the four points corresponding to the four basic, idealized, social orientations: altruistic orientations correspond to the point (50,100), cooperative orientations correspond to the point (50+50 $\sqrt{2}$,50+50 $\sqrt{2}$), individualist orientations correspond to the point (100,50), and, finally, competitive orientations correspond to the point (50+50 $\sqrt{2}$, 50-50 $\sqrt{2}$). By connecting these four points, one finds six lines that can be used to determine the allocations for self and other of six mini-dictator games. Then one computes the mean allocation for self, A_s , and the mean allocation for other, A_o . Next one subtracts 50 from each of these values. Finally, one defines SVO° to be the arctan of $(A_o - 50) / (A_s - 50)$. This SVO angle is different from the one introduced by Liebrand (1984). Indeed, possible values for this angle spans from -16.26° (for a perfect competitor) to 61.39° (for a perfect altruist). Negative values correspond to competitors, an angle of 0° correspond to a perfect individualist, and an angle of

45° corresponds to giving equal weights to own and others' outcomes. The angle of 22.45° is usually taken as the threshold to separate individual orientations from prosocial orientations.

Basics of game theory

After reviewing the empirical evidence, I will introduce a game-theoretical framework to organize the experimental regularities. This framework requires some basic notions of Game Theory. Since one of my aims is to keep this article self-contained, in this section I will review the notions that are needed to develop the theoretical framework. These concepts will be used only for introducing the model. Therefore, a reader who, at this moment, is not yet interested in the theoretical framework, can skip this section without affecting the reading of the following sections.

The first notion needed for the theoretical model is the very definition of a game. Intuitively speaking, a game is any situation where a number of people interact among each other, by making choices, and where the payoff of each player depends on the configuration of the choices made by each player. More formally, a *finite game in normal form G* is given by:

- A set of players $P = \{1, ..., n\},\$
- For each player i, a finite non-empty strategy set S_i ,
- For each player i, a payoff function u_i from $S = S_1 x ... x S_n$ to the real numbers.

Note that S_i can contain only one element, in which case player i is not actually choosing among multiple options. This is useful to represent situations in which player i is affected by the decisions of other players, but i actually makes no active decisions (e.g., the recipient of the dictator game). Elements of S are called "strategy profiles". Given a strategy profile $s = (s_1, ..., s_n) \in S$, as usual, I will use the notation $s = (s_i, s_{-i})$ when I will need to highlight the strategy played by player i versus the strategy profile played by the other players, denoted -i, for brevity.

A Nash equilibrium of G is any strategy profile $(s_1, ..., s_n) \in S$ such that, for all $i \in P$, one has $u_i(s_i, s_{-i}) \ge u_i(t_i, s_{-i})$, for all $t_i \in S_i$.

Intuitively speaking, a Nash equilibrium is a strategy choice for each player such that no player has an incentive to change strategy.

By the classical Nash theorem, every finite game has an equilibrium (Nash, 1950).

Another basic notion is that of discounted payoffs, which will be useful to define the payoff of an infinitely repeated game. Indeed, finitely repeated games create no definitional problems, as one can define the payoff at round T simply as the sum of the payoffs until round T. This definition, however, becomes problematic in infinitely repeated games, as it may not converge to a finite number. There are several ways to avoid this technical difficulty. A popular one, especially used in situations in which there is reason to assume that future payoffs count less than present payoffs, is by discounting future payoffs by a factor θ . More formally, the payoff of player i at round T is defined as:

$$U_i = \sum_{t=1}^T e^t u_i$$

The cognitive basis of cooperation

Probably, the most important form of social behaviour is cooperation. Many scholars have come to argue that our uniquely-human capacity to cooperate with unrelated others has been the secret of our enormous evolutionary success as an animal species (Ostrom, 2000; Fehr & Gächter, 2002; Milinski, Semman & Krambeck, 2002; Gintis, Bowles, Boyd & Fehr, 2003; Fehr & Fischbacher, 2004; Nowak, 2006; Bowles & Gintis, 2011; Rand & Nowak, 2013; Perc et al, 2017). Along these lines, research in psychology has suggested that the psychological basis of

cooperation, *shared intentionality*, is what makes humans uniquely humans, as it is possessed by children, but not by great apes (Tomasello et al., 2005).

Measures of cooperation

Cooperation can be operationally defined in several different ways. Here, I follow the literature on social dilemmas. A social dilemma is any situation in which there is a conflict between individual interest and group interest (Olson, 1965; Hardin, 1968; Dawes, 1980; Kollock, 1988; Fehr & Fischbacher, 2004; Nowak, 2006; Rand & Nowak, 2013; Capraro, 2013; Perc et al, 2017). The main social dilemmas considered in the literature are the prisoner's dilemma and the public goods game.

The prisoner's dilemma

The prisoner's dilemma is a simultaneous-move, symmetric, game, with two players, each of whom can either cooperate or defect. If both players cooperate, they receive, as payoff, the *reward* for cooperation, R; if both player defect, they receive, as payoff, the *punishment*, P. If one player cooperates and the other player defects, the cooperator receives the *sucker*'s payoff, S, and the defector receives the *temptation* payoff, T. Payoffs are assumed to satisfy the inequalities T > R > P > S. These inequalities guarantee that both players would be better off if they both cooperate than if they both defect. However, unilateral defection brings the highest individual payoff. This generates the conflict between individual payoff and group payoff that is at the core of the social dilemma.

A particular version of the prisoner's dilemma, often used by biologists, is formulated in terms of cost and benefit of cooperation. Specifically, cooperation means paying a cost c to give a benefit b > c to the other person; defection means doing nothing. This corresponds to a prisoner's dilemma with T = b, R = b - c, P = 0, S = -c.

The classical prisoner's dilemma game is symmetric (i.e., the values T, R, P, S do not depend on the players). However, some scholars have also studied the cognitive basis of cooperation in asymmetric games (e.g., Lotz, 2015). Asymmetric social dilemmas are important in light of practical applications, because in reality there is often power imbalance among people; this unequal distribution of power can be formalized in terms of asymmetries among players between costs and benefits of cooperation.

The public goods game

The public goods game is a simultaneous-move, symmetric, game with N players, each of whom starts with an initial endowment e > 0. Each player has to choose how much of it, if any, to contribute to a public pool. Let c_i be player i's contribution. The payoff of player i is defined as $p_i(c_1,...c_N) = e - c_i + m(c_1 + ... + c_N)$, where m is a constant, named marginal return for cooperation, which is assumed to be strictly greater than 1/n and strictly smaller than 1. The fact that $m \in (1/n, 1)$ implies that, as in the prisoner's dilemma, players would be better off if they all cooperate, compared to when they all defect; however, each individual is better off when s/he defects, independently of the other players' strategies. Therefore, also in this case there is a conflict between individual payoff and group payoff. In this case too, one can easily define an asymmetric variant of the game, in order to take into account potential power imbalances among players.

Other social dilemmas

The prisoner's dilemma and the public goods game are not the only games characterized by a conflict between individual payoff and group payoff. Other examples are the traveller's dilemma (Basu, 1994) and the Bertrand competition. However, to the best of my knowledge, no one has explored the cognitive underpinnings of cooperation in these social dilemmas.¹¹

Therefore, in this review, I will not consider these games. Somewhat related to social dilemmas, one paper has looked at the effect of intuition and deliberation on the Stag-Hunt game (Belloc et al. 2018). In this game there is conflict between a risk-dominant, "safe", strategy and one that is efficient; however, both these strategies are Nash equilibria, and therefore, the Stag-Hunt game is not a social dilemma, because there is no conflict between individual and group interest. Another reason for excluding the Stag-Hunt game from this review is that, in fact, there is only one study looking at the effect of intuition and deliberation on this game (Belloc et al. 2018); more research is needed before one could write a review on the cognitive basis of the trade-off between safety and efficiency.

Other definitions of cooperation

As I mentioned above, there are other ways to define cooperative behaviour. Of particular interest, for the current review, is the distinction between *pure cooperation* and *strategic cooperation* proposed by Rand (2016).

Pure cooperation refers to any game in which: (i) a player pays a cost to give a benefit to one or more other players; (ii) the payoff maximizing strategy in the iterated variant of the game is different from the payoff maximizing strategy in its one-shot, anonymous, variant. One can easily see that the prisoner's dilemma and the public goods game (as well as the traveller's dilemma and the Bertrand competition) verify these two properties. But there are also other games satisfying this property, such as the trust game. Here, Player 1 is given a certain amount of money and has to decide how much of it, if any, to invest. The amount invested is multiplied by a factor greater than 1 and given to Player 2. Then Player 2 decides how much of the amount s/he receives to return to Player 1. Clearly, in one-shot anonymous interactions, Player 2 has no incentive to return any money, and therefore Player 2's payoff-maximizing strategy in one-shot

anonymous interactions is to keep the whole amount s/he received. However, in repeated interactions it might be optimal for Player 2 to return part of the money in order to avoid that Player 1 in the next round does not invest any money, leading to a long-term loss for Player 2. Therefore, the trust game (played in the role of Player 2) is an example of pure cooperation in the sense of Rand (2016).

Strategic cooperation refers to those games in which one player pays a cost to give a benefit to another player; however, whether this choice maximizes the player's payoff depends on the decision of the other player. The two prototypical cases considered by Rand (2016) are the one-shot, anonymous, trust game (player 1) and the one-shot, anonymous, ultimatum game (player 1). Indeed, in the one-shot, anonymous, trust game, whether investing money maximizes Player 1's payoff depends on the behaviour of Player 2: if Player 1 believes that Player 2 will return a sufficient amount of money, then it is optimal for Player 1 to invest their money; otherwise, it is not. In the ultimatum game, Player 1 has to decide how to split a certain sum of money and Player 2 has to decide whether to accept or reject the proposed split. In case Player 2 accepts the offer, then the two players are paid according to the agreed split; in case Player 2 rejects the offer, both players end up with nothing. Therefore, Player 1 can make a larger offer (i.e., paying a cost to give a benefit to Player 2), but this choice will be optimal only if Player 1 believes that Player 2 will accept the corresponding offer and will reject any smaller offer. Other cases of strategic cooperation are the repeated public Goods game and the repeated prisoner's dilemma.

Review of the empirical evidence

A number of empirical studies have explored the causal link between cognitive process and cooperation, especially using time constraints. The seminal work by Rand, Greene & Nowak

(2012) showed that time pressure increases cooperation, relative to time delay, in the public goods game. This work sparked enormous interest in the field and was immediately followed by a number of works. Tinghög et al. (2013) criticized the methodology used by Rand et al. (2012), because it excluded from the analysis participants who failed to comply to the time constraints, leading to potential selection bias. Specifically, Tinghög et al. (2013) conducted five studies and, without excluding participants who fail to respond within the given time constraints, they found no overall effect of time pressure on cooperation. As a consequence, they advanced the hypothesis that the earlier results of Rand et al. (2012) were driven by selection bias. In their response paper, Rand et al. (2013) showed that the results of Rand et al. (2012) hold also when including in the analysis all subjects, and not only those who complied with the time pressure, suggesting that their original findings had not been driven by selection bias. Furthermore, Rand et al. (2014) replicated this result with a greater sample. Additionally, they found that the effect was driven by participants with no previous experience in the public goods game. This observation led to the so-called Social Heuristics Hypothesis (see below for further details). The positive effect of time pressure on cooperation was replicated also in the context of competitively framed public goods games (Cone & Rand, 2014) and in the context of public goods games played with out-group members (Rand, Newman & Wurzbacher, 2015). In parallel, however, a number of works failed to replicate the basic positive effect of time pressure on cooperation. Verkoeijen & Bouwmeester (2014), Strømland, Tjotta and Torsvik (2016) and Lohse (2016) all found no overall effect of time pressure on cooperation; if anything, Lohse (2016) found that time pressure decreases cooperation, but only among subjects who score high in the CRT. Bouwmeester et al. (2017) conducted a large (21 studies) preregistered replication of Study 7 in Rand et al. (2012). They replicated the effect that time pressure increases cooperation, but only

when participants who fail to comply with the time constraint are excluded from the analysis; they concluded that the positive effect of time pressure on cooperation is best explained in terms of selection bias. Two works even found the opposite effect: Capraro & Cococcioni (2016) found that extreme time pressure leads to less cooperation in the one-shot prisoner's dilemma, Goeschl and Lohse (2018) found that time pressure causes more free-riding in the one-shot public goods game. However, the former work had the limitation that time constraints were added in the instruction screen, and this might have interacted with participants' level of comprehension, whereas the latter work had very high non-compliance rate (43%). More recently, Everett, Ingbretsen, Cushman and Cikara (2017) made an important step in the direction of eliminating the methodological limitations of time pressure studies. They used a prisoner's dilemma preceded by an extensive warm-up. Moreover, the time constraint in the time pressure condition was set at 15 seconds, instead of 10 seconds, as in the majority of previous studies. In this way, the authors were able to reduce non-compliance to time pressure from about 50% in Rand et al. (2012) to just 2%. In doing so, they confirmed the original result by Rand and colleagues that time pressure has a positive effect on cooperation. Moreover, they replicated the result that time pressure increases cooperation even when playing versus an out-group member. Additionally, they also found that the same effect is present in the prisoner's dilemma played in a loss frame. A similar warm-up procedure was followed by Isler, Maule and Starmer (2018), who also found that time pressure increases cooperation.

Some works have explored the role of potential moderators. Rand and Kraft-Todd (2014) found that time pressure increases cooperation in a public goods game, but only among participants who are both inexperienced and have high trust on people with whom they are interact in their everyday life; symmetrically, Capraro and Cococcioni (2015) found that time

pressure does not affect cooperation in a prisoner's dilemma played in India, where the rate of cooperation is much lower than in a similar experiment conducted in the US. These papers were interpreted as supporting the evidence that time pressure makes people more likely to use cooperative heuristics learned in their everyday interaction. Another work investigated the cognitive foundations of cooperation as a function of participants' SVO. Specifically, Alós-Ferrer and Garagnani (2018) implemented incentive compatible time constraints as follows: participants under time delay were paid €1 if they responded after 10 seconds; participants under time pressure lost 30 cents for every second spent on the decision screen. In doing so, they found no overall effect of time pressure on cooperation in a one-shot public goods game; however, they found that time constraints interacted with SVO, in such a way that time pressure made prosocials become more prosocials, left individualists unaffected, and made competitors become more competitive. They replicated this result also in a repeated public goods game with random re-matching and no feedback. They concluded that intuition enhances intrinsic predispositions. Bird et al. (2019) explored the moderating effect of risk aversion in a sample of men. In doing so, they found that time pressure increases cooperation, but only among low risk men, while the effect even reverses among high risk men, who become less cooperative under time pressure. Another potential important moderator is the benefit to cost ratio of cooperation. Kessler, Kivimaki and Niederle (2017) have reported that time pressure increases cooperation in prisoner's dilemmas with a relatively small benefit to cost ratio (smaller than 2, in their experiment), whereas time pressure decreases cooperation when the benefit to cost ratio is relatively high (greater than 4, in their experiment, with the exception of b/c=6, which gave a non-significant effect).

A handful of papers explored the effect of conceptual primes on cooperation. Rand et al. (2012) found that priming intuition favours cooperation, compared to priming deliberation. Lotz (2015) conceptually replicated this result using asymmetric prisoner's dilemmas, whereas Levine et al. (2018) did so using a conceptual priming of emotion versus reason. Consistent with these works, Urbig, Terjesen, Procher, Muehlfeld and van Witteloostuijn (2016) found that Dutch participants cooperate less in a public goods game when the instructions are presented in English than when they are presented in their mother tongue. In contrast, Camerer et al. (2018) failed to replicate the original study in Rand et al. (2012). Gargalianou, Urbig and Witteloostuijn (2017) found that Dutch women become *more* cooperative in a prisoner's dilemma when the instructions are presented in English or French, compared to when they are presented in their native language; for men, instead, the effect was trending in the opposite direction, although not significantly so.

As for ego-depletion, Capraro and Cococcioni (2016) found that depleting participants' self-control with three tasks (Stroop task, e-hunting task, and give-the-wrong-answer task) impairs cooperative behaviour in a one-shot prisoner's dilemma, but only among participants with no previous experience about the game, while it has no effect on experienced participants. Rantapuska et al. (2017) found that hunger does not affect cooperation in a public goods game; along the same lines Clark and Dickinson (2017) found that sleep restriction does not affect contributions to the public goods game.

Only one study used neurostimulation techniques to explore the cognitive basis of cooperation. Li et al. (2018) used an iterated asymmetric public goods game with no feedback after every iteration. They used transcranial direct current stimulation to stimulate or reduce the rDLPFC. In doing so, the authors found that all participants, regardless of their (asymmetric)

starting endowment, cooperate more when the rDLPFC is activated than when it is inhibited. The baseline, "sham", condition sits in the middle.

Finally, one study used the two-response paradigm: Bago, Bonnefon and De Neys (2019) found that very few subjects change their initial strategy, made under time pressure, even when they are asked to think for at least 10 seconds before deciding whether to switch strategy. As a consequence, most of the cooperative choices made under intuition, remained cooperative also under deliberation. See Table 1 for a short summary of each study.

In sum, the evidence about the effect of cognitive processing manipulations on cooperation in one-shot, anonymous, social dilemmas is mixed. To shed light on this question, Rand (2016) turned to meta-analytic techniques. He used random-effect meta-analysis to analyze 67 experiments in which participants' cognitive process was manipulated while making a pure cooperation decision, that is, he included in his meta-analysis, also studies using the trust game (only player 2), and sequential prisoner's dilemmas (only second movers). He also included two field experiments in which subjects had to decide whether to return a lost glove under different time constraints (Artavia-Mora, Bedi & Rieger, 2017; Artavia-Mora, Bedi & Rieger, 2018) and one study in which cooperating with one player implies competing with a third party (De Dreu, Dussel & Ten Velden, 2015). In doing so, he reported an overall positive effect of intuition on cooperation. Moreover, this positive effect was not present in strategic cooperation. He used this result to support the Social Heuristics Hypothesis (SHH). The SHH (Rand et al, 2012; Rand et al, 2014; Bear & Rand, 2016) contends that people internalize strategies that are optimal in everyday interactions, which are typically repeated, and use them as default strategies in novel settings, when their capacity to deliberate is impaired. Then, after deliberation, people override these social heuristics and tailor their behaviour towards the one that is optimal in the given

interaction. Therefore, the SHH predicts that intuition favours pure but not strategic cooperation, which is exactly what Rand found in his 2016 meta-analysis. However, this overall conclusion was recently challenged by Kvarven et al. (2019), who conducted an updated meta-analysis (82 experiments). In their meta-analysis, Kvarven et al. (2019) included only social dilemmas, that is, they did not include the trust game, the sequential prisoner's dilemma, and the two natural field experiments by Artavia-Mora and colleagues (2017, 2018). In doing so, they found an overall effect of intuition on cooperation, but, critically, they found that this effect was entirely driven by conceptual primes explicitly asking people to rely on their emotion. They argue that these primes might have induced demand effect. Moreover, the same result was obtained when adding, in the analysis, the trust game and the sequential prisoner's dilemmas, but not Artavia-Mora's field experiments, because they interpreted these works as studying as altruistic behaviour and not as cooperative behaviour (the person who lost the glove cannot reciprocate the action). In a response paper, Rand (2019) argued that there is no theoretical reason to exclude the trust game, the sequential prisoner's dilemma from the analysis, and the two field experiments, because they all meet the criteria for being classified as "pure cooperation". Adding these papers, Rand (2019) found a positive effect of intuition on pure cooperation, which remains significant also when restricting the analysis to studies which do not use the explicit primes.

Outlook and open problems

In this section, I shortly summarize the previous findings and I list a set of open problems that I think deserve attention in future research.

As described above, the evidence regarding a potential causal link between cognitive process and cooperation is mixed. Rand's (2016) meta-analysis sheds light on the topic by providing support for the hypothesis that, overall, Type 1 processing favours cooperative

behaviour, whereas Type 2 processing promotes non-cooperative behaviour. However, this conclusion was challenged by a larger meta-analysis, with different inclusion criteria (Kvarven et al., 2019), which, on turn, was challenged by an even larger meta-analysis (Rand, 2019). The overall conclusion that one can draw from this debate is that there is a need for more studies. In particular, most of the experiments meta-analysed in Rand (2016) used, as cognitive manipulation, time constraints (27 out of 51 studies) and conceptual primes of intuition (11 out of 51 studies). Among the remaining studies, only three of them used ego-depletion (Capraro & Cococcioni, 2016), one of them used foreign language (Urbin et al. 2016), while none of them used cognitive load. The remaining studies included in the meta-analysis either used the trust game, or they were obtained as last-rounds or average-across-rounds of iterated games (Duffy & Smith, 2014; Døssing, Piovesan, & Wengström, 2017, de Haan & van Veldhuizen, 2015; Osgood & Muraven, 2015), or regarded situations in which cooperating with one player implies competing with a third party (De Dreu et al. 2015). A similar asymmetry affects also the other two meta-analyses (Kvarven et al., 2019; Rand, 2019), which essentially differs from the original one primarily in the 21 time-pressure studies included in the preregistered replication by Bouwmeester et al. (2017). Additionally, the only study using neurostimulation (Li et al. 2018) is not included in any of these meta-analyses. Since both most time constraint and conceptual prime studies have been criticized, I believe that the current empirical evidence on the cognitive underpinnings of cooperation should be complemented by further experiments using, as cognitive process manipulations, ego-depletion, cognitive load, the two-response paradigm, and neurostimulation.

Open problem. Exploring the cognitive basis of cooperation using ego-depletion, cognitive load, the two-response paradigm, and neurostimulation.

33

Another important question regards the role of potential moderators. Alós-Ferrer and Garagnani (2018) found that the effect of time pressure on cooperation depends on the social value orientation, such that pro-socials tend to become more cooperative when deciding under pressure, whereas individualists tend to be unaffected by the manipulation and competitive subjects even tend to become more selfish. 12 To explain these results, they proposed that time pressure makes people more likely to behave according to their social value orientation, while time delay makes people readapt this initial response towards what they believe to be the norm in the given situation. I believe that testing this hypothesis is an important direction for future research. Note that, if confirmed, this hypothesis would be at odds with the SHH in terms of the predicted effect of deliberation on cooperation, but not in terms of the predicted effect of intuition. This is because the SHH makes predictions about the moderating role of the environment in which participants live. More precisely, the SHH contends that Type 1 processing promotes cooperation among participants who live in a cooperative setting, but not among those who live in a non-cooperative setting, who are predicted to remain unaffected by the cognitive manipulation. In agreement with this prediction, it has been found that trust in daily life moderates the positive effect of time pressure on cooperation (Rand & Kraft-Todd, 2014) and that time pressure does not have affect cooperative behaviour among participants living in a noncooperative setting (Capraro & Cococcioni, 2015). Although I am not aware of any study correlating a participant's SVO with the cooperativeness of the environment in which this participant lives, it would not be surprising if these two measures were correlated. Therefore, the predictions of the SHH are very similar to those of Alós-Ferrer and Garagnani's hypothesis, in terms of the effect of intuition. By contrast, while the SHH predicts no positive effect of

deliberation on cooperation on any subset of subjects, Alós-Ferrer and Garagnani's hypothesis predicts a positive effect of deliberation on cooperation among pro-selves.

Open problem. Exploring more in depth the effect of intuition and deliberation on cooperation as a function of participants' SVO.

Another open problem regards whether the effect of promoting intuition vs. deliberation on cooperation depends on the benefit-to-cost ratio. Kessler et al. (2017) found that this might be the case. In their experiment, time pressure increased cooperation for relatively small b/c, but the effect was reversed for bigger values. If confirmed, this effect would have important theoretical implications, because it would not be consistent with any of the frameworks previously mentioned, and it would require a new theorization. As a potential explanation, Kessler et al. (2017) proposed that Type 2 processing puts more weight on social efficiency.

Open problem. Investigating the effect of intuition and deliberation on cooperation as a function of the benefit-to-cost ratio.

Another important line of research regards the existence of a potential inverted-U effect of cognitive effort on cooperation. Study 3 in Moore and Tenbrunsel (2014) found that cognitive complexity has an inverted-U effect on cooperation in a non-incentivized, social dilemma based on the Shark Harvesting and Resource Conservation exercise. Capraro & Cococcioni (2016) used this result to explain why extreme cognitive depletions (extreme time pressure and egodepletion) seem to give rise to lower rates of cooperation: it is possible that the effect of cognitive effort on cooperative behaviour is non-linear. The theoretical rational for this assumption is that cooperation might be partly driven by an abstract desire to do the morally right thing (Capraro & Rand, 2018); and it has been long argued that the application of abstract moral rules is not immediate, but appears only at later stages of moral reasoning (Kohlber, 1963).

Therefore, exploring a potential inverted-U effect of cognitive reflection on cooperation is a fascinating and non-trivial question, with obvious theoretical and practical implications. One important observation, in my opinion, is that this question is inherently difficult because implementing extreme cognitive manipulations (e.g., extreme time pressure, or very hard cognitive load or ego depletion tasks) may interfere with participants understanding of the game or increase drop-out rates, leading to potential selection biases. Studies using neurostimulation might overcome these limitations (although they have others). The only study using neurostimulation that has been conducted so far found no evidence of an inverted-U effect: the control condition of Li et al. (2018) gave rise to a rate of cooperation between the two cognitive process manipulations. But, clearly, more evidence is needed to allow general conclusions.

Open problem. Exploring a potential inverted-U effect of cognitive reflection on cooperation. The use of neurostimulation could be particularly useful in this case.

I conclude with an observation regarding asymmetric social dilemmas. These dilemmas are important in light of practical applications, because, in reality, there is often power imbalance among people. This power imbalance can be modelled in terms of asymmetries between costs and benefits of cooperation. In spite of this practical importance, however, only two studies explored the cognitive underpinnings of cooperation in asymmetric dilemmas, and they led to opposite results: Lotz (2015) used conceptual primes and found that intuition promotes cooperation; conversely, Li et al. (2018) used neurostimulation and found that activity in the DLPFC is linked to cooperation. Clearly, more studies are needed on this topic.

Open problem. Exploring the cognitive underpinnings of cooperation in asymmetric social dilemmas.

The cognitive basis of altruism

We often pay costs to help others even in situations where the recipient of the help has no way to reciprocate our action. A typical example is when we buy some food for the homeless person sitting at the entrance of the supermarket. The notion of altruism formalizes the core of this type of interactions. Formally speaking, altruism is when a decision-maker pays a cost to help someone. The recipient of the altruistic action need not be a single individual; it could be also a group of individuals, as, for example, a humanitarian organization. The prototypical measure of altruism is the dictator game.

Measures of altruism: the dictator game

In the dictator game, the *dictator* is given a certain amount of money, and has to decide how much of it, if any, to donate to the *recipient*, who is initially given nothing. The recipient has no choice and only gets what the dictator decides to give.

Clearly, the dictator has no monetary incentive to give any part of their endowment to the recipient. Therefore, their payoff-maximizing strategy is to give nothing. In contrast to this prediction, however, laboratory experiments have repeatedly shown that a significant proportion of dictators actually give part of their endowment. According to Engel's (2011) meta-analysis of 616 dictator game experiments, the average donation is 28.3% of the endowment and the distribution of donations is typically bimodal, with main modes at "giving nothing" and "giving half". One might wonder whether altruistic behaviour, as measured using the dictator game, represents actually a different kind of behaviour than cooperative behaviour measured through social dilemmas. The answer is positive: although they are typically correlated (Capraro et al. 2014; Peysakhovich et al. 2014), the causal link holds only in one direction: people who give in the DG typically cooperate in the PD, but not the converse (Capraro et al. 2014).

In the standard version of the dictator game the recipient is a person. However, one can consider variants in which the recipient is, for example, an organization. A variant that has received considerable attention takes the name of charitable giving, where the recipient of the donation is a charitable organization. In this review, I will focus on both standard dictator game experiments and charitable giving experiments.

Another variant of the dictator game can be obtained by changing the ratio between the cost and the benefit of the altruistic action. For example, when the cost of helping is smaller than the benefit created, the altruistic action does not only help the other person, but also increases social efficiency. This might introduce a confounding factor, especially in light of some preliminary research suggesting that equity and efficiency might be cognitively different (see the final section). For this reason, I opted for dividing the review of the empirical evidence regarding the cognitive basis of altruism in two subsections. In the next section, I will review only the studies implementing a dictator game where the cost of the altruistic action is equal to its benefit; in the section after, I will review also the literature on the cognitive basis of altruism in dictator games with varying cost-to-benefit ratio.

Review of the empirical evidence

Several scholars explored the effect of time constraints on altruistic behaviour.

Carlson, Aknin and Liotti (2016) conducted a series of charitable giving experiments under time pressure versus time delay and found no differences in donations depending on the cognitive process manipulation. Along the same lines, Tinghög et al (2016) found that time pressure does not affect charitable donations to four different humanitarian organizations (Save the Children, WWF, Doctors Without Borders, and UNICEF). Jarke and Lohse (2016) found that time pressure has no effect on altruistic decisions in a binary dictator game in which dictators have to choose

between (5,5) and (8,2). Kessler, Kivimaki and Niederle (2017) found that time pressure has no effect on giving in the standard dictator game. Similarly, Andersen, Gneezy, Kajackaite and Marx (2018) found that allowing dictators to reflect about their decision for one day does not alter their amount of giving. However, Mrkva (2017) found that time delay, compared to baseline, increases charitable donations, whereas Gärtner (2018) found that time pressure increases self-regarding decisions in the dictator game, but only when the dictator game is presented without a status-quo; by contrast, Gärtner (2018) found that time pressure has no effect on decisions when the statusquo is set on the self-regarding choice or on the altruistic choice. Chuan, Kessler and Milkman (2018) found that 30-day time delay between the provision of medical care and a donation solicitation decreases the likelihood of a donation by 30%. Grolleau, Sutan, El Harbi and Jedidi (2018) found that time delay decreases dictator game giving in Tunisia. Artavia-Mora, Bedi and Rieger (2017) and Artavia-Mora, Bedi and Rieger (2018) found that time pressure increases help in a natural field experiment in which participants had to return a lost glove. Finally, Strømland & Torsvik (2019) found no effect of time pressure on dictators' giving on a large representative sample of the Norwegian population; they also measured participants level of experience in the game and found no interaction between time pressure and experience in determining the donation level.

Several studies have investigated the effect of cognitive load on altruism. An earlier paper by Roch, Lane, Samuelson, Allison and Dent (2000) found that cognitive load increases the amount that people take from a common pool. Cornelissen, Dewitte and Warlop (2011) explored the differential effect of cognitive load on individualists and pro-socials. They found that cognitive load favours altruistic choices in the dictator game, but only among pro-social individuals; in contrast, non-cognitively loaded pro-socials gave as much as non-cognitively

loaded individualists, whose behaviour was not affected by the cognitive manipulation. Schulz, Fischbacher, Thöni and Utikal (2014) conducted a series of 20 mini dictator games with varying inequality but keeping the social efficiency constant (100 points for 10 games and 94 for 10 games). In doing so, they found that cognitive load increases generosity. However, several other works failed to find a significant effect of cognitive load on altruism. Benjamin, Brown and Shapiro (2013) found no effect of cognitive load on dictator game giving. Kessler and Meier (2014) conducted a cognitive load manipulation on charitable giving to the Red Cross and found no effect. This result was replicated in several studies. Tinghög et al (2016) found no effect of cognitive load on charitable giving. Hauge, Brekke, Johansson, Johansson-Stenman and Svedsäter (2016) found no effect on cognitive load on dictator game donations, both when the game is played in its standard "give frame" and when it is played in its "give frame" variant. Grossman and Van der Weele (2017) found no effect of cognitive load on charitable giving.

Only three studies applied conceptual primes during a decision involving altruistic behaviour. An earlier paper by Small et al. (2007) found that priming intuition versus deliberation increases giving to a charity, but only when the victim is identifiable and not statistical. Consistent with this result, Zhong (2011) found that participants give more to a charity organization (Child Family Health International) when they are asked how much they *feel* to give compared to when they are asked how much they *decide* to give. However, Kinnunen and Windmann (2013) found no differences in donations to Greenpeace.

The effect of ego depletion on giving was explored by a multitude of studies. Halali, Bereby-Meyer and Oxenfels (2013) found that depleted participants choose the equal split in the dictator game less often than non-depleted participants. Along the same lines, Achtziger, Alós-Ferrer and Wagner (2015) found that depleted dictators give less than non-depleted ones.

Itzchakov, Uziel and Wood (2018) found that depleted participants donate less to charity than non-depleted ones, but only when they see a persuasion message; in contrast, there was no effect in the baseline, when the persuasion message was not present. Particularly interesting is the work of Banker, Ainsworth, Baumeister, Ariely and Vohs (2017). They found that depleted participants give less than non-depleted ones in the standard dictator game; however, they also found that the effect flips in the dictator game in the take frame, where the endowment is initially given to the recipient. They interpret this finding as suggesting that ego depletion neither makes people more selfish nor more altruistic: it simply makes them more likely to adopt the status quo. Dickinson and McElroy (2017) found that sleep restriction decreases altruistic behaviour in the dictator game. Ferrara et al. (2015) found that sleep deprivation decreases dictator game giving, but only for females. Rantapuska et al. (2017) found that hunger does not affect charitable giving. A recent study found that self-control depletion increases helping, sharing, and volunteering in high-crisis scenarios, such as in earthquake simulation (Wang, Zhang, Li & Xie, 2019). Another recent study found that sleep deprivation decreases charitable giving, as well as other forms of civic engagement, such as voting and signing petitions (Holbein, Schafer & Dickinson, 2019).¹³

Three studies explored the effect of brain stimulation on altruistic behaviour. Ruff,
Ugazio and Fehr (2013) found that transcranial direct current stimulation of the rLPFC decreased
dictator game giving, while disruption of the rLPFC increased giving, relative to the placebo,
"sham", condition. Strang et al. (2015) used the same technique, transcranial magnetic
stimulation, but they focussed on a specific subregion of the rLPFC, the rDLPFC. They found
that disrupting the rDLPFC led to less giving, compared to two different baselines: disruption of
the lDLPFC and the placebo, "sham", condition. A more recent study argued that disruption of

the rLPFC does not simply increase altruism or selfishness, but it has the effect of making individuals more likely to follow rules of behaviour (Gross et al., 2018).

Finally, only one study used the two-response paradigm to explore the cognitive basis of altruism: Bago, Bonnefon and De Neys (2019) found that very few subjects change their initial strategy, made under time pressure, even when they are asked to think for at least 10 seconds before deciding whether to switch strategy. As a consequence, most of the altruistic choices made under intuition, remained altruistic also under deliberation. See Table 2 for a short summary of each study.

Given the mixed evidence above, scholars have recently turned to meta-analytic techniques, looking also at the effect of potential moderators. In particular, one moderator that has attracted considerable attention is gender. There is a theoretical reason for this interest. Previous work suggests that men and women tend to behave according to stereotypes regarding their social roles, with women being more communal and unselfish, and men being more agentic and independent (Eagly, 1987). In agreement with this view, empirical research repeatedly found that women tend to be more altruistic than men in the dictator game (Croson & Gneezy, 2009; Brañas-Garza, Capraro & Rascón-Ramírez, 2018), that women are expected to be more altruistic than men in the dictator game (Aguiar, Brañas-Garza, Cobo-Reyes, Jimenez & Miller, 2009; Brañas-Garza, Capraro & Rascón-Ramírez, 2018), and that when women behave in a way that is perceived to be insufficiently altruistic, they are liked less, less likely to be helped, hired, promoted, paid fairly, and given status (Heilman & Okimoto, 2007). Therefore, to the extent to which Type 1 reactions reflect internalized strategies that people quickly access when they have no opportunity to reason about the specific situation at hand, one might expect that intuition promotes altruism for women but not for men.

Following this idea, Rand, Brescoll, Everett, Capraro and Barcelo (2016) meta-analysed 22 dictator game experiments (total N=4,366) manipulating cognitive process. They found that, indeed, Type 1 promotes altruism for women but not for men. Moreover, in line with the aforementioned interpretation, in a second study published in the same work, the authors also reported that the more women described themselves using masculine attributes relative to feminine attributes, the more Type 2 reduced their altruism. This suggests that the specific negative effect of deliberation on intuition among women was partly explained by identification with the other gender.

However, a more recent meta-analysis (Fromell, Nosenzo & Owens, 2018) found an interaction between gender and cognitive mode, in a slightly different direction than Rand et al. (2016). Specifically, Fromell et al. (2018) meta-analysed 19 studies (total N = 10,898) and, in line with Rand et al. (2016), they found that there is an interaction between gender and cognitive process manipulation; however, in contrast to Rand et al. (2016), they found that promoting Type 1 has no effect on women, whereas it has a marginally significant negative effect on men. 14

In sum, two meta-analyses of the cognitive basis of altruistic behaviour agree on the existence of a differential effect on women and men, but they disagree on the main effects: Rand et al. (2016) argue that Type 1 favours altruism for women but not for men; Fromell et al. (2018) argue that Type 1 disfavours altruism for men but not for women.

Altruism with varying cost-to-benefit ratio

For completeness, in this section I will review also the line of research exploring the cognitive basis of altruism in situations in which the cost of the altruistic action is different than its benefit.

The majority of these studies used time constraints. Capraro, Corgnet, Espín and Hernán-González (2017) built on a previous correlational study (Corgnet et al. 2015) and conducted a series of mini-dictator games that were used to classify participants into self-regarding, spiteful, inequity averse, or socially efficient. The classification was done in two ways: using a generalized version of the Fehr & Schmidt's (1999) model; and using consistency across choices. Independently of the classification method being used and both in the USA and in India, Capraro et al. (2017) found that time pressure (and low scores in the CRT) were associated with spitefulness and inequity aversion, while time delay (and high scores in the CRT) were associated with social efficiency. In line with these findings, Kessler, Kivimaki and Niederle (2017) found that deliberation favours generosity, when generosity is socially efficient. Chen and Krajbich (2018) conducted a sequence of 200 binary dictator games, with varying cost to benefit ratios. They classified participants into pro-socials and individualists, by using Fehr & Schmidt's (1999) model. In doing so, Chen and Krajbich (2018) found that time pressure made pro-socials more prosocial and individualists more self-regarding. Krawczyk and Sylwestrzak (2018) conducted a battery of dictator games with varying cost to benefit ratios and using the tworesponse paradigm: participants were incentivized to give a fast choice, which then they were able to revise. In doing so, they found that time pressure increased envy, that is, pressured participants were more likely to decrease the payoff of the other person, but only when this payoff was higher than their own. Finally, Merkel & Lohse (2019) conducted a time pressure vs time delay vs no time constraint experiment using four binary dictator games with varying costto-benefit ratios. The cost to benefit ratio was, in some cases, smaller than 1, and, in other cases, larger than 1. They found that time pressure does not affect the rate of altruism, independently of the cost-to-benefit ratio.

Beyond time constraints, I am aware of only two studies using ego-depletion. Friehe and Schildberg-Hörisch (2017) found no effect of ego-depletion on a taking game in which the recipient is given 20 points and the dictator can take any amount from the recipient, knowing that any amount taken would be divided by 4. Balafoutas, Kerschbamer and Oexl (2018) found no effect of ego-depletion on a series of mini-dictator games built to measure aversion to advantageous and disadvantageous inequality.

Outlook and open problems

As described above, the work looking at the effect of cognitive manipulations on altruistic behaviour shows mixed results. However, at the meta-analytic level, there is agreement that intuition and deliberation have no overall effect on altruistic behaviour. This, clearly, does not imply that the effect is not present in any subclass of people. Instead, this leads to a general research question regarding the role of potential moderators in determining for which classes of people altruism turns out to be the automatic or the deliberative response.

One important moderator is gender. Two meta-analyses reached a similar conclusion: the cognitive basis of altruism among women differ from the cognitive basis of altruism among men. However, these two meta-analyses disagree on the direction of the main effects: Rand et al. (2016) argue that intuition favours altruism for women but not for men; Fromell et al. (2018) argue that intuition disfavours altruism for men but not for women. Therefore, a primary direction for future research is to clarify the interaction between cognitive process and gender in determining altruistic behaviour.

Open problem. Clarifying the interaction between cognitive process and gender in determining altruistic behaviour.

Another promising moderator is the SVO, or similar methods to classify participants according to their social preferences. Only one study investigated this moderator. Cornelissen et al. (2011) explored the differential effect of cognitive load on individualists vs pro-socials (classified according to their SVO measured using Leibrand's (1984) ring measure) and

found that cognitive load favours dictator game giving, but only among pro-socials; by contrast, non-cognitively loaded pro-socials individuals gave as much as non-cognitively loaded individualists. Another study looked at a similar question, but with dictator games with varying cost to benefit ratios and reported similar results (Chen and Krajbich, 2018). Therefore, this preliminary evidence suggests that the role of intuition on altruism might be moderated by social preferences. Future research should explore more in depth this possibility.

Open problem. Exploring the interaction between cognitive process and SVO (or similar constructs) in determining altruistic behaviour.

The aforementioned work by Banker et al. (2017) suggests that ego-depletion does not actually affect altruistic behaviour, but simply makes people more likely to choose the statusquo. However, with the exception of Gärtner (2018), no one has tried to explore the same potential confound using other cognitive manipulations, and possibly looking at the interaction between with gender and SVO. It is, indeed, in principle possible that the overall effects reported in the meta-analyses are confounded with the status-quo bias.

Open problem. Exploring the cognitive underpinnings of altruistic behaviour by controlling for status-quo bias.

The cognitive basis of lying

The conflict between lying and truth-telling is at the core of all social and economic interactions that include communication with asymmetric information. In these cases, people

may be tempted to misreport their private information for their benefit. Indeed, previous empirical work, as well as everyday experience, suggests that some people take advantage of their private information and use it for their own benefit. As a consequence, lying has an enormous negative impact on people, companies, and the society as a whole. For example, tax evasion costs about 100 billion a year only to the US government (Gravelle, 2009). Understanding whether intuition favour or disfavour truth-telling is therefore a topic of primary importance.

Measures of lying

There are several ways to measure people's tendency to misreport the truth. These measures can be classified in two types: those where lying harms an "abstract" person (e.g., the experimenter), and those where lying harms a "concrete" person (e.g., another participant in the experiment). 15

Lying tasks (when lying harms "abstract others")

A simple way to measure people's (dis)honesty in situations in which lying harms an abstract person is to have them complete a task and then pay them according to the self-reported performance in this task. In this case, the abstract person that is harmed by the lie is the experimenter.

A popular task that has been used for this purpose is the so-called "matrix task" (Mazar, Amir & Ariely, 2008; Gino, Ayal & Ariely, 2013). In its proto-typical version, participants are presented with a series of twenty 4x3 matrices, whose entries are all three-digit numbers between 0 and 10 (e.g, 1.34). Participants are given four minutes to find two numbers per matrix that add up to 10. Finally, participants are paid according to the self-reported performance in the task. Of course, this is not the only task that fits this purpose. For example, Muraven, Pogarsky and

Shmueli (2006) used logic puzzles that were modified in such a way to be unsolvable; Zhong (2011) used maths problems; Pitesa, Thau and Pillutla (2013) used performance in the Raven matrix test.

Another popular lying task is the dice-under-cup task (or dice-rolling task). Here, participants roll a dice, in private (under a cup), and then they are paid according to the reported outcome (Fischbacher & Föllmi-Heusi, 2013).

The sender-receiver game (when lying harms "concrete others")

Lying when it harms concrete others is typically measured through the sender-receiver game. This game has been initially introduced by Gneezy (2005). In this game, the *sender* and the *receiver* are informed that there are two possible allocations of money between the sender and the receiver; however, crucially, only the sender is informed about the exact monetary payoffs corresponding to these two options. Then, the sender sends a message to the *receiver*, chosen among two possible messages: Message A: "Option A gives you a better payment than Option B"; Message B: "Option B gives you a better payment than Option A". Finally, the receiver decides which of the two options to implement as a payment.

The sender-receiver game generates a potential confound, known as "sophisticated deception" (Sutter, 2009): there might be senders who tell the truth not because they want to be honest, but because they want to deceive the receiver, but, at the same time, they believe that the receiver would not believe them. In order to eliminate this confound, scholars have introduced several variants of the sender-receiver game in which either the receiver does not make any decision or his decision has no effect on the sender's payoff (Biziou van Pol, Haenen, Novaro, Occhipinti-Liberman & Capraro, 2015; Cappelen, Sørensen and Tungodden, 2013; Gneezy, Rockenbach & Serra-Garcia, 2013).

One of the reasons why the sender-receiver game is particularly useful is because it allows researchers to distinguish four types of lies, depending on their monetary consequences. Specifically, following the taxonomy of Erat and Gneezy (2012), I define:

- Black lies, those that benefit the sender at a cost to the receiver
- Spiteful lies, those that harm both the sender and the receiver
- Altruistic white lies, those that help the receiver at a cost to the sender
- Pareto white lies, those that help both the sender and the receiver

Nevertheless, the vast majority of the studies that explored the cognitive underpinnings of lying using the sender-receiver game focused on black lies. Therefore, unless explicitly stated differently, in what follows I will always assume that lying in the sender-receiver game increases the sender's payoff at a cost to the receiver.

Review of the empirical literature

A number of studies explored the effect of time constraints in the decision to lie. Gunia, Wang, Huang, Wang and Murnighan (2012) found that three minutes of contemplation increases honesty in a sender-receiver game, compared to the baseline, with no time constraint.

Symmetrically, Shalvi, Eldar and Bereby-Meyer (2012) found that a 20-second time pressure increases cheating in a dice-rolling task, compared to a condition with no time constraint.

However, this effect was not replicated in two preregistered replications (Van der Cruyssen, D'hondt, Meijer & Verschuere, 2019). These results have also been challenged on a theoretical ground by four more recent studies. Foerster, Pfister, Schmidts, Dignath and Kunde (2013) criticized previous experiments because participants could in principle generate their answer before the cognitive manipulation and therefore the truth need not be inhibited. To avoid this confound, Foerster et al. (2013) implemented a modified design in which participants have to

report the outcomes of a series or dice roll either immediately after the roll or after a delay. They found that participants who are asked to report the outcome immediately are more honest than those under time delay. Along the same lines, Capraro (2017) and Capraro, Schulz and Rand (2019) found that a 5-second time pressure, compared to a 30-second time delay, increases honesty in a sender-receiver game in which the private information was communicated at the same time as the timer started. Similarly, Lohse, Simon and Konrad (2018) found that time pressure increases honesty in a situation in which participants were put, without prior knowledge, in a condition in which they could misreport an information for their benefit; moreover, they showed that the effect was due to a lack of awareness of the opportunity to lie. By contrast, Andersen et al. (2018) explored the effect of a long reflection time (one-day) on cheating on a game similar to a dice-rolling task; in doing do, they found no effect.

Only one study used cognitive load to manipulate cognitive process during a lying task: van't Veer, Stel and Van Beest (2014) found that cognitive load increases honesty in a dicerolling experiment.

In contrast, several studies investigated the effect of conceptual primes on truth-telling. Zhong (2011) found that promoting deliberation, through a math problem-solving task, versus promoting emotional-based choices, through a task in which participants have to report their feeling in a number of situations (Hsee & Rottenstreicht, 2004; Small, Loewenstein and Slovic, 2007), favours lying in a sender-receiver game. These results thus suggest that emotions favour honesty, while deliberation favours dishonesty. Zhong (2011) also replicated this finding with a different manipulation consisting of framing a choice as a decision rather than an intuitive reaction. Cappelen et al. (2013) studied the cognitive basis of lying in a sender-receiver game with Pareto white lie payoffs. The choice of setting the payoffs in the Pareto white lie domain

was justified by the aim of looking at intrinsic lying aversion, cleared by any confound that could potentially be brought in by the monetary consequences of lying. In fact, in Cappelen et al. (2013), lying maximized the individual payoff and the social welfare and, additionally, it minimized the inequity between the sender and the receiver. Consequently, in Cappelen et al. (2013), telling the truth would be entirely driven by lying aversion. In doing so, the authors found that promoting intuition favours honesty, but only for males. Finally, Bereby-Meyer et al. (in press) found that using a second language increases honesty.

The effect of ego-depletion on lying has also been largely explored. Results generally agree that ego-depletion favours dishonest behaviour in lying tasks. ¹⁶ Muraven, Pogarsky and Shmueli (2006) found that ego-depletion promotes lying in a task in which participants are paid according to the number of problems they report having solved. Mead, Baumeister, Gino, Schweitzer and Ariely (2009) found that ego-depletion increases dishonesty in a matrix task. Additionally, they also found that depleted participants are more likely than non-depleted ones to expose themselves to the possibility of cheating. Gotlib and Converse (2010) found that egodepletion increases dishonesty (in the form of leaving a room before the designated end time). Gino, Schweitzer, Mead and Ariely (2011) replicated the result that ego-depletion increases dishonesty in a matrix task and, additionally, they showed that resisting temptations to behave unethically both required and depleted self-control resources. Pitesa, Thau and Pillutla (2013) found that impairing cognitive control increases cheating in a situation in which participants have to pay themselves according to the performance in a Raven matrix task, but only when the negative consequences of cheating on other people are not made salient. If these consequences were made salient, then the effect even reverses: impairing cognitive control increases honesty. Along these lines, Yam, Chen and Reynolds (2014) studied the effect of ego-depletion on

dishonesty as a function of social consensus about the badness of the dishonest action. They found that ego-depletion increases dishonesty only in situations in which there is low social agreement that the dishonest action is actually bad; conversely, when there is high consensus about the badness of the dishonest action, there is a reversal of the effect, such that ego-depletion increases honesty. Chiou, Wu and Cheng (2017) found that lowering men's self-control, by showing them pictures of attractive women, increases dishonesty in a matrix task (the same egodepletion technique did not seem to work for women, whose self-control resources were not affected by pictures of attractive men). Yam (2018) found that suppressing ethics-unrelated thoughts – a technique known to deplete self-control (Richerson & Shelton, 2004; Gordijn et al. 2004) – leads to more cheating in a task in which participants are paid according to the performance they report. Interestingly, when asked to suppress ethics-related thoughts, participants becomes more honest, suggesting that suppressing ethics-related thoughts can increase moral awareness. The only study finding no effect is by Welsh, Ellis, Christian and Mai (2014), who found no effect of sleep deprivation on deceiving behaviour in a sender-receiver game.

Two studies explored the effect of brain stimulation on honesty. Maréchal, Cohn, Ugazio and Ruff (2017) found that transcranial direct current stimulation of the rDLPFC increases honesty in a die-rolling task, where lying benefits the liar; by contrast, the same effect was not present in the variant of the same task in which lying benefits another person. Gao, Yang, Shi, Lin and Chen (2018) used transcranial direct current stimulation over the DLPFC during a sender-receiver game. In doing so, they found that, in the sham condition, males are more honest than females, while stimulation of the right DLPFC makes gender differences to disappear by increasing honesty for females, but not for males. However, the authors themselves noted that

one potential limitation of their study is that the rate of honesty among males in the sham condition was already very high, and this implied that there was little space for statistical changes. See Table 3 for a short summary of each study.

As just reviewed, the evidence on the cognitive underpinnings of truth-telling is mixed. Interestingly, however, a recent meta-analysis does find an overall effect, but this effect depends on whether lying harms abstract others or concrete others. Specifically, Köbis, Verschuere, Bereby-Meyer, Rand and Shalvi (2019) meta-analysed all the studies manipulating cognitive process during any task involving lying and looked at variability in the proportion of liars (73 studies, N=12,711) and in the magnitude of liars (50 studies, n=6,473). In doing so, they found that promoting Type 1 processing increases the proportion of liars and the magnitude of lying in cases in which lying harms abstract others, but not in cases in which lying harms concrete others. In the latter case, promoting Type 1 over reflection has no effect on dishonest behaviour. Köbis et al. (2019) interpreted these results in lights of the Social Heuristics Hypothesis: when lying harms abstract others, the negative consequences of lying are not salient and participants tend to act in a self-interested way;¹⁷ in contrast, when lying harms concrete others and the negative consequences of lying are salient, self-interest conflicts with the intuitive social heuristics "do not harm", that pushes in the opposite direction, and nullifies the effect of promoting intuition on dishonesty.

Outlook and open problems

Virtually all studies explored the cognitive underpinnings of black lies. Only one study implemented Pareto white lies (Cappelen et al, 2013) and another one investigated altruistic white lies (Maréchal et al, 2017). Another study explored the role of ego-depletion on the proclivity to tell an altruistic white lie in hypothetical situations (Cantarero & Van Tilburg, 2014).

Exploring the cognitive basis of lies other than black lies is an important direction for future research for different reasons. Pareto white lies are of great theoretical interest, because they represent the setting in which one can measure intrinsic lying aversion, depurated from any other confound (e.g. in the case of black lies, telling the truth coincides with being altruistic). It is noteworthy that the only study that manipulated cognitive process in Pareto white lie situations found honesty to be intuitive, providing a first piece of evidence that lying aversion is automatic. However, and especially in light of the Reproducibility Crisis (Open Science Collaboration, 2015), more studies should be conducted on this topic to strengthen the conclusions. It would also be interesting to explore the effect of promoting intuition versus deliberation in the decision to tell altruistic white lies and spiteful lies. In this case, predictions are not obvious. Previous research suggests that time pressure increases spiteful behaviour (Capraro et al. 2017). Therefore, it might be possible that it also increases spiteful lying. Similarly, previous work suggests that promoting intuition has no overall effect on altruism (Rand et al. 2016; Fromell et al. 2018). This suggests that there might be no overall effect of intuition on altruistic white lies. In any case, these are interesting questions that are worth exploring in future research.

Open problem. Exploring the cognitive basis of honesty as a function of lie type (black lies, spiteful lies, altruistic white lies, and, especially, Pareto white lies).

As reviewed above, there has been some research looking at the cognitive basis of honesty as a function of gender. Gender is known to influence lying and, by now, there are three meta-analyses showing that men lie more than women (Capraro, 2018; Abeler, Nosenzo & Raymond, 2019; Gerlach, Teodorescu & Hertwig, 2019). Therefore, it is natural to ask whether gender interacts with cognitive mode in determining the decision whether to lie. However, only two studies explored this topic. They both found gender differences in the cognitive basis of

honesty, but they disagreed on the details of this difference. Specifically, Cappelen et al. (2013) found that conceptual priming of intuition favours honesty (in a sender-receiver game with Pareto white lies payoffs), but only among males. Gao et al. (2018) found that transcranial direct current stimulation over the DLPFC increased honesty (in a sender-receiver game with black lies payoffs), but only for females. More studies should contribute to this debate.

Open problem. Exploring the interaction between gender and cognitive mode in lying tasks.

In most experiments on lying, the payoffs maximizing strategy is known to the participants since before the starting of the cognitive manipulation. For example, in the typical dice-rolling task, it is evident that the payoff maximizing strategy is to report that the outcome of the roll is 6. A similar situation happens with the typical sender-receiver game, in which the sender is informed about the payoffs associated with the two available options before the cognitive manipulation. This has two consequences. First, participants can in principle make a decision before the cognitive manipulation. Second, the (generalization of the) Social Heuristics Hypothesis cannot be applied to this case. According to the SHH, people bring to the laboratory heuristics learned in their everyday life and use them when they are in a situation in which they are depleted of the cognitive resources needed to compute the payoff maximizing strategy. If the payoff maximizing strategy is already known, there is no room for the SHH. Motivated by this observation, two recent studies implemented a variant of the sender-receiver game in which participants were not initially informed about the payoff-maximizing option, but they had to infer it during the cognitive manipulation (Capraro, 2017; Capraro et al. 2019). Both these studies found evidence that time pressure increased honesty. A conceptually similar result was obtained by Lohse et al. (2018), who found that, when people are not initially aware of the possibility of

lying, time pressure increases honest behaviour. These results provide a first piece of evidence for a general form of the SHH, according to which, when participants do not have access to the payoff maximizing option, they tend to act honestly, arguably because honesty is payoff maximizing in the long-term, and therefore people internalize it as a useful heuristics. In any case, more works should be done to validate this theory by, for example, implementing lying tasks where the cognitive process manipulation is crossed with a manipulation of the moment in which participants are informed about the payoff maximizing option.

Open problem. Exploring the cognitive basis of honesty in situations in which participants do not initially know their payoff-maximizing strategy.

The cognitive basis of reciprocity

In the previous sections, I reviewed the cognitive basis of human sociality in situations in which either there is only one decision-maker (e.g., dictator games and lying tasks) or, in cases in which multiple decision-makers make decisions either simultaneously (e.g., prisoner's dilemmas, public goods games) or without knowing each other's payoff maximizing strategy (e.g., sender-receiver games). However, in reality, many interactions possess an element of reciprocity: the first mover makes a choice knowing that the second mover can respond to this choice.

Particularly relevant is the case in which the first mover has to choose between a self-regarding choice and an other-regarding choice, knowing that the second mover (or a third-party) will have the opportunity to punish or reward the choice made by the first mover. In this section, I focus on this type of interactions.

Measures of reciprocity

Measures of negative reciprocity

Negative reciprocity refers to situations in which one person has the choice to pay a cost to decrease the payoff of a decision-maker in response to a choice made by the decision-maker. The standard ways to measure negative reciprocity are through the ultimatum game (player 2) and through a social dilemma game followed by a punishment phase. In this latter case, the punisher might or might not be the target of the initial action. Situations in which the punisher is the target of the initial action take the name of second-party punishment; situations in which the punisher is not the target of the initial action take the names of third-party punishment. Below, I describe these two kinds of interactions in more details.

Second-party punishment

There are two standard ways to measure second-party punishment. The first one is through the second mover in the ultimatum game. In this game, Player 1, the *proposer*, has to propose how to divide a sum of money between her/himself and Player 2, the *responder*. The responder can either accept or reject the offer. In case the responder accepts the offer, then the money is split between the two players as agreed; in case the responder rejects the offer, neither the proposer nor the responder gets any money. Therefore, rejecting the proposer's offer corresponds to paying a cost (equal to the offer received) to decrease the payoff of the proposer.

The other standard way to measure second-party punishment is through a social dilemma followed by a punishment phase, in which participants can pay a cost to decrease the payoff of other players. The details can be implemented in different ways. For instance, as a social dilemma, instead of the public goods game, one can use the prisoner's dilemma; or, in the punishment phase, punishers, instead of targeting one particular player to be punished, they can punish all other players indiscriminately.

Third-party punishment

Third-party punishment is similar to second-party punishment, with the difference that the punisher is not the target of the first action, but only observes the behaviour of other participants, and then decide whether to punish them or not.

Measures of positive reciprocity

Positive reciprocity refers to situations in which one person can pay a cost to increase the payoff of a decision maker in response to a choice made by the decision-maker. Also in this case there are two types of positive reciprocity depending on whether the decision maker is or is not the recipient of the action to be rewarded.

Second-party rewarding

There are two standard ways to measure second-party rewarding. The first one is through the second mover in the trust game. In this game, Player 1, the *investor*, has to decide whether to transfer a sum of money to Player 2, the *investee*. Any amount invested is multiplied by a factor greater than 1 (typically equal to 3) and given to the investee. The investee has to decide how much, if any, to return to the investor. Therefore, returning money corresponds to paying a cost (equal to the amount returned) to increase the payoff of the investor.

The other standard way to measure second-party rewarding is through a social dilemma followed by a rewarding-stage, in which players can pay a cost to increase each other's payoff in response to the social dilemma interaction.

Third-party rewarding

Third-party rewarding differs from second-party rewarding in that the rewarder is not the recipient of the action to be rewarded, but only observes it.

Measures of expectation of reciprocity

The complementary notion of reciprocity is *expectation of reciprocity*, situations in which one person makes a decision knowing that someone else will have the chance to respond to this choice, by punishing it or rewarding it. Clearly, also this notion spits into two parts: *expectation of negative reciprocity* and *expectation of positive reciprocity*. Expectations of reciprocity can be measured through the ultimatum game (player 1), the trust game (player 1), and social dilemmas followed by a punishment or rewarding stage.

Review of the empirical evidence on negative reciprocity

Several studies explored the effect of time constraints on ultimatum game rejections. The empirical evidence consistently showed that time pressure makes responders more likely to reject low offers. Sutter, Kocher and Strauß (2003) found that responders under a 20 seconds time pressure are more likely to reject low offers, compared to responders with a 100 seconds time delay. Symmetrically, Grimm and Mengel (2011) found that a 10 minutes time delay makes responders more likely to accept lower offers, compared to the baseline condition, with no time constraint. A similar result was obtained by Wang et al. (2011), who found that a 5minute time delay decreases rejections of low offers, compared to baseline. These findings have been conceptually replicated respectively by Cappelletti, Güth and Ploner (2011), who found that a 30 seconds time pressure increases rejection rates compared to a time delay of 180 seconds, and by Neo, Yu, Weber and Gonzalez (2013), who found that a 15-minute time delay decreases rejection rates of low offers, compared to the baseline with no time constraint. More recently, scholars have started exploring the effect of potential moderators. Ferguson, Maltby, Bibby and Lawrence (2014) found that whether rejection rates are intuitive or deliberative depends on the offer: for the mildly unfair offer (60:40), time delay favours rejection; for the unfairer offers (90:10), (80:20), and (70:30), time delay has no effect on rejection rates. Oechssler, Roider and

Schmitz (2015) found that a 24-hour cooling off period decreases the rejection rate of UG unfair offers, but only when the offers were made using lottery tickets, and not when they were made using cash. Finally, Balafoutas and Tarek (2018) found that time pressure increases rejection rates in the Impunity Game, which is the variant of the Ultimatum Game in which rejecting the proposer's offer has no effect on the proposer's payoff: if the responder rejects the offer, the responder gets nothing, while the proposer still gets their endowment minus the offer.

Only one study used conceptual primes. Interestingly, its results contrast those obtained using time constraints. Indeed, Hochman, Ayal and Ariely (2015) found that asking responders to choose following their gut feelings increases the acceptance rate of unfair offers, compared to when they are asked to thoroughly consider the available information.

Two studies explored the effect of cognitive load. Both of these studies found no effect (Cappelletti et al, 2011; Olschewski, Rieskamp and Scheibehenne, 2018). Additionally, Olschewski et al. (2018) found that cognitive load significantly decreases choice consistency and they argue that this effect might confound with the effect of promoting intuitive choices: it is possible that effects that were previously attributed to increased intuition are actually driven by increased choice inconsistency. Relevant is also the work of Duffy and Smith (2014), who explored the evolution of cooperation in an iterated prisoner's dilemma played under cognitive load and compared it with the one under no load. They found that the behaviour of people under cognitive load converges to the Nash equilibrium at a slower rate than the behaviour of people under low load. Interpreting defection in the iterated prisoner's dilemma as a form of punishment, this finding is in line with the view that people under high cognitive load punish less than those under low load.

Numerous studies investigated the effect of ego-depletion. An earlier series of works explored the effect on ultimatum game responders of depleting or enhancing serotonin, a neurotransmitter implicated in self-control (Carver, Johnson & Joormann, 2008). Crockett et al. (2008) found that participants with depleted levels of serotonin rejected a greater proportion of unfair but not fair offers. The fact that serotonin depletion is associated to punishment in the ultimatum game has been also replicated by Crockett et al. (2010a). Symmetrically, Crockett et al. (2010b) found that enhancing serotonin makes participants less likely to reject unfair offers. Consistent with this line of research, Anderson and Dickinson (2010) found that a night of sleep deprivation increases the minimum acceptable offer among ultimatum game responders; Morewedge, Krishnamurti and Ariely (2014) found that alcohol intoxication makes participants more likely to reject low offers; Halali, Bereby-Meyer and Meiran (2014) found that participants depleted through a Stroop task are more likely to reject low offers, a finding that has been also replicated by Liu, He and Dou (2015). There are, however, also some works pointing towards the opposite direction: Achtziger, Alós-Ferrer and Wagner (2016) found that depleting responders through an e-crossing task makes them more likely to accept low offers in the first round of a sequence of ultimatum games with random re-matching after every round. In a subsequent work, the same authors found no effect of depleting responders through an e-crossing task (Achtziger, Alós-Ferrer & Wagner, 2018). Finally, Clark and Dickinson (2017) found that sleep deprived participants do not punish more than well-rested participants after a public goods game.

There have been also numerous papers exploring the effect of neurostimulation of the DLPFC on ultimatum game rejections. Knoch et al. (2006) found that disruption of the rDLPFC (but not of the lDLPFC) by transcranial magnetic stimulation reduces rejection rates in the ultimatum game. Interestingly, in their experiment, responders still judged low offers to be

unfair, but they were unable to reject them. Knoch et al. (2008) found a similar result by disrupting the rDLPFC using cathodal transcranial direct current stimulation, compared to sham. They concluded that the rDLPFC plays a decisive role to overcome the prepotent, selfish, impulse of accepting any offer. A similar conclusion was obtained by Baumgartner, Knoch, Hotz, Eisenegger and Fehr (2011), who found that disrupting the rDLPFC via transcranial magnetic stimulation make responders more likely to accept low offers, compared to disruption of the IDLPFC.

Finally, only one study used the two-response paradigm. Bago, Bonnefon and De Neys (2019) found that very few subjects change their initial strategy, made under time pressure, even when they are asked to think for at least 10 seconds before deciding whether to switch strategy. As a consequence, most of the rejection choices made under intuition, remained rejections also under deliberation.

As for third-party punishment, Wang et al. (2011) implemented a trust game followed by third party punishment. They found that a 150 seconds time delay decreases punishment, compared to a 30 seconds time pressure. Consistent with this finding, Liu et al. (2015) implemented an ultimatum game with a third party whose role is to punish proposers. In doing so, they found that depleted third-parties are more willing to punish unfair offers. In contrast with these findings, however, Buckholtz et al. (2015) found that disrupting the DLPFC via transcranial magnetic stimulation has, compared to sham, the effect of decreasing the rate of punishment of norm-transgressors in hypothetical scenarios. Yudkin, Rothmund, Twardawski, Thalla and Van Bavel (2016) found that cognitively loaded participants punish out-group members more severely than in-group members. Artavia-Mora, Bedi and Rieger (2017) found that time delay, compared to time pressure, does not affect third-party punishment in a field

experiment in which participants had the possibility of withholding help (by not returning a lost glove) from someone who had litter.

Finally, there is also one paper exploring the cognitive basis of anti-social punishment. Pfattheicher, Keller and Knezevic (2017) found that conceptual priming of intuition, compared to baseline, increases antisocial punishment in a series of one-shot public goods games followed by a punishment stage. In a subsequent study published in the same paper, Pfattheicher and colleagues also found that inhibiting intuition, compared to baseline, decreases antisocial punishment. The same results were not found when activating or inhibiting deliberation. These results provide a first piece of evidence suggesting that antisocial punishment is driven by the Type 1 processing. See Table 4 for a short summary of each study.

Review of the empirical evidence on positive reciprocity

The cognitive basis of positive reciprocity has been explored much less than the cognitive basis of negative reciprocity.

Two studies used time constraints. Neo et al. (2013) found no effect of a 15-minute time delay (compared to baseline) on the amount returned in the trust game. However, Cabrales et al. (2017) found that time delay increases return rates in the same game.

Only one work using conceptual primes. Urbig et al. (2016) found that foreign language (Dutch students taking the experiment in English) make people more likely to not reciprocate others' contribution in a sequential public goods game, suggesting that deliberation impairs positive reciprocity.

Three experiments used ego-depletion. Halali et al. (2014) found that depleted investees return more than non-depleted ones. The authors obtained this result with two different ego-depletion manipulations, a Stroop task and an *e*-hunting task. In contrast to this view, Dickinson

and McElroy (2017) found that sleep restriction decreases trustworthiness in the trust game. Yet, Rantapuska et al. (2017) found that hunger does not affect trustworthiness.

Finally, one paper used neurostimulation. Knoch, Schneider, Schunk, Hohmann and Fehr (2009) found that disrupting the rLPFC via transcranial magnetic stimulation has the effect of decreasing the amount returned in the trust game, especially when future investors can observe investees' choices. The authors interpreted this finding as a piece of evidence that, in order to return money in the trust game, investees have to overcome a selfish impulse to keep the whole amount. See Table 5 for a short summary of each study.

Review of the empirical evidence on expectations of negative reciprocity

The cognitive basis of the expectations of negative reciprocity has also been explored much less than the cognitive basis of negative reciprocity.

Only one study used time constraints. Cappelletti, Güth and Ploner (2011) found that a 15-second time pressure increases proposers' offers, compared to a 180-second time window.

Five studies used ego depletion. Halali et al. (2013) found that ego-depletion increases the proportion of fair ultimatum offers. They also showed that this effect is driven by strategic considerations, as the same increase was not observed in dictator game giving. Similarly, Morewedge, Krishnamurti and Ariely (2014) found that alcohol intoxication does not affect proposers' offers. Along these lines, Achtziger, Alós-Ferrer and Wagner (2016) found that depleted proposers offer more than non-depleted ones, and that this effect is driven by fear to be rejected. Clark and Dickinson (2017) found weak evidence that sleep deprived participants contribute to the public good more than well-rested participants, when they know that the contribution phase will be followed by a punishment phase. Finally, Achtziger, Alós-Ferrer and Wagner (2018) found that depleted proposers are more selfish than non-depleted ones.

Finally, two studies used neurostimulation. Ruff et al. (2013) used anodal and cathodal transcranial direct current stimulation to either promote or impair the rLPFC during a voluntary donation in which the opponent can punish the donor (dictator game with a punishment option). In doing so, they found that increasing the activity in the rLPFC increases donations, while decreasing the activity in the rLPFC decreases them. This result was replicated by Strang et al. (2015). See Table 6 for a short summary of each study.

Review of the empirical evidence on the expectations of positive reciprocity

Two studies used time constraints. Neo et al. (2013) found no effect of a 15-minute time delay on the amount sent by the investor in the trust game. Jaeger et al. (2019) found that a 5 seconds time pressure, compared to a 10 seconds time delay, increases investments in the trust game.

Five studies explored the effect of ego depletion. Anderson and Dickinson (2010) found that one night of sleep deprivation reduces trust. This result was also replicated in a subsequent work (Dickinson & McElroy, 2017). Consistent with this view, Ainsworth, Baumeister, Ariely and Vohs (2014) found that ego-depletion decreases trust. However, Evans, Dillon, Goldin and Krueger found that ego-depletion has no overall effect on trust, but it makes people more likely to follow the default choice, whichever that is. A lack of an overall effect of ego-depletion on trust has been suggested also by Rantapuska et al. (2017), who found that hunger does not affect trust.

Two studies used cognitive load. Samson and Kostyszyn (2015) found that cognitive load (implemented in two different ways: through a disturbing noise and a memorisation task) decreases trust. In contrast to this result, Bonnefon, Hopfensitz and De Neys (2013) found that

cognitive load does not interact with capacity to detect trustworthiness, suggesting that trust is, in fact, intuitive.

Finally, one study used conceptual primes. Urbig et al. (2016) conducted a sequential 3-player public goods game where two players independently act as first movers and the third player can condition their choice on the first two contributions. Therefore, first mover choices can be considered as a measure of trust. In doing so, the authors found no effect on trust of second language (English for Dutch participants) compared to first language (Dutch for Dutch participants). See Table 7 for a short summary of each study.

Outlook and open problems

As reviewed above, the evidence regarding the effect of cognitive process on negative reciprocity is quite clear as most studies report a positive effect of Type 1 processing on both second- and third-party punishment. However, one notices that all three neurostimulation studies point towards the opposite result, suggesting that the DLPFC is necessary for the implementation of punishment. This suggests that neurostimulation of the DLPFC works fundamentally differently than the other cognitive manipulations, at least in the case of negative reciprocity. Were this be confirmed, this would be an indication that neurostimulation of the DLPFC does something different from simply manipulating participants' cognitive processing. I will elaborate more on this in the final section.

Open problem. Gain a better understanding of the effect of neurostimulation of the DLPFC on punishment, by conducting more studies and eventually meta-regressing the studies over the cognitive manipulation, to understand whether neurostimulation works fundamentally differently from the other manipulations.

Regarding positive reciprocity, as reviewed above, the results are very split. This might provide a first piece of evidence that there is no overall effect. However, since there are only seven studies on the topic, more evidence and a formal meta-analysis are needed to draw general conclusions.

Open problem. To conduct more studies on the cognitive basis of positive reciprocity.

Studies exploring the cognitive basis of expectations of negative reciprocity appear to depend on the cognitive manipulation in a similar fashion as those exploring the cognitive basis of negative reciprocity. Specifically, time pressure and ego-depletion studies tend to point towards a positive role of intuition, such that intuition makes people more pro-social when they know that they might be punished. This is the mirror image of the empirical regularity mentioned above that intuition promotes negative reciprocity in virtually all studies using time pressure and ego depletion. However, in the domain of neurostimulation, as in the case of negative reciprocity, the effect seems to flip, as intuition seems to make people less pro-social. This potential reversal of the causal effect between cognitive manipulation and punishment depending on the cognitive process manipulation suggests that neurostimulation on one hand and time constraints and ego depletion on the other hand work fundamentally differently at least in the case of second- and third-party punishment. As I mentioned before, to confirm this with more studies and formal meta-analyses is a key direction for future research, which can have an important theoretical impact.

Open problem. Conducting more studies and/or to meta-regressing the current studies in order to explore the moderating role of cognitive manipulation on the effect of intuition and deliberation on expectations of negative reciprocity.

Another interesting question regards the potential moderator role of gender in determining the cognitive basis of (expectations of) positive reciprocity. Indeed, while previous research find no evidence of gender differences in negative reciprocity (Solnik, 2001; Croson & Gneezy, 2009), the evidence regarding positive reciprocity and, in particular, trust game behaviour, seems to converge on the finding that men trust more than women, although they are equally trustworthy (Buchan, Croson & Solnick, 2008; Dittrich, 2015). This might suggest that there is an interaction between gender and cognitive process, at least in the case of expectations of positive reciprocity.

Open problem. Exploring the interaction between gender and cognitive manipulation, at least in the case of expectations of positive reciprocity.

The cognitive basis of act utilitarianism

Many choices have a moral component. How to guide decisions in these situations have intrigued philosophers and psychologists for centuries, leading to two main traditions: consequentialism and deontology. Consequentialist judgments are those that are in line with moral theories according to which the rightness or wrongness of an action depends only on its consequences. A popular consequentialist theory is act utilitarianism, according to which "actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (Mill, 1863; Bentham 1789/1983). Act utilitarianism can be logically rejected in many ways. The main non-utilitarian theory discussed in contemporary research is "deontology". Broadly speaking, a deontological moral theory states that the rightness or wrongness of an action depends on whether it fulfils certain moral norms, rules, or duties, independently of the consequences that this action brings about (e.g., Kant, 1797/2002).

Numerous works in the last two decades have explored the cognitive underpinnings of deontology and act utilitarianism.

Measures of the conflict between deontology and act utilitarianism

Sacrificial moral dilemmas

Classically, the conflict between deontology and act utilitarianism has been measured using sacrificial moral dilemmas, hypothetical scenarios in which participants have to report whether it is morally acceptable to sacrifice one person to save a greater number of people. One example, considered, among others, in Greene et al.'s (2001) pioneering work is the "footbridge dilemma". This dilemma is typically presented as follows: "A runaway trolley is headed for five people who will be killed if it proceeds on its present course. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and five people. The only way to save the five people is to push this stranger off the bridge, onto the tracks below. The stranger will die if you do this, but his body will stop the trolley from reaching the others. Should you save the five others by pushing the large man off the footbridge?" A participant who responds that it is morally acceptable to sacrifice the large stranger is said to be making an utilitarian judgment, because their judgement appears to be in line with act utilitarianism; a participant who responds that it is not acceptable to sacrifice the large stranger is said to be making a deontological judgement, because their judgment appears to reject the utilitarian duty to act so as to maximize the total welfare.

The process-dissociation approach

More recently, Conway and Gawronski (2013) proposed a different approach, based on Jacoby's (1991) process dissociation procedure. Conway and Gawronski's starting observation is that classical sacrificial dilemmas treat deontology and act utilitarianism as if they were inversely

related. For example, in the footbridge dilemma above, selecting the deontological option coincides with rejecting the utilitarian option; conversely, selecting the utilitarian option coincides with rejecting the deontological option. This appears to be inconsistent with the literature showing that participants experience high conflict when facing sacrificial dilemmas (Greene et al, 2001; Greene, 2007; Koenigs et al, 2007), suggesting that deontological and utilitarian inclinations are not inversely related (otherwise they would not be in conflict), but they are operating at the same time. This has important theoretical implications: suppose, for example, that one finds that time pressure makes people less likely to sacrifice the large man, then it is impossible to say whether time pressure has increased deontological inclinations, or it has decreased utilitarian inclinations.

Conway and Gawronski (2013) suggested one way to overcome this limitation. The key idea of their procedure is to compare answers to *incongruent dilemmas* (i.e., moral dilemmas in which the deontological and the utilitarian responses are misaligned) with answers to *congruent dilemmas* (i.e., moral dilemmas in which the deontological and the utilitarian responses are aligned). An example of congruent dilemma is torturing a person to prevent a *paint* bomb from exploding. An example of incongruent dilemma is torturing a person to prevent a *pipe* bomb from exploding. By comparing answers to congruent versus incongruent dilemmas, the relative influence of deontological and utilitarian inclinations can be computed algebraically. In doing so, Conway and Gawronski (2013) found that deontology and (act) utilitarianism are indeed dissociable: while both inclinations were correlated to moral identity, deontological inclinations were associated to empathic concern, perspective-taking, and religiosity, whereas utilitarian inclinations were associated to need for cognition.

A two-dimensional model of act utilitarianism

Another approach has been recently proposed by Guy Kahane and colleagues (Kahane & Shackel, 2010; Kahane, 2015; Kahane et al, 2015; Kahane et al, 2018). In this line of work, Kahane and colleagues contend that sacrificial dilemmas are limited in that they measure only one dimension of utilitarianism and, arguably, the least important one. Their claim originates from Peter Singer's (1979) observation that the core dimension of act utilitarianism is commitment to *impartial beneficence*, that is, the moral requirement that one must promote the greater good of all human beings in a radically impartial way, without regard to physical, emotional, or relational distance between the actor and the beneficiary. Following this idea, Kahane et al. (2018) introduced a two-dimensional model of utilitarian psychology. In this model, the two dimensions correspond to instrumental harm (IH) and impartial beneficence (IB). To measure people's position in this two-dimensional space, the authors introduced, refined, and validated a novel scale, the Oxford Utilitarianism Scale (OUS). After refinement, this scale consists of nine items, five in the IH dimension and four in the IB dimension. Kahane et al. (2018) used this scale to show that the IH and the IB dimensions are psychologically independent: empathic concern, identification with all humanity, and concern for future generations were found to be positively correlated with IB, but negatively correlated with IH. These correlations also imply that the two dimensions proposed by Kahane et al. (2018) do not map exactly onto those proposed by Conway and Gawronski (2013).¹⁸

The nine OUS items are reported below. Answers are collected using a 7-item Likers scales from "strongly disagree" to "strongly agree".

OUS_IB1. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.

OUS_IH1. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.

OUS_IB2. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.

OUS_IH2. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.

OUS_IB3. From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favour the well-being of people who are especially close to them either physically or emotionally.

OUS_IH3. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.

OUS_IB4. It is just as wrong to fail to help someone as it is to actively harm them yourself.

OUS_IH4. Sometimes it is morally necessary for innocent people to die as collateral damage—if more people are saved overall.

OUS_IB5. It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal

Review of the empirical evidence on the cognitive basis of deontology and utilitarianism

As mentioned earlier, scholars have proposed three measures of deontological and utilitarian choices. In this review, I will follow a chronological order. First, I will review the studies that used sacrificial dilemmas (see also Isler and Yilmaz, 2019, for another review); then

I will review those using the process dissociation approach; finally, I will review those using the OUS.

The work using sacrificial dilemmas provides converging evidence that intuition favours deontological judgments. Valdesolo and DeSteno (2006) found that a positive emotion induction makes participants more likely than control participants to sacrifice the large man in the footbridge dilemma. In line with this research, Greene, Morelli, Lowenberg, Nystrom and Cohen (2008) found that cognitive load selectively interferes with utilitarian judgments, by making them slower. Suter and Hertwig (2011) found that an 8-second time pressure, compared to a 3minute time delay, increases deontological judgments in sacrificial moral dilemmas. Additionally, Suter and Hertwig (2011) replicated their result using, instead of time pressure, conceptual primes that instruct participants to answer swiftly vs. to deliberate thoroughly. Cummins and Cummins (2012) found that time pressure reduces the proportion of utilitarian judgments in eighteen sacrificial moral dilemmas. Paxton, Ungar and Greene (2012) found that subjects taking the cognitive reflection test prior to making a sacrificial moral dilemma judgment tend to become more utilitarian, arguably because the CRT primes them to reflect about their choice. Trémolière, De Neys and Bonnefon (2012) found that ego-depletion (implemented by mortality salience) increases deontological judgments in sacrificial dilemmas. Kvaran, Nichols and Sanfey (2013) found that participants primed to think analytically tend to make more utilitarian choices than control participants, whereas emotional primes tend to make participants less utilitarian. Trémolière and Bonnefon (2014) found that time pressure increases deontological choices in sacrificial moral dilemmas, but only when the save/kill ratio is relatively small (five); by contrast, the authors found no effect when the save/kill ratio is large (five hundred). They found the same result with cognitive load at the place of time pressure. Costa et al. (2014) found

that second language increases utilitarian judgments. A similar result has been obtained by Cipolletti et al. (2016), Corey et al. (2017), and Geipel et al. (2015a). Baron and Gürçay (2017) meta-analysed all response time studies using sacrificial dilemmas and found that, after controlling for subjects' ability (probability to provide the utilitarian response) and dilemma difficulty (probability of eliciting the deontological response), there is no relation between response time and moral judgments; they concluded that there is no evidence for a sequential two-system model of moral judgment. Spears, Fernández-Linsenbarth, Okan, Ruz and González (2018) manipulated participants cognitive processes presenting dilemmas using an easy-to-read versus hard-to-read font. In doing so, they found that participants in the hard-to-read-font condition are more likely to endorse utilitarian judgments. Timmons and Byrne (2018) used an ego-depletion task similar to the e-crossing task and found that depleted participants tend to be more deontological than non-depleted ones. Bago and De Neys (2018) implemented a tworesponse paradigm and, using several sacrificial dilemmas, found that about 70% of the utilitarian deliberative choices were already utilitarian under intuition. Moreover, they found that switches from deontological to utilitarian choices are only slightly more frequent than switches from utilitarian to deontological choices.

Therefore, research consistently showed that Type 1 processing is associated to deontological judgments in sacrificial moral dilemmas. Is this association driven by a reduction of utilitarian inclinations or by a strengthen of deontological intuitions? Four papers attempted to answer this question using the process dissociation approach. Conway and Gawronski (2013) found that cognitive load selectively reduces utilitarian judgments, while leaving deontological inclinations unaffected. A similar result has been obtained by Li, Xia, Wu and Chen (2018), who found that a combination of conceptual priming of intuition and cognitive load decreases

utilitarian inclinations, while leaving deontological inclinations unaffected. Therefore, these papers suggest that promoting Type 1 processing reduces utilitarian judgments. Two more papers explored the effect of impairing Type 1 processing by having participants making moral judgments in a second language. Hayakawa, Tannenbaum, Costa, Corey and Keysar (2017) reported six experiments showing that second language reduces deontological judgments, but it does not increase utilitarian judgments. Moreover, in three out of this six experiments, second language actually *decreased* utilitarian judgments. This somewhat surprising result was subsequently replicated by Muda, Niszczota, Białek and Conway (2018), who found that second language reduces both deontological and utilitarian judgments. The authors interpreted this finding suggesting that second language does not simply reduce Type 1 processing, but it more generally reduces harm aversion both towards the person to be sacrificed and the people to be saved.¹⁹

Only one study explored the cognitive underpinnings of moral judgments using the Oxford Utilitarianism Scale. Capraro, Everett and Earp (2019) found that priming intuition increases non-utilitarian judgments in the instrumental harm dimension, but not in the impartial beneficence dimension. This suggests that the two dimensions of act utilitarianism, instrumental harm and impartial beneficence, are cognitively distinct. See Table 8 for a short summary of each study.

Outlook and open problems

As just reviewed, there is large and consistent evidence that Type 1 processing is associated with deontological judgments in sacrificial moral dilemmas. Two additional studies suggest that this effect is driven by a reduction of utilitarian inclinations, rather than a strengthen of deontological ones. On the other hand, only one study explored the effect of cognitive

manipulations on moral judgments using the OUS: Capraro et al. (2019) found that conceptual priming of intuition favours non-utilitarian judgments in the domain of instrumental harm but not in the domain of impartial beneficence. This is result has important theoretical implications, because it suggests that the two dimensions of utilitarianism are cognitively distinct. However, one limitation of this work is that it uses explicit conceptual primes of intuition and deliberation, which have been criticized as they can generate potential demand effects, such that participants respond according on how they believe they should respond (Rand, 2016; Kvarven et al., 2019). Therefore, it would be important to conceptually replicate this finding using cognitive manipulation techniques other than explicit conceptual primes.

Open problem. Exploring the cognitive underpinnings of moral judgments in the OUS dilemmas using cognitive manipulations other than explicit conceptual primes.

A major open question regards whether real moral dilemmas have the same cognitive basis of hypothetical moral dilemmas. Conducting experiments using real dilemmas is difficult for obvious ethical reasons. However, very recently, Bostyn, Sevenhant and Roets (2018) explored decisions in real-life moral dilemmas using electroshocks (which participants did not know they were bogus) to be administered to one vs five mice. Surprisingly, Bostyn et al. (2018) found that moral judgments in classical hypothetical moral dilemmas are not predictive of behaviour in real-life dilemmas, but "they are predictive of affective and cognitive aspects of the real-life decision". Although Bostyn et al.'s (2018) results have been criticized (Białek, Turpin & Fugelsang, 2019; Colman, Gold & Pulford, 2019; Evans & Brandt, 2019; Plunkett & Greene, 2019), they anyway raise a major question. What are the cognitive underpinnings of deontology and act utilitarianism in real-life moral dilemmas?

Open problem. Exploring the underpinnings of deontology and act utilitarianism in reallife moral dilemmas.

Another interesting question regards potential gender differences in the cognitive basis of moral judgments. There is indeed evidence that women differ from men in moral judgments in the instrumental harm dimension (Capraro & Sippel, 2017; Friesdorf, Conway & Gawronski, 2015; Fumagalli et al. 2010). Therefore, it would be interesting to see whether there are gender differences in the cognitive underpinnings of utilitarianism.

Open problem: Exploring potential gender differences in the cognitive basis of deontology and act utilitarianism.

The General Social Heuristics Hypothesis

The Social Heuristics Hypothesis (SHH) has been originally introduced by Rand and colleagues (Rand et al. 2014), in the domain of cooperative behaviour, to explain their finding that intuition promotes cooperation in one-shot, anonymous, social dilemmas. Using their original words, the SHH contends that "cooperation is typically advantageous in everyday life, leading to the formation of generalized cooperative intuitions. Deliberation, by contrast, adjusts behaviour towards the optimum for a given situation. Thus, in one-shot anonymous interactions where selfishness is optimal, intuitive responses tend to be more cooperative than deliberative responses". More recently, the SHH has been adapted to other behavioural domains, such as altruism (Rand et al. 2016), honesty (Capraro, 2017), and reciprocity (Hallsson et al. 2018).

In spite of these applications, there is at the moment no formal model of the SHH and its generalizations. Some scholars introduced specific models to study the evolution of cooperation among dual-process individuals (Bear & Rand, 2016; Bear, Kagan & Rand, 2017, Jagau & Van Veelen, 2018; Mosleh & Rand, 2018). However, being focused on cooperative behaviour, these

models, although interesting, are of limited applicability. Therefore, the development of a general framework would be very desirable at this stage, especially in light of the recent observation that the lack of a unified theory limits the capacity of a theory to make explicit predictions, which, in turn, limits its falsifiability, slows down new developments, and might even result in a replicability crisis (Muthukrishna & Henrich, 2019).

The goal of this section is to develop a formal, general framework – inspired to the original SHH – that can be applied to potentially every one-shot, anonymous game.

Theoretical framework

Let G be a one-shot, anonymous game, with $N \ge I$ anonymous players. As standard, here I call "games" also decision problems with only one decision maker, such as the dictator game. Individuals that are affected by the decisions but do not make any actual decisions (e.g., recipients in the dictator game) are counted among the N players. G needs not be simultaneous.

Descriptively, the key assumptions of the General Social Heuristics Hypothesis (GSHH) are:

GSHH1: Type 1 processing allows people to quickly access the heuristics that they have internalized in those everyday interactions that resemble game G.

GSHH2: Type 2 processing allows people to compute the strategies that maximize their payoff in the given, one-shot, anonymous, interaction, and implement one of them, by overriding the heuristic suggested by Type 1.

To formalize GSHH2 is easy: Type 2 processing pushes people towards a Nash equilibrium of the game G. To formalize GSHH1 is less easy, as one needs to give mathematical sense to the informal notion of "everyday interactions that resemble game G".

Everyday interactions differ from one-shot, anonymous, laboratory interactions in several regards. Here, I focus on two points:

- (i) Payoff structure. While, in the lab, payoffs are set by the experimenter, in everyday interactions they depend on a number of factors that are beyond the control of the experimenter. Therefore, it is well possible that, when a participant reads the instructions of the experimental game G, she or he automatically thinks about another game G' that is, from a strategic point of view, similar to the original game G, but it differs from it from the point of view of the particular payoffs.
- (ii) *Dynamicity*. One-shot laboratory experiments are, by construction, one-shot and anonymous. Real interactions are different: in reality, after one interaction, people can usually decide whether to interact again, or not.

To formalize point (i) into a theoretical framework, I use the notion of weak embedding of a game into another game. This concept is closely related to the notion of weak isomorphism between games, introduced to describe games that are strategically equivalent, but differ from one another for the specific payoffs (Osborne & Rubinstein, 1994; Gabarró, García & Serna, 2011).

In what follows, G will typically represent the experimental game as set up by the experimenter, while G' will typically represent the game to which an experimental subject automatically thinks about when reading the instructions of G. Note that, while G, obviously, does not depend on the experimental subject, G' will generally do.

Let G, G be two games, with the same number of players and the same strategy sets. Let $p_{i,G}(s)$ be the material payoff of player i in game G when the strategy profile s is played. I say

that G is weakly embeddable into G' if the following property is satisfied: whenever s and t are two strategy profiles such that $p_{i,G}(s) > p_{i,G}(t)$, then one has $p_{i,G'}(s) > p_{i,G'}(t)$.

Therefore, weak embeddings preserve the order of payoffs, but they do not preserve equivalence of payoffs: if $p_{i,G}(s) = p_{i,G}(t)$, it might happen that $p_{i,G'}(s) \neq p_{i,G'}(t)$.

Weak embeddings are useful because they allow to take into account two potential sources of mismatch between game G (the experimental game as set up by the experimenter) and game G' (the game to which an experimental subject automatically thinks about when reading the instructions of G).

As for the first source of mismatch, assume that the experimenter wants to test the behaviour of participants in a prisoner's dilemma with benefit-to-cost ratio b/c=2. Assume that, in the experimental sample, there are subjects who were raised in a very cooperative society in which cooperation typically has a large benefit and a small cost and, consequently, when these participants read the instructions of the prisoner's dilemma, their first reaction is to think about a prisoner's dilemma with a greater benefit-to-cost ratio, say, b/c=10. Weak embeddings allow to take this into account, because all prisoner's dilemmas can be weakly embedded into each other.

The second source of mismatch between G and G' is that G' might contain "secondary dimensions" that are important for the experimental subject when making a decision, but they were not explicitly included in G. As an explicit example, consider the case of instrumental harm in sacrificial dilemmas. The typical lab experiment assumes that the consequences for the decision maker when s/he decides to push the large man off the bridge are equal to the consequences when s/he decides not to push the large man off the bridge. Therefore, in G, the payoff function of the decision maker does not distinguish "pushing the large man off the bridge" from "not pushing the large man off the bridge". However, in reality (i.e., in G') there are several

"secondary dimensions", that are not represented in G, that might imply that the consequences of pushing the large man off the bridge are not the same as the consequences of not pushing him. For example, the decision maker might think that, in reality, if s/he tries to harm another person, then the person target of the harm would react, leading to a fight that is likely to harm also the decision maker. Another dimension regards a potential reputation cost: real actions are typically observed by others, who might judge the decision maker badly if observed harming other people. Weak embeddings are useful to take into account the presence of potential "secondary dimensions": two strategy profiles s and t that are distinguishable in G, are distinguishable also in G; however, two strategy profiles that are not distinguishable in G, might become distinguishable in G, thanks to these "secondary dimensions".

In sum, weak embeddings allow to take into account the existence of (at least) two potential sources of mismatch between G and G'. It is important to note that, while this gives elasticity to the theory, the fact that weak embeddings preserve the order of payoffs guarantee that, in many cases, the predictions of the theory do not depend on G'.

I can now formalize GSHH2. I assume that, under Type 1, players do not play G, but they play its "real-life analogue", G^{real} , defined as follows:

- In the first round, players play a game G' such that G weakly embeds into G'
- At the end of the first round, all decision-makers vote whether to play another round of
 G'
- If all the decision-makers vote for playing G' again, then they play G', and the game starts again. Otherwise the game stops.²²

Two technical notes are needed at this stage. First, note that G^{real} might be an infinite game, if players keep voting to re-play G' after every round. Therefore, to guarantee

convergence of payoffs in case the game turns out to be infinite, I define the payoff of player i at round t to be the payoff of the game, discounted by θ^t . Second, note that, in the definition, I did not require that all players vote whether to play another round, but I required that all decision-makers vote to play another round. Therefore, for example, in the dictator game, only the dictator votes whether to play another round.

Having said this, I can finally state the GSHH.

The General Social Heuristics Hypothesis (GSHH) contends that Type 1 processing promotes a strategy that maximizes the payoff of G^{real} , while Type 2 processing promotes a strategy that maximizes the payoff of G.

Testing the GSHH on social decisions

In this section, I extrapolate the predictions of the GSHH in all the social decisions considered in this review, and I compare them with the empirical findings discussed above.

Cooperation

Assume that G is a one-shot, anonymous, prisoner's dilemma (the public goods game is similar). Let G be a game such that G weakly embeds into G. It is easy to see that G is itself a prisoner's dilemma game, because weak embeddings preserve the order of payoffs. The only difference is that G needs not be a symmetric prisoner's dilemma. Now assume that G is such that the payoff for mutual cooperation is positive for both players. If so, one of the Nash equilibria of G^{real} is to cooperate at every round and then vote to play another round if and only if the other player cooperated in the round before. On the other hand, the only strategy that maximizes the payoff in G is defection, being G a one-shot prisoner's dilemma. Therefore, the GSHH predicts that the rate of cooperation under Type 1 is *greater than or equal to* the rate of

cooperation under Type 2. This overall prediction has been supported by Rand's (2016), Kvarven et al.'s (2019), and Rand's (2019) meta-analyses.²³

Note that the GSHH does not predict that the rate of cooperation under Type 1 is always strictly greater than the rate of cooperation under Type 2. Whether the inequality is, in general, strict is still highly debated (Kvarven et al. 2019; Rand, 2019). By contrast, the GSHH predicts that whether the inequality is strict depends on the particular sample. For example, people who live in a cooperative society (high cost to benefit ratio and positive payoffs for mutual cooperation) are more likely than people who live in a non-cooperative society to play the cooperative equilibrium of G^{real} , therefore they are more likely to have internalized the cooperative heuristic. This prediction has been supported by two studies (Rand & Kraft-Todd, 2014; Capraro & Cococcioni, 2015).

Altruism

Assume that G is a one-shot, anonymous, dictator game. Since weak embeddings preserve the ordering of the payoffs, then G' is strategically equivalent to a dictator game. The only differences between G and G' can be in the initial endowments of the two players and in the cost to benefit ratio of the altruistic action, which need not be I in G', because weak embeddings preserve only payoff orders but not payoff ratios. From this observation, it follows that the only strategy that maximizes the dictator's payoff in G^{real} is to be selfish, because her or his payoff, at every round, can only decrease. Since also the strategy that maximizes the dictator's payoff in G is to be selfish, in the case of the dictator game the GSHH predicts that Type 1 and Type 2 have no effect on altruistic behaviour. Two meta-analyses agree on this finding (Rand et al. 2016; Fromell et al. 2018).²⁴

Honesty

Since Köbis et al.'s (2019) meta-analysis suggests that the effect of Type 1 and Type 2 processing on honest behaviour might depend on whether lying harms concrete or abstract others, in the analysis below, I distinguish these two cases.

Consider a one-shot, anonymous sender-receiver game, G, where lying, comparing to telling the truth, harms the receiver and benefits the sender (black lie). I now show that, in this situation, there exist games G' such that: (i) G weakly embeds into G'; and (ii) being honest is an equilibrium of the corresponding G^{real} . To demonstrate this, I first note that G' is a senderreceiver-like game, which differs from G only on the specific payoff consequences of lying versus telling the truth, but not on the order of the payoffs. This means that, since G is a senderreceiver game in the black lie condition, so it is G'. Therefore, among the possible G' there are some in which the receiver's payoff when the sender lies is negative, whereas the sender's payoff when s/he tells the truth is positive. In this situation, if the sender lies at the first round of G^{real} , the receiver might prefer voting to leave the interaction, rather than voting to play another round, in order to avoid getting one more negative payoff in the next round; on the other hand, since the sender receives a positive payoff by telling the truth, s/he prefers telling the truth and then voting to play another interaction, rather than lying at the first interaction and then stop, because the receiver refuses to play another round. In sum, to be honest is among the equilibria of G^{real} . On the other hand, since G is a sender-receiver game in the black lie condition, the only Nash equilibrium of G is to lie. Therefore, the GSHH predicts that, when lying harms concrete others, the rate of honesty under Type 1 is greater than or equal to the rate of honesty under Type 2. Köbis et al.'s (2019) meta-analysis found that the rate of honesty under Type 1 is equal to the rate of honesty under Type 2. Therefore, the predictions are in line with the empirical data. Additionally, there are two studies exploring the case in which senders are not initially informed

about their payoff maximizing strategy (which arguably leaves more room to Type 1 processing), all of which found greater honesty under Type 1 (Capraro, 2017; Capraro et al. 2019). Therefore, the empirical results are in line with the GSHH predictions, especially when the experimental designs make space for Type 1 processing.

However, the same does not hold true for cheating tasks in which lying harms abstract others. The peculiarity of this task compared to the sender-receiver game is that there is no second player that can vote to leave the interaction. Therefore, both in G' and G' it is optimal to lie. Thus, the GSHH predicts that there will be no effect of Type 1 and Type 2. Köbis et al.'s (2019) meta-analysis shows that this prediction is not supported by the empirical data.

Negative reciprocity

Assume that G is a one-shot, anonymous, ultimatum game. Since weak embeddings preserve the ordering of the payoffs, then G' is strategically equivalent to an ultimatum game, apart from the specific payoffs of the two players. Considers G' such that the payoff associated to rejection is 0 for both participants (e.g., G' = G). In this case, G'^{real} admits a number of equilibria of the following form: the responder rejects any offer under a certain threshold and always vote to play another round of G' (even when s/he receives an offer smaller than its threshold); the proposer makes an offer equal to the responder's threshold and always vote to play another round of G'. On the other hand, in G, the responder's payoff-maximizing strategy is to accept any offer. Therefore, the GSHH predicts that the rate of rejections of low offers under Type 1 is greater than or equal to the rate of rejections of low offers under Type 2. As reviewed above, this is in line with the empirical data.

Now, consider the same game G, from the point of view of the proposer. A similar line of reasoning as above shows that, in G^{real} , making a positive offer and then voting to play another

round is certainly part of the Nash equilibrium in which the responder rejects any offer below that positive offer. In contrast, in G the proposer knows that the responder's payoff maximizing strategy is to accept any offer, therefore the proposer's payoff maximizing strategy is to offer the minimum non-zero offer. In other words, the GSHH predicts that offers in G under Type 1 are greater than or equal to the offers under Type 2. As reviewed above, this prediction is supported by the empirical data.

Positive reciprocity

Let G be a one-shot, anonymous, trust game. Since weak embeddings preserve the ordering of payoffs, then G' differs from G only on the particular payoffs received by the players, and not on the strategic structure. Consider G' such that all the payoffs are non-negative (e.g., G' = G). In this case, G^{real} admits the following equilibrium: the investor invests the money and then vote to play another round of G' if and only if the investee returns the money; the investee returns the money and always vote to play another round of G'. By contrast, in G, the only payoff maximizing strategy is to return no money. Therefore, the GSHH predicts that return rates in the trust game under Type 1 are higher than or equal to the return rates under Type 2. This prediction does not seem to be supported by the data. Seven experiments have been conducted so far, leading to inconclusive results: two studies suggest a positive effect of Type 1 on return rates, two studies suggest a null effect, and three studies suggest a negative effect.

Now, consider the same game G, from the point of view of the investor. A similar line of reasoning as above shows that, in G^{real} , investing is optimal, because investees have an incentive to return money and then vote to repeat the interaction. However, in G the investor knows that the investee's payoff maximizing strategy is to return no money, therefore the investor's payoff maximizing strategy is to invest no money. Therefore, the GSHH predicts that Type 1 promotes

trust, while Type 2 inhibits trust. This prediction is not supported by the data, which actually trend in the opposite direction. Indeed, as reviewed before, there are ten empirical studies exploring the cognitive underpinnings of expectations of positive reciprocity: six of them found no effect, four studies found that Type 1 processing disfavours prosocial behaviour, and one of them found that Type 1 processing favours prosocial behaviour.

Deontology vs. act utilitarianism

In this section, I would like to apply the GSHH to moral dilemmas. With some exceptions, the vast majority of empirical studies on moral dilemmas used hypothetical situations to test which of two conflicting actions people rate to be more morally appropriate. Therefore, moral dilemmas regard judgments rather than actual behaviour. Recent research highlights that judgments in moral dilemmas might not correspond to actual behaviour (Bostyn et al., 2018; FeldmannHall et al., 2012). Therefore, this section should be taken with a grain of salt. However, I think it is interesting to see that, under a "realistic assumption" the GSHH makes also testable predictions regarding moral dilemmas, and these predictions are roughly in line with the empirical observations.

In a typical sacrificial dilemma (a moral dilemma in the instrumental harm dimension, in the language of Kahane et al. 2018), people have to choose whether to sacrifice one person to save a greater number of people, typically five, from harm. One can approximate this situation by means of an economic decision problem G in which the decision maker has to decide between the allocation D = (0,0,-1,-1,-1,-1,-1) and the allocation U = (0,-1,0,0,0,0,0), where -1 represents "harm", 0 represents "not harm", the first component represents the payoff of the decision maker (who is assumed to be indifferent between "harming" and "not harming"), the second component represents the payoff of the person to be sacrificed, and the remaining components represent the

payoffs of the people to be saved from harm. The notation D (U) stands for Deontological (Utilitarian) choice.

Consider now G', which is, by definition, a game such that G weakly embeds into G', and it represents the game similar to G that the decision maker usually plays in real life. In reality, harming a person often comes with a cost, given simply by the fact that people, in general, do not like to be harmed; therefore, in reality, a person target of the harm is likely to try to defend her/himself, with the consequence of causing some harm to the decision-maker. Therefore, it seems "realistic" to assume that, in G', the decision-maker's payoff in the utilitarian scenario U' is negative and smaller that its payoff in the deontological scenario D'. From this assumption, it immediately follows, that, in G^{real} , the only strategy that maximizes the decision-maker's payoff is the deontological decision, D'. On the other hand, in G, given these payoffs, the decision maker is indifferent between the two choices. Therefore, under the assumption that, in G', harming someone is costly for the decision maker, then the GSHH predicts that the rate of deontological choices under Type 1 is greater than the rate of deontological choices under Type 2. This prediction is in line with the empirical research reviewed above.

I now move to the dimension of impartial beneficence. In this case, predictions seem less obvious. For example, consider the first item of the Oxford Utilitarianism Scale (OUS): "If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice." If we represent this situation with a decision problem \underline{G} in which the decision maker has to choose between two allocations of money, (0,-1) and (-1/2,0), it is clear that not sacrificing the leg is the only payoff maximizing choice both in G and in G' (because weak embeddings preserve orders). Therefore, in this case, the GSHH predicts that the rate of impartial beneficence under Type 1 should be the same as the rate of

impartial beneficence under Type 2 (and it should be zero). A similar argument applies to the second item of the OUS (from a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy), and also to the fifth item (it is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal). The predictions are less obvious in case of the third and the fourth items, that are more abstract (from a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favour the well-being of people who are especially close to them either physically or emotionally; it is just as wrong to fail to help someone as it is to actively harm them yourself).

In any case, it is interesting that, overall, with minimal and realistic assumptions on how to represent moral dilemmas using economic games and keeping the abstract items of the OUS aside, then the GSHH makes predictions that are in line with the empirical finding that intuition promotes deontological judgments in the IH dimension, but not in the IB dimension (Capraro, Everett & Earp, 2019). See Table 9.

Limitations of the GSHH

In sum, the GSHH successfully predicts the overall effects of Type 1 and Type 2 on decision-making in a number of behavioural contexts, but not all. Here I describe the main limitations of this framework.

The first main limitation regards its precise predictions in the domain of altruism and impartial beneficence. In both cases, the GSHH correctly predicts that there is no overall effect of Type 1 and Type 2 on behaviour. However, looking at the details of the GSHH predictions,

one sees that it predicts that the rate of altruism and impartial beneficence (at least in the OUS items 1, 3 and 5) should be zero. This is not what we see in the experiments. Decades of experimental research have shown that a significant proportion of people give part of their endowment in the one-shot, anonymous dictator game (Engel, 2011), and this proportion even increases when the benefit created is greater than the cost of the altruistic action (Andreoni & Vesterlund, 2001), a behaviour that could be classified as impartial beneficence.

A simple way to overcome this limitation is by assuming that some people are not completely selfish, but they think about the good of other people (altruism) or the good of the society as a whole (impartial beneficence). Formal theories of social preferences for equity and/or efficiency are very well studied in economics (Levine, 1998; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Engelmann & Strobel, 2004). Although these frameworks do not account for the differential effect of Type 1 and Type 2 processing, they can account for a baseline level of altruism and impartial beneficence, that is independent of the two cognitive processes. Similarly, assuming that some people are lying averse (Gneezy, Kajackaite & Sobel, 2018; Abeler, Nosenzo & Raymond, 2019), one can take into account a baseline level of honesty. Therefore, combining these theories with the GSHH, one can easily explain the baseline levels of altruism, impartial beneficence, and honesty. Yet, one cannot explain why Type 2 processing favours honesty in cases in which lying harms concrete others (Köbis et al. 2019). The inability of the GSHH to explain this result is, in my opinion, its greatest limitation. In principle, one might try to accommodate it by assuming that Type 2 processing favours the emergence of moral behaviour through cognitively costly, abstract moral principles. But, if so, why does Type 2 favours the emergence of honesty, but not the emergence of altruism

and cooperation? These behaviours too can be supported by moral principles. Explaining this asymmetry is, in my opinion, one of the most important directions for future research.

The second main limitation of the GSHH is that it predicts that Type 1 processing promotes trust, while the empirical results point towards the diametrically opposite effect. Is it possible to explain this discrepancy? Looking at the details of the review of the literature reported above, one notices that, among the four studies showing that Type 1 processing disfavours trust, three of them used ego-depletion as a mechanism to manipulate cognitive process. Previous work suggests that ego-depletion does not simply make people more likely to adopt Type 1, but it also makes them more likely to prefer immediate rewards over delayed rewards (Bayer & Osher, 2018). This generates a confound that is potentially problematic in the trust game, because, being it a non-simultaneous game, it does not only contain a conflict between prosociality and selfishness, but it also contains a conflict between short-term reward and long-term reward. To clarify this point, I think it is important that future studies explore the cognitive basis of trust using manipulations other than ego-depletion.

General open problems

In the previous sections, I have mentioned several open problems whose solution would further advance our understanding of the cognitive basis of a specific social decision. I would like to conclude this review by mentioning two open problems that do not regard one social decision in particular, but they rather regard this field of research in general.

The need of uniformizing measures and manipulations

The underlying assumption of this field of research is that, given a social decision, all cognitive process manipulations have qualitatively similar effects on this social decision. This assumption is, for example, implicitly included in all the meta-analyses that have been conducted

thus far, because they analyse all cognitive manipulations together²⁵ (Rand, 2016; Rand et al. 2016; Fromell et al. 2018; Köbis et al. 2019), often without controlling on the cognitive manipulation itself (although the two most recent meta-analyses have started including this control: Kvarven et al. 2019; Rand, 2019). An important question is, therefore: is this assumption satisfied? Are we sure that all cognitive manipulations have qualitatively similar effects? There is some evidence that time pressure, ego depletion, conceptual primes of intuition, and cognitive load have similar effects on cooperation (Kvarven et al. ²019; Rand 2019). However, both these meta-analyses agree that the effect of explicit conceptual primes of emotion have a stronger effect. Moreover, these meta-analyses do not include neurostimulation studies. Finally, these meta-analyses are limited to cooperative behaviour. In sum, whether all cognitive manipulations have qualitatively similar effects on *all* social decisions is an open problem.

The case of neurostimulation is particularly interesting. As reviewed in the previous sections, there are two behavioural domains (cooperation and negative reciprocity) where the effect of neurostimulation seems to be different from the effect of other cognitive manipulations. Three meta-analyses (Rand, 2016; Kvarven et al., 2019; Rand, 2019) agree that time pressure, ego depletion, conceptual primes of intuition, and cognitive load, all together, promote cooperative behaviour in social dilemmas; by contrast, Li et al. (2018) found that activation of the rDLPFC *increases* cooperation, compared to the sham condition, and that inhibition of the rDLPFC *decreases* cooperation, relative to the sham condition. More worrisomely, the majority of studies agree that promoting Type 1 has a positive effect on negative reciprocity and on expectations of negative reciprocity; by contrast, all six studies using neurostimulation found that inhibiting the DLPFC *decreases* negative reciprocity and expectations of negative reciprocity

(Knoch et al. 2006; Knoch et al. 2008; Baumgartner et al. 2011; Ruff et al. 2013; Buckholtz et al. 2015; Strang et al. 2015).

However, the effect of neurostimulation seems to be in line with the effect of the other cognitive manipulations in the other behavioural domains. In the case of altruism, two meta-analyses (Rand et al. 2016; Fromell et al. 2018) agree that there is no overall effect of time pressure, ego-depletion, cognitive load, and conceptual primes of intuition, on altruism; consistent with this view, the evidence reported in the three studies using neuro-stimulation is mixed (Ruff et al. 2013; Strang et al. 2015; Gross et al. 2018). In the case of honesty, Köbis et al.'s (2019) meta-analysis found that, when lying harms abstract others, then honesty is deliberative, which is the same result reported in the only study using neuro-stimulation (Maréchal et al., 2017). In the same meta-analysis, Köbis et al. (2019) found that, when lying harms concrete others, then there is no overall effect of Type 2 on honesty; somewhat consistent with this view, the only study using neurostimulation in the sender-receiver game founds that activation of the DLPFC increases honesty, but only for women (Gao et al. 2018).

Leaving the case of neurostimulation aside, there is, in my view, an even more general problem *within* cognitive manipulations. Consider the case of time pressure studies. Some scholars used a time pressure of five seconds, others a time pressure of ten seconds, yet others a time pressure of fifteen or even twenty seconds. Even assuming that time pressure indeed promotes Type 1 processing, it is hard to compare different studies using different time constraints. A similar methodological issue can be observed within all other cognitive manipulation techniques: subjects are cognitively loaded using different parallel tasks; they are depleted using different tasks, and so on. For example, how can we compare the effect of the *e*-hunting task with the effect of the Stroop task? It is like if different researchers are trying to

measure the length of the same stick, but one researcher is using meters, another one is using miles, another one is using yards, and another is using... kilograms. Uniformizing the units of measurement is the step number zero to build a truly scientific theory. Until the units of measurement will not be uniformized, the entire field of research will be a conceptual chaos. I definitely think that this is *the* primary direction for future research.

Extending the dual-process approach to other types of decisions

The type of decisions discussed in this review cover a large class of social decisions. Indeed, the reason why I chose to focus on these particular decisions is that they have been the most studied ones, with at least ten experimental papers each. However, they do not cover the totality of social decisions. In fact, there are other types of social decisions that deserve to be studied in future research.

One example is the equity-efficiency trade-off. The natural distribution of a resource is often unequal, and creating equity is often costly. This generates a fundamental conflict between equity and efficiency, that has been named as "the fundamental problem in distributive justice" (Hsu et al., 2008) and "the big-trade-off" (Okun, 2015). See also Kohlberg (1931), Rawls (1971), and Sen (2008). People who are in power of distributing resources often face this conflict. This leads to a conflict between equity and efficiency. For example, according to Okun's "leaky bucket" argument, taxation is an instance of this conflict, because it increases equity, but it is costly for the institution that implements the tax. Understanding the cognitive basis of the conflict between equity and efficiency is therefore a topic of primary importance. However, very little is known about it. In spite of the importance of this type of decision, I am aware of only one study exploring the cognitive basis of the equity-efficiency trade-off (Capraro, Corgnet, Espín & Hernán-González, 2017). This study found that time pressure promotes equity, while time delay

favours efficiency. This result is somewhat in line with previous correlational studies: Hsu et al (2008) used functional Magnetic Resonance Imaging (fMRI) to study how people allocate meals to children in an orphanage in Uganda. In doing so, they found that equity choices are associated with activation in the insula, an area of the brain that is implicated with aversive social interactions and norm violation (Rilling and Sanfey, 2011). Hsu et al. (2008) used this finding to suggest that "emotional responses related to norm violations may underlie individual differences in equity considerations". Corgnet, Espín and Hernán-González (2015) analysed US and Spain samples and found that low scores in the CRT are correlated with equity, while high scores in the CRT are correlated with "mild altruism", which is a form of efficiency, where people pay a small cost to generate a great group benefit. Conceptually related, Kessler et al. (2019) found that when dictator game giving increases social efficiency, time delay promotes giving. Similarly, when cooperation has high benefit-to-cost ratio, time delay favours cooperation. Therefore, this preliminary work seems to suggest that Type 1 processing promotes equity, while Type 2 processing favours efficiency. Future work should explore this hypothesis more in depth.

Another example is ingroup favouritism. Ingroup favouritism is considered to be one of the most fundamental bias among humans. People tend to favour people in their own group, even when groups are formed at random (Tajfel, 1970; Tajfel et al, 1971; Tajfel, 1980). Ingroup favouritism has been observed in many economic games, including public goods games (Krupp et al., 2008), dictator games (Whitt & Wilson, 2007), charitable donation games (Pavey et al., 2011), ultimatum games (Kubota et al., 2013; McLeish and Oxoby, 2011), prisoner's dilemmas (Ahmed, 2007), among many others (see Everett et al., 2015, for a review). In spite of the importance of this type of behaviour, only three studies explored the cognitive underpinnings of ingroup favouritism, and the results are quite mixed. Rand, Newman and Wurzbacher (2015)

found that time pressure makes people more cooperative both with in-group members and with out-group members. De Dreu, Dussel and Ten Velden (2015) found that cognitive load increases ingroup favouritism. Everett et al. (2017) found that time delay decreases cooperation towards out-groups. Further work should clarify the cognitive basis of in-group favouritism.

One more example is aggression. People sometimes are aggressive, especially when they feel they are under threat. Is aggression intuitive or requires deliberation? Researchers tend to distinguish between two types of aggression, one more deliberative and goal-oriented, and one that is more impulsive, reactive, defensive (Nelson & Trainor, 2007; De Dreu, Scholte, van Winden & Ridderinkhof, 2015). Recently, Everett, Ingbretsen, Cushman and Cikara (2018) introduced a novel economic game, that they named pre-emptive strike game, to study whether defensive aggression is intuitive or deliberative. They found that time pressure increases aggression, especially against outgroup members. Future work should explore the cognitive underpinnings of aggression more in depth.

Finally, there are several situations in which cooperation is not costly, but it is risky, in the sense that a cooperator receives the benefit of cooperation only if the other partner cooperates (Hardin, 1968; Skyrms, 2004). However, mutual cooperation is an equilibrium: none of the players has an incentive to change strategy, if the other player is cooperating. What are the cognitive basis of risky cooperation? Only one studied explore this question. Belloc, Bilancini, Boncinelli and D'Alessandro (2019) found that a 10-second time pressure increases cooperation. Future work should explore the robustness of this finding.

Conclusion

I reviewed the existing literature on the cognitive basis of cooperation, altruism, honesty, positive and negative reciprocity and their expectations, and moral judgments (deontology vs act

utilitarianism). For each of these domains, I listed a number of open problems that I believe to be key to further advance our understanding of the cognitive basis of human sociality. I concluded by making an attempt to introduce a theoretical framework to explain the main empirical regularities. The framework is promising, as it makes predictions that are generally in line with the empirical evidence, in all but two cases: honesty (but only when lying harms abstract others) and positive reciprocity. Future work should shed light on the areas in which empirical evidence is inconclusive and theoretical predictions are inaccurate. I did my best to keep the review self-contained, exhaustive and research-oriented. My hope is that it can contribute to bring further attention in an area of research that has emerged in the last decades as one of the most vibrant and exciting areas of research in social science.

References

- Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87, 1115-1153.
- Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2015). Money, depletion, and prosociality in the Dictator Game. *Journal of Neuroscience, Psychology, and Economics*, 8, 1-14.
- Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2016). The impact of self-control depletion on social preferences in the ultimatum game. *Journal of Economic Psychology*, *53*, 1-16.
- Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2018). Social preferences and self-control. *Journal of Behavioral and Experimental Economics*, 74, 161-166.
- Aguiar, F., Brañas-Garza, P., Cobo-Reyes, R., Jimenez, N., & Miller, L. M. (2009). Are women expected to be more generous? *Experimental Economics*, 12, 93-98.
- Ahmed, A. M. (2007). Group identity, social distance and intergroup bias. *Journal of Economic Psychology*, 28, 324-337.
- Ainsworth, S. E., Baumeister, R. F., Ariely, D., & Vohs, K. D. (2014). Ego depletion decreases trust in economic decision making. *Journal of Experimental Social Psychology*, *54*, 40-49.
- Almås, I., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science*, *328*, 1176-1178.
- Alós-Ferrer, C., & Garagnani, M. (2018). The cognitive foundations of cooperation. *Available at https://www.econ.uzh.ch/dam/jcr:2e76448c-37da-45fd-a189-fc7f85e4e74a/econwp303.pdf*
- Andersen, S., Gneezy, U., Kajackaite, A., & Marx, J. (2018). Allowing for reflection time does not change behavior in dictator and cheating games. *Journal of Economic Behavior and Organization*, 145, 24-33.
- Anderson, C., & Dickinson, D. L. (2010). Bargaining and trust: The effect of 36hr sleep deprivation on socially interactive decisions. *Journal of Sleep Research*, 19, 54-63.
- Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, 116, 293-312.

- Arndt, J., Greenberg, J., Solomon, S., Pyszczynski, T., & Simon, L. (1997). Suppression, accessibility of death-related thoughts, and cultural worldview defense: Exploring the psychodynamics of terror management. *Journal of Personality and Social Psychology*, 73, 5.
- Artavia-Mora, L., Bedi, A. S., & Rieger, M. (2017). Intuitive help and punishment in the field. *European Economic Review*, 92, 133-145.
- Artavia-Mora, L., Bedi, A. S., & Rieger, M. (2018). Help, prejudice and headscarvers. *Working Paper available at http://ftp.iza.org/dp11460.pdf*.
- Bago, B., Bonnefon, J. F., & De Neys, W. (2019). Deliberation is irrelevant to prosociality. *Manuscript* in preparation.
- Bago, B., & De Neys, W. (2018). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*. Online first: http://dx.doi.org/10.1037/xge0000533.
- Balafoutas, L., Kerschbamer, R., & Oexl, R. (2018). Distributional preferences and ego depletion. *Journal of Neuroscience, Psychology, and Economics*, 11, 147-165.
- Balafoutas, L., & Tarek, J. L. (2018). Impunity under pressure: On the role of emotions as a commitment device. *Economics Letters*, *168*, 112-114.
- Ball, L., Thomson, V., & Stupple, E. (2017). Conflict and dual process theory: The case of belief bias. InW. De Neys (Ed.), *Dual Process Theory 2.0*. Oxon, UK: Routledge.
- Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas:

 A meta-analysis. *Group Processes & Intergroup Relations*, 12, 533–547.
- Banker, S., Ainsworth, S. E., Baumeister, R. F., Ariely, D., & Vohs, K. D. (2017). The sticky anchor hypothesis: Ego depletion increases susceptibility to situational cues. *Journal of Behavioral Decision Making*, 30, 1027-1040.
- Banks, A. P., & Hope, C. (2014). Heuristic and analytic processes in reasoning: An event-related potential study of belief bias. *Psychophysiology*, *51*, 290-297.
- Barbey, A.K., Koenigs, M., & Grafman, J. (2013). Dorsolateral pre-frontal contributions to human

- working memory. Cortex, 49, 1195–205.
- Baron, J., Gürçay, B. (2017). A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Memory & Cognition*, 45, 566-575.
- Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory: *The American Economic Review*, 84, 391-395.
- Baumeister, R., F., Heatherton, T. (1996). Self-regulation failure: An overview. *Psychological Inquiry*, 7, 1-15.
- Baumeister, R. F., Heatherton, T., & Tice, D. M. (1994). Losing control: How and why people fail at self-regulation. San Diego, CA: Academic Press.
- Baumeister, R. F., & Tierney, J. (2011). Willpower: Rediscovering the greatest human strength. New York, NY: Penguin Press.
- Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, *14*, 1468e74.
- Bayer, Y. M., & Osher, Y. (2018). Time preference, executive functions, and ego-depletion: An exploratory study. *Journal of Neuroscience, Psychology, and Economics, 11*, 127-134.
- Bear, A., Kagan, A., & Rand, D. G. (2017). Co-evolution of cooperation and cognition: The impact of imperfect deliberation and context-sensitive intuition. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20162326.
- Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, 113, 936-941.
- Becker, E. (2007). The denial of death. Simon and Schuster.
- Belloc, M., Bilancini, E., Boncinelli, L., & D'Alessandro, S. (2019). Intuition and deliberation in the Stag Hunt game. *Scientific Reports*, *9*, 14833.
- Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is "behavioral"? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, 11, 1231-1255.
- Bentham, J. (1983). The collected works of Jeremy Bentham: Deontology, together with a table of the

- springs of action; and the article on utilitarianism. Oxford, England: Oxford University Press. (Original work published 1789).
- Bereby-Meyer, Y., Hayakawa, S., Shalvi, S., Corey, J. D., Costa, A., & Keysar, B. (in press) Honesty speaks a second language. *Topics in Cognitive Science*.
- Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. Current Opinion in Psychology, 6, 196-198.
- de Berker, A. O., Bikson, M., & Bestmann, S. (2013). Predicting the behavioral impact of transcranial direct current stimulation: Issues and limitations. *Frontiers in Human Neuroscience*, 7, 613.
- Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reaosners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, 12, 148-167.
- Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, 1383-1385.
- Bird, B. M., Geniole, S. N., Procyshyn, T. L., Ortiz, T. L., Carré, J. M., & Watson, N. V. (2019). Effect of exogenous testosterone on cooperation depends on personality and time pressure. *Neuropsychopharmacology*, *44*, 538.
- Biziou van Pol, L., Haenen, J., Novaro, A., Occhipinti-Liberman, A., & Capraro, V. (2015). Does telling white lies signal pro-social preferences? *Judgment and Decision Making*, 10, 538-548.
- Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90, 166-193.
- Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General*, 142, 143-150.
- Brañas-Garza, P., Capraro, V., & Rascón-Ramírez, E. (2018). Gender differences in altruism on Mechanical Turk: Expectations and actual behaviour. *Economics Letters*, 170, 19-23.
- Boschin, E. A., Piekema, C., & Buckley, M. J. (2015). Essential functions of primate frontopolar cortex in cognition. *Proceedings of the National Academy of Science*, 112, E1020-E1027.
- Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. Journal of

- Experimental Psychology: General, 146, e1-e7.
- Bostyn, D. H., Sevenhant, S., Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*, 1084-1093.
- Bouwmeester, S., Verkoeijen, P. J. L., Acze, B., Barbosa, F., Bègue, L., et al. (2017). Registered replication report: Rand, Greene & Nowak (2012). *Perspectives on Psychological Science*, 12, 527-542.
- Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.
- Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford, UK: Oxford University Press.
- Buchan, N. R., Croson, R. T. A., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior & Organization*, 68, 466-476.
- Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., Marois, R. (2015).
 From blame to punishment: Disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, 87, 1369-1380.
- Byrd, N., & Conway, P. (in press). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition*.
- Cabrales, A., Espín, A., Kujal, P., & Rassenti, S. (2017). Humans' (incorrect) distrust of reflective decisions. *Available at SSRN: https://ssrn.com/abstract=2938434*.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social* psychology, 42, 116.
- Camerer, C. F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637.
- Cantarero, K., & Van Tilburg, W. A. (2014). Too tired to taint the truth: Ego-depletion reduces other-benefiting dishonesty. *European Journal of Social Psychology*, 44, 743-747.

- Cappelen, A. W., Nielsen, U. H., Tungodden, B., Tyran, J. R., & Wengström, E. (2016). Fairness is intuitive. *Experimental Economics*, 19, 727-740.
- Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2013). When do we lie? *Journal of Economic Behavior and Organization*, 93, 258-265.
- Cappelletti, D., Güth, W., & Ploner, M. (2011) Being of Two Minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology*, *32*, 940-950.
- Capraro, V. (2013). A model of human cooperation in social dilemmas. PLoS ONE, 8, e72427.
- Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Economics Letters*, *158*, 54-57.
- Capraro, V. (2018). Gender differences in lying in sender-receiver games: A meta-analysis. *Judgment and Decision Making*, 13, 345-355.
- Capraro, V., & Cococcioni, G. (2015). Social setting, intuition and experience in laboratory experiments interact to shape cooperative decision-making. *Proceedings of the Royal Society: Biological Sciences*, 282, 20150237.
- Capraro, V., & Cococcioni, G. (2016). Rethinking spontaneous giving: Extreme time pressure and egodepletion favor self-regarding reactions. *Scientific Reports*, *6*, 27219.
- Capraro, V. (2019). Gender differences in the equity-efficiency trade-off. Available at SSRN.
- Capraro, V., Corgnet, B., Espín, A. M., & Hernán-González, R. (2017). Deliberation favours social efficiency by making people disregard their relative shares: Evidence from USA and India. *Royal Society Open Science*, 4, 160605.
- Capraro, V., Everett, J. A. C., & Earp, B. D. (in press). Priming intuition disfavors instrumental harm but not impartial beneficence. *Journal of Experimental Social Psychology*.
- Capraro, V., Jordan, J. J., & Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports*, *4*, 6790.
- Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity and efficiency per se, drive human prosociality. *Judgment and Decision*

- Making, 13, 99-111.
- 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.
- Capraro, V., & Sippel, J. (2017). Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive Processing*, 18, 399-405.
- Carlson, R. W., Aknin, L. B., & Liotti, M. (2016). When is giving an impulse? An ERP investigation of intuitive prosocial behavior. *Social Cognitive and Affective Neuroscience*, 11, 1121-1129.
- Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control dfoes not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, 144, 796-815.
- Carter., E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: Has the evidence for ego-depletion been overestimated. *Frontiers in Psychology*, *5*, 823.
- Carver, C. S., Johnson, S. L., Joormann, J. (2008). Serotonergic function, two-mode models of self-regulation, and vulnerability to depression: What depression has in common with impulsive aggression. *Psychological Bulletin*, 134, 912-943.
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39, 752-766.
- Chaiken S, & Trope, Y. (1999). Dual-Process Theories in Social Psychology. New York: Guilford.
- Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117, 817-869.
- Chen, F., & Fischbacher, U. (2019). Cognitive processes underlying distributional preferences: A response time study. *Experimental Economics*. Online first: https://doi.org/10.1007/s10683-019-09618-x
- Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*, 3557.
- Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y.

- Trope (Eds.), *Dual-process theories in social psychology* (pp. 73-96). New York, NY, US: Guilford Press.
- Chiou, W. B., Wu, W. H., & Cheng, W. (2017). Self-control and honesty depend on exposure to pictures of the opposite sex in men but not women. *Evolution and Human Behavior*, 38, 616-625.
- Chuan, A., Kessler, J. B., Milkman, K. L. (2018). Field study of charitable giving reveals that reciprocity decays over time. *Proceedings of the National Academy of Sciences USA*, 115, 1766-1771.
- Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, 2, 84-92.
- Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign-language effect. *Philosophical Psychology*, 29, 23–40.
- Clark, J., & Dickinson, D. L. (2017). The impact of sleep restriction on contributions and punishment: First evidence. *Available at https://www.econstor.eu/bitstream/10419/170807/1/dp10823.pdf*
- Colman, A. M., Gold, N., & Pulford, B. D. (2019). Comparing hypothetical and real-life trolley problems: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, 1386-1388.
- Cone, J., & Rand, D. G. (2014). Time pressure increases cooperation in competitively framed social dilemmas. *Plos One*, *9*, e115756.
- Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making:

 A process dissociation approach. *Journal of Personality and Social Psychology*, 104, 216-235.
- Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, 179, 241-265.
- Corey, J. D., Hayakawa, S., Foucart, A., Aparici, M., Botella, J., Costa, A., & Keysar, B. (2017). Our moral choices are foreign to us. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1109-1128.
- Corgnet, B., Espín, A. M., Hernán-González, R. (2015). The cognitive basis of social behavior: Cognitive

- reflection overrides antisocial but not always prosocial motives. *Frontiers in Behavioral Neuroscience*, *9*, 287.
- Cornelissen, G., Dewitte, S., & Warlop. L. (2011). Are social value orientations expressed automatically?

 Decision making in the Dictator Game. *Personality and Social Psychology Bulletin*, *37*, 1080-1090.
- Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014a). "Piensa" twice: On the foreign language effect in decision making. *Cognition*, 130, 236–254.
- Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B.(2014b). Your morals depend on language. *PLoS ONE*, 9, e94842.
- Costa, A., Vives, M., & Corey, J. D. (2017). On language processing shaping decision making. *Current Directions in Psychological Science*, 26, 146–151.
- Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, *320*, 1739-1739.
- Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010a). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion*, 10, 855-862
- Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010b). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107, 17433-17438.
- Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47, 448-474.
- Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment.

 Frontiers in Psychology, 3, 328.
- Curtin, J., & Fairchild, B. A. (2003). Alcohol and cognitive control: Implications for regulation of behavior during response conflict. *Journal of Abnormal Psychology*, 112, 424-436.
- Danziger, S., Levav, J., Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108, 6889-6892.

- Dawes, R. (1980). Social dilemmas. Annual Review of Psychology, 31, 169-193.
- Debey, E., Verschuere, B., & Crombez, G. (2012). Lying and executive control: An experimental investigation using ego depletion and goal neglect. *Acta Psychologica*, *140*, 133-141.
- Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, 78, 97-119.
- De Dreu, C. K. W., Dussel, B. D., & Ten Velden, F. S. (2015) In intergroup conflict, self-sacrifice is stronger among pro-social indviduals, and parochial altruism emerges especially among cognitively taxed individuals. *Frontiers in Psychology*, *6*, 572.
- De Dreu, C. K. W., Scholte, H. S., van Winden, F. A. A. M., & Ridderinkhof, K. R. (2015). Oxytocin tempers calculated greed but not impulsive defense in predator-prey contests. *Social Cognitive and Affective Neuroscience*, 10, 721-728.
- De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7, 28-38.
- De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing.

 Current Directions in Psychological Science. Online first:

 https://doi.org/10.1177/0963721419855658.
- Dickinson, D. L., & McElroy, T. (2017) Sleep restriction and circadian effects on social decisions. *European Economic Review*, 97, 57-71.
- Dittrich, M. (2015). Gender differences in trust and reciprocity: Evidence from a large-scale experiment with heterogeneous subjects. *Applied Economics*, 47, 3825-3838.
- Døssing, F., Piovesan, M., & Wengström, E. (2017). Cognitive load and cooperation. *Review of Behavioral Economics*, *4*, 69-81.
- Duffy, S., & Smith, J. (2014). Cognitive load in the multiplayer prisoner's dilemma game: Are there brains in games? *Journal of Behavioral and Experimental Economics*, 51, 47-56.
- Durante, R., Putternam, L., & Van der Weele, J. (2014). Preferences for redistribution and perception of fairness: An experimental study. *Journal of the European Economic Association*, 12, 1059-1086.

- Eagly, A. H. (1987). Sex differences in social behavior: A social-role interpretation. Mahwah, New Jersey: L. Erlbaum Associates.
- Engel, C. (2011). Dictator games: A meta-study. Experimental Economics, 14, 583-610.
- Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin prefernces in simple distribution experiments. *American Economic Review*, 94, 857-869.
- Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, 49, 709-724.
- Epstein, S., & Pacini, R. (1999). Some basic issues regarding dual-process theories from the perspective of cognitive–experiential self-theory. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 462-482). New York, NY, US: Guilford Press.
- Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, 71, 390-405.
- Erat, S., & Gneezy, U. (2012). White lies. Management Science, 58, 723-733.
- Evans, A. M., & Brandt, M. J. (2019). Comparing the effects of hypothetical moral preferences on reallife and hypothetical behavior: commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, 30, 1380-1382.
- Evans, A. M., Dillon, K. D., & Goldin, G., & Krueger, J. I. (2011). Trust and self-control: The moderating role of the defaul. *Judgment and Decision Making*, *6*, 697-705.
- Evans, A. M., Dillon, K. D., & Rand, D. G. (2015). Fast but not intuitive, slow but not deliberative:

 Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology:*General, 144, 951-966.
- Evans, A. M., & Rand, D. G. (2018). Cooperation and decision time. *Current Opinion in Psychology*, 26, 67-71.
- Evans, J. St. B. T. (1989). Bias in Human Reasoning: Causes and Consequences. Brighton, UK: Erlbaum.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: extension and evaluation.

 *Psychonomic Bulletin and Review, 13, 378-395.

- Evans, J. St. B. T. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*. Hove, England: Psychology Press.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 29, 255-278.
- Evans, J. St. B. T., & Over, D. E. (1996). Rationality and Reasoning. Hove, UK: Psychol. Press.
- Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8, 223-241.
- Everett, J. A. C., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism.

 Frontiers in Behavioral Neuroscience, 9, 15.
- Everett, J. A. C., Ingbretsen, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior Even towards competitive outgroups, even when using control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology*, 73, 76-81.
- Everett, J. A. C., Ingbretsen, Z., Cushman, F., & Cikara, M. (2018). Aggression, fast and slow: Intuition favors defensive aggression. *Available at https://psyarxiv.com/4v39b/*
- Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145, 772-787.
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. Nature, 425, 785-791.
- Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8, 185-190.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137-140.
- Fehr, E., Glätzle-Rützler, D., & Sutter, M. (2013). The development of egalitarianism, altruism, spite and parochialism in childhood and adolescence. *European Economic Review*, 64, 369-383.
- Fehr, E., Naef, M., & Schmidt, K. M. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *The American Economic Review*, *96*, 1912-1917.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114, 817-868.

- FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*, 434-441.
- Fellows, L. K., Heberlein, A. S., Morales, D. A., Shivde, G., Waller, S., & Wu, D. H. (2005). Method matters: An empirical study of impact in cognitive neuroscience. *Journal of Cognitive Neuroscience*, 17, 850-858.
- Fennis, B. M., Janssen, & Vohs, K. (2008). Acts of benevolence: A limited-resource account of compliance with charitable requests. *Journal of Consumer Research*, *35*, 906-924.
- Ferguson, E., Maltby, J., Bibby, P. A., & Lawrence, C. (2014). Fast to forgive, slow to retaliate: Intuitive responses in the ultimatum game depend on the degree of unfairness. *PLoS ONE*, *9*, e96344.
- Ferrara, M., Bottasso, A., Tempesta, M., Carrieri, M., De Gennaro, L., & Ponti, G. (2015). Gender differences in sleep deprivation effects on risk and inequality aversion: Evidence from an economic experiment. *PLoS ONE*, *10*, e0120029.
- Fillmore, M., Carscadden, J., Vogel-Sprott, M. (1998). Alcohol, cognitive impairment, and expectancies. *Journal of Studies on Alcohol*, *59*, 174-179.
- Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise An experimental study on cheating. *Journal of the European Economic Association*, 11, 525-547.
- Fodor, J. (1983). The Modularity of Mind. Scranton, PA: Crowell.
- Fodor, J. (2001). The Mind Doesn't Work That Way. Cambridge, MA: MIT Press.
- Foerster, A., Pfister, R., Schmidts, C., Dignath, D., & Kunde, W. (2013). Honesty saves time (and justifications). *Frontiers in Psychology*, *4*, 473.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25-42.
- Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: a process dissociation analysis. *Personality and Social Psychology Bulletin*, 41, 696-713.
- Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is ego depletion real?

- An analysis of arguments. Personality and Social Psychology Review, 23, 107-131.
- Friehe, T., & Schildberg-Hörisch, H. (2017). Self-control and crime revisited: Disentangling the effect of self-control on risk taking and antisocial behavior. *International Review of Law and Economics*, 49, 23-32.
- Fromell, H., Nosenzo, D., & Owens, T. (2018). Altruism, Fast and Slow? Evidence from a meta-analysis and a new experiment. *Available at https://www.nottingham.ac.uk/cedex/documents/papers/cedex-discussion-paper-2018-13.pdf*
- Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., ... & Cappa, S. (2010). Gender-related differences in moral judgments. *Cognitive processing*, *11*, 219-226.
- Gabarró, J., García, A., & Serna, M. (2011). The complexity of game isomorphism. *Theoretical Computer Science*, 6675-6695.
- Gailliot, M., Baumeister, R., DeWall, C., Maner, J., Plant, E., Tice, D., Brewer, L., & Schmeichel, B. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, 92, 325-336.
- Gargalianou, V., Urbig, D., & van Witteloostuijn, A. (2017). Cooperating or competing in three languages: Cultural accommodation or alienation? *Cross Cultural & Strategic Management, 24*, 167-191.
- Gärtner, M. (2018). The prosociality of intuitive decisions depends on the status quo. *Journal of Behavioral and Experimental Economics*, 74, 127-138.
- Geipel, J., Hadjichristidis, C., & Surian, L.(2015a). The foreign language effect on moral judgment: The role of emotions and norms. *PLoS ONE*, *59*, 8–17.
- Geipel, J., Hadjichristidis, C., & Surian, L.(2015b). How foreign language shapes moral judgment. *Journal of Experimental Social Psychology*, 10, e0131529.
- Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*, 1-44.
- Gevins, A. S., & Cutillo, B. C. (1993). Neuroelectric evidence for distributed processing in human

- working memory. Electroencephalography and Clinical Neurophysiology, 87, 128-143.
- Gilbert, D. T., Giesler, R. B., & Morris, K. A. (1996). When comparisons arise. *Journal of Personality* and Social Psychology, 69, 227-236.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60, 509-517.
- Gilbert, D. T., Tafarodi, R. W., Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221-233.
- Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, 93, 285-292.
- 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, 191-203.
- Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution* and Human Behavior, 24, 153-172.
- Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks:

 Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, 3, 215-228.
- Glöckner, A., & Witteman, C. L. M. (2010). Beyond dual-process models: A categorization of processes underlying intuitive judgment and decision making. *Thinking & Reasoning*, 16, 1-25.
- Goeschl, T., & Lohse, J. (2018). Cooperation in public good games. Calculated or confused? *European Economic Review*, 107, 185-203.
- Gordijn, E. H., Hindriks, I., Koomen, W., Dijksterhuis, A., & Van Knippenberg, A. (2004).

 Consequences of stereotype suppression and internal suppression motivation: A self-regulation approach. *Personality and Social Psychology Bulletin*, 30, 212-224.
- Gotlib, T., & Converse, P. (2010). Dishonest Behavior: The impact of prior self-regulatory exertion and personality. *Journal of Applied Social Psychology*, 40, 3169–3191.

- Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, 95, 384-394.
- Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108, 419-453.
- Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, 93, 293-300.
- Gravelle, J. G. (2009). Tax havens: Internation tax avoidance and evasion. *National Tax Journal*, *68*, 727-753.
- Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In *Public self and private self* (pp. 189-212). Springer, New York, NY.
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11, 322-323.
- Greene, J. D., Morelli, S. A., Lowenberg. K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107, 1144-1154.
- 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, 2105-2108.
- Griesinger, D. W., & Livingston, J. W. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral Science*, *18*, 173–188.
- Grimm, V., & Mengel, F. (2011). Let me sleep on it: delay reduces rejection rates in ultimatum games. *Economics Letters*, 111, 113-115.
- Grolleau, G., Sutan, A., El Harbi, S., Jedidi, M. (2018). Do we need more time to give less? Experimental evidence from Tunisia. *Bulletin of Economic Research*, 70, 400-409.
- Gross, J., Emmerling, F., Vostroknutov, A., & Sack, A. T. (2018). Manipulation of pro-sociality and rule-following with non-invasive brain stimulation. *Scientific Reports*, 8, 1827.
- Grossman, Z., & Van der Weele, J. J. (2017). Dual-process reasoning in charitable giving: Learning from non-results. *Games*, 8, 36.

- Gunia, B. C., Wang, L., Huang, L., Wang, J. and Murnighan, J. K. (2012). Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management Journal*, 55, 13–33.
- de Haan, T., & van Veldhuizen, R. (2015). Willpower depletion and framing effects. *Journal of Economic behavior and Organization*, 117, 47-61.
- Hadjichristidis, C., Geipel, J. & Savadori, L. (2015). The effect of foreign language on perceived risk and benefit. *Journal of Experimental Psychology: Applied, 21*, 117-129.
- Hadjichristidis, C., Geipel, J. & Surian, L. (2017). How foreign language affects decisions: Rethinking the brain-drain model. *Journal of International Business Studies*, 48, 645-651.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546-573.
- Halali, E., Bereby-Meyer, Y., Meiran, N. (2014). Between self-interest and reciprocity: The social bright side of self-control failure. *Journal of Experimental Psychology: General*, 143, 745-754.
- Halali, E., Bereby-Meyer, Y., Ockenfels, A. (2013). Is it all about the self? The effect of self-control depletion on ultimatum game proposers. *Frontiers in Human Neuroscience*, *7*, 240.
- Hallsson, B. G., Siebner, H. R., & Hulme, O. J. (2018). Fairness, fast and slow: A review of diual process models of fairness. *Neuroscience & Biobehavioral Reviews*, 89, 49-60.
- Hammond, K. R. (1996). Human Judgment and Social Policy. New York: Oxford Univ. Press.
- Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 28-43.
- Hardin, G. (1968). The tragedy of the commons. Science, 162, 1243-1248.
- Hare T.A., Camerer C.F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, *324*, 646–648.
- Hauge, K. E., Brekke, K. A., Johansson, L. O., Johansson-Stenman, O., & Svedsäter, H. (2016).

- Keeping others in our mind or in our heart? Distribution games under cognitive load. *Experimental Economics*, 19, 562-576.
- Hayakawa, S., Costa, A., Foucart, A., & Keysar, B. (2016). Using a foreign language changes our choices. *Trends in Cognitive Sciences*, 20, 791–793.
- Hayakawa, S., Tannenbaum, D., Costa, A., Corey, J. D., & Keysar, B. (2017). Thinking more or feeling less? Explaining the foreign language effect on moral judgment. *Psychological Science*, 28, 1387–1397.
- Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology*, 92, 81-92.
- Hill, K. Altruistic cooperation during foraging by the Ache, and the evolved human predisposition to cooperate. *Human Nature*, *13*, 105-128.
- Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 298-306.
- Hochman, G., Ayal, S., & Ariely, D. (2015). Fairness requires deliberation: The primacy of economic over social considerations. *Frontiers in Psychology*, *6*, 747.
- Holbein, J. B., Schafer, J. P., & Dickinson, D. L. (2019). Insufficient sleep reduces voting and other prosocial behaviours. *Nature Human Behaviour*, *3*, 492-500.
- Horne, J. A. (1993). Human sleep, sleep loss and behaviour. Implications for the prefrontal cortex and psychiatric behaviour. *British Journal of Psychiatry*, *162*, 413-419.
- Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General*, 133, 23-30.
- Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science*, *320*, 1092-1095.
- Hutcherson, C. A., Plassmann, H., Gross, J. J., & Rangel, A. (2012). Cognitive regulation during decision making shifts behavioral control between ventromedial and dorsolateral prefrontal value systems. *Journal of Neuroscience*, 32, 13543–13554.

- Inzlicht, M., Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science*, *7*, 450-463.
- Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, 18, 127-133.
- Isler, O., Maule, J., & Starmer, C. (2018). Is intuition really cooperative? Improved tests support the social heuristics hypothesis. *PloS one*, *13*, e0190560.
- Isler, O., & Yilmaz, O. (2019). Intuition and deliberation in morality and cooperation: An overview of the literature. In Liebowitz, J. (Ed.), *Developing informed intuition for decision-making* (pp. 101-115).
 Boca Raton, FL, US: CRC Press.
- Itzchakov, G., Uziel, L., & Wood, W. (2018). When attitudes and habits don't correspond: Self-control depletion increases persuasion but not behavior. *Journal of Experimental Social Psychology*, 75, 1-10.
- Jacoby, L. L. (1991). A process dissociation framework: Separating auto- matic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
- Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology: General*, 148, 1008-1021.
- Jagau, S., van Veelen, M. (2018). A general evolutionary framework for the role of intuition and deliberation in cooperation. *Nature Human Behaviour*, 1, 0152.
- Janssen, L., Fennis, B. M., Pruyn, Th. H. A., & Vohs, K. (2009). The path of least resistance: Regulatory resource depletion and the effectiveness of social influence techniques. *Journal of Business Research*, 61, 1041-1045.
- Jarke, J., & Lohse, J. (2016). I'm in a hurry, I don't want to know! The effects of time pressure and transparency on self-serving behavior. *Available at SSRN:* https://ssrn.com/abstract=2823678.
- Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., & Pascual-Leone, A. (2014). TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. *Frontiers in Neuroscience*, 8, 18.

- Jordan, J. J., McAuliffe, K., & Rand, D. G. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19, 741-763.
- Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*, 551-560.
- Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 134, 193-209.
- Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J.
 (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, 125, 131-164.
- Kahane, G., & Shackel, N. (2010) Methodological issues in the neuroscience of moral judgment. *Mind & Language*, 25, 561-582.
- Kahneman, D. (2011). Thinking, fast and slow. New York, NY: Farrar, Straus and Giroux.
- Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In K. J. Holyoak & R. G.
 Morrison (Eds.), The Cambridge handbook of thinking and reasoning (pp. 267-293). New York, NY,
 US: Cambridge University Press.
- Kamphuis, J., Meerlo, P., Koolhaas, J. M., & Lancel, M. (2012). Poor sleep as a potential causal factor in aggression and violence. *Sleep Medicine*, *13*, 327-334.
- Kant, I. (2002). *Groundwork for the metaphysics of morals*. New Haven, CT: Yale University Press. (Original work published 1797).
- Kaplan, H., Hill, J., Lancaster, J., & Hurtado, A. M. (2000). A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology*, *9*, 156-185.
- Kessler, J. B., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: Generosity over time.

 Available at
 - https://users.nber.org/~kesslerj/papers/KesslerKivimakiNiederle GenerosityOverTime.pdf
- Kessler, J. B., & Meier, S. (2014). Learning from (failed) replications: Cognitive load manipulation and

- charitable giving. Journal of Economic Behavior & Organization, 102, 10-13.
- Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*, 661-668.
- Kinnunen, S. P., & Windmann, S. (2013). Dual-processing altruism. Frontiers in Psychology, 4, 193.
- Köbis, N., Verschuere, B., Bereby-Meyer, Y., Rand, D. G., & Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic. *Perspectives on Psychological Science*, 14, 778-796.
- Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: Evidence from the Ultimatum Game. *Journal of Neuroscience*, 27, 951-956.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007).

 Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908-911.
- Kohlberg, L. (1931). The philosophy of moral development: Moral stages and the idea of justice (Vol. 1).

 San Francisco: Harper & Row.Kohlber, L. (1963). The development of children's orientation towards moral order. *Human Development*, *6*, 11-33.
- Kollock, P. (1988). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24*, 183-214.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*, 829-832.
- Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008).
 Studying the neurobiology of social interaction with transcranial direct current stimulation The example of punishing unfairness. *Cerebral Cortex*, 18, 1987-1990.
- Knoch, D., Schneider, F., Schunk, D., Hohmann, M., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences*, 106, 20895-20899.
- Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*, 7455.
- Krawczyk, M., & Sylwestrzak, M. (2018). Exploring the role of deliberation time in non-selfish behavior:

- The double response method. Journal of Behavioral and Experimental Economics, 72, 121-134.
- Krupp, D. B., DeBruine, L. M., & Barclay, P. (2008). A cue of kinship promotes cooperation for the public good. *Evolution and Human Behavior*, 29, 49-55.
- Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological Science*, *24*, 2498-2504.
- Kuene, M., Heimrath, K., Heinze, H.-J., & Zaehle, T. (2015). Transcranial direct current stimulation of the left dorsolateral prefrontal cortex shifts preference of moral judgments. *Plos One, 10,* e0127061.
- Kuo, W.-J., Sjöström, Chen, Y.-P., Wang, W.-H., & Huang, C.-Y. (2009). Intuition and deliberation: Two systems for strategizing in the brain. *Science*, *324*, 519-522.
- Kvaran, T., Nichols, S., & Sanfey, A. (2013). The effect of analytic and experiential modes of thought on moral judgment. *Progress in Brain Research*, 202, 187-196.
- Kvarven, A., Strømland, E., Wollbrant, C., Andersson, D., Johannesson, M., Tinghög, G., Västfjäll, D., & Ove, K. (2019). The intuitive cooperation hypothesis revisited: A meta-analytic examination of effect-size and between-study heterogeneity. *Available at https://osf.io/preprints/metaarxiv/kvzg3/*
- Levine, D. K. (1998). Modeling altruist and spitefulness in experiments. *Review of Economic Dynamics*, 1, 593-622.
- Levine, E., Barasch, A., Rand, D. G., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General*, 5, 702-719.
- Li, J., Liu, X., Yin, X., Wang, G., Niu, X., & Zhu, C. (2018). Transcranial direct current stimulation altered voluntary cooperative norms compliance under equal decision-making power. *Frontiers in Human Neuroscience*, 12, 265.
- Li, Z., Xia, S., Wu, X., & Chen, Z. (2018). Analytical thinking style leads to more utilitarian moral judgments: An exploration with a process-dissociation approach. *Personality and Individual Differences*, *131*, 180–184.
- Lian, H., Yam, K. C., Ferris, D. L., & Brown, D. (2017). Self-control at work. *Academy of Management Annals*, 11, 703-732.

- Libet, B. (2009). Mind time: The temporal factor in consciousness. Harvard University Press.
- Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. R. Williams, & W. von Hippel (Eds.). *Social Judgments: Implicit and Explicit Processes* (pp. 44-67). New York: Cambridge Univ. Press.
- Liebrand, W. B. G. (1984). The effect of social motives, communication and group size on behaviour in an n-person multi stage mixed motive game. *European Journal of Social Psychology*, 14, 239–264.
- Liu, S., & Hao, F. (2011). An application of a dual-process approach to decision making in social dilemmas. *American Journal of Psychology*, 124, 203-212.
- Liu, Y., He, N., & Dou, K. (2015). Ego-depletion promotes altruistic punishment. *Open Journal of Social Science*, *3*, 62-69.
- Lohse, J. (2016). Smart or selfish When smart guys finish nice. *Journal of Behavioral and Experimental Economics*, 64, 28-40.
- Lohse, J., Goeschl, T., & Diederich, J. H. (2017). Giving is a question of time: Response times and contributions to an environmental public good. *Environmental and Resource Economics*, 67, 455-477.
- Lohse, T., Simon, S. A., & Konrad, K. A. (2018). Deception under time pressure: Conscious decision or a problem of awareness? *Journal of Economic Behavior & Organization*, *146*, 31-42.
- Lotito, G., Migheli, M., & Ortona, G. (2013). Is cooperation instinctive? Evidence from the response times in a public goods game. *Journal of Bioeconomics*, *15*, 123-133.
- Lotz, S. (2015). Spontaneous giving under structural inequality: Intuition promotes cooperation in asymmetric social dilemmas. *PLoS ONE*, *10*, e0131562.
- Martinsson, P., Myrseth, K. O. R., & Wollbrant, C. (2012). Reconciling pro-social vs selfish behavior: On the role of self-control. *Judgment and Decision Making*, 7, 304-315.
- Martinsson, P., Nordblom, K., Rützler, D., & Sutter, M. (2011). Social prefenreces during childhood and the role of gender and age An experiment in Austria and Sweden. *Economics Letters*, 110, 248-251.

- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45, 633-644.
- 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 USA*, 114, 4360-4364.
- McLeish, K. N., & Oxoby, R. J. (2011). Social interactions and the salience of social identity. *Journal of Economic Psychology*, 32, 172-178.
- 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, 594–597.
- Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Science*, 22, 280-293.
- Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, 18, 193-197.
- Merkel, A., & Lohse, J. (2019). Is fairness intuitive? An experiment accounting for subjective utility differences under time pressure. *Experimental Economics*, 22, 24-50.
- Messick, D. M., & McClintock, C. G. (1968). Motivational Bases of Choice in Experimental Games. *Journal of Experimental Social Psychology*, 4, 1–25.
- Meuwese, R., Crone, E. A., de Rooij, M., Güroğlu. (2015) Development of equity preferences in boys and girls across adolescence. *Child Development*, 86, 145-158.
- Milinski, M., Semman, D., & Krambeck, H. J. (2002). Reputation helps solve the "tragedy of the commons". *Nature*, *415*, 424-426.
- Mill, J. S. (1863). *Utilitarianism*. London, England: parker, Son, and Bourne.
- Mischkowski, D., & Glöckner, A. (2016). Spontaneous cooperation for prosocials, but not proselfs: Social value orientation moderates spontaneous cooperative behavior. *Scientific Reports*, *6*, 21555.
- Mischkowski, D., Glöckner, A., & Lewisch, P. (2018). From spontaneous cooperation to spontaneous

- punishment Distinguishing the underlying motives driving spontaneous behavior in first and second order public games. *Organizational Behavior and Human Decision Processes*, 149, 59-72.
- Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.
- Moller, A. C., Deci, E. L., & Ryan, R. M. (2006). Choice and ego-depletion: The moderating role of autonomy. *Personality and Social Psychology Bulletin*, *32*, 1024-1036.
- Moore, C., Tenbrunsel, A. E. (2014). "Just think about it"? Cognitive complexity and moral choice.

 Organizational Behavior and Human Decision Processes, 123, 138-149.
- Morewedge. C. K., Krishnamurti, T., & Ariely, D. (2014). Focused on fairness: Alcohol intoxication increases the costly rejection of inequitable rewards. *Journal of Experimental Social Psychology*, 50, 15-20.
- Moskowitz, H., Burns, M. M., Williams, A. F. (1985). Skills performance at low blood alcohol levels. *Journal of Studies on Alcohol*, 46, 482-485.
- Mosleh, M., & Rand, D. G. (2018). Population structure promotes the evolution of intuitive cooperation and inhibits deliberation. *Scientific Reports*, *8*, 6293.
- Motro, D., Ordonez, L. D., Pittarello, A., & Welsh, D. T. (2018). Investigating the Effects of Anger and Guilt on Unethical Behavior: A Dual-Process Approach. *Journal of Business Ethics*, *152*, 133-148.
- Mrkva, K. (2017). Giving, fast and slow: Reflection increases costly (but not uncostly) charitable giving. *Journal of Behavioral Decision Making*, 30, 1052-1065.
- Muda, R., Niszczota, P., Białek, M., & Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *Journal of Experimental Psychology, Learning, Memory, & Cognition, 44*, 321-326.
- Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psyhological Bulletin*, *126*, 247-259.
- Muraven, M., Pogarsky, G., & Shmueli, D. (2006). Self-control depletion and the general theory of crime. *Journal of Quantitative Criminology*, 22, 263–277.

- Murphy, R. O., Ackermann, K. A., Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, 6, 771-781.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human behaviour*, 3, 221-229.
- Muzur, A., Pace-Schott, E. F., & Hobson, J. A. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Sciences*, 6, 475-481.
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*, 48-49.
- Neo, W. S., Yu, M., Weber, R., & Gonzalez, C. (2013). The effects of time delay in reciprocity games. *Journal of Economic Psychology, 34*, 20-35.
- Nielsen, U. H., Tyran, J.-R., & Wengström, E. (2014). Second thoughts on free riding. *Economics Letters*, 122, 136-139.
- Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic vs. analytic cognition. *Psychological Review*, 108, 291–310.
- Nowak, M. A. (2006). Five rules for the evolution of cooperation. Science, 314, 1560-1563.
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437, 1291-1298.
- Oechssler, J., Roider, A., & Schmitz, P. W. (2015). Cooling off in negotiations: Does it work? *Journal of Institutional and Theoretical Economics*, 171, 565-588.
- Okun, A. M. (2015). Equality and efficiency: The big tradeoff. Brookings Institution Press.
- Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *Journal of Experimental Psychology:*General, 147, 462-484.
- Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943-950.
- Osborne, M., & Rubinstein, A. (1994). A course in game theory. MIT Press.

- Osgood, M. J., & Muraven, M. (2015). Self-control depletion does not diminish attitudes about being prosocial but diminish prosocial behaviors. *Basic Applied Social Psychology*, *37*, 68-80.
- Ostrom, E. (2000). Collective action and the evoltion of social norms. *Journal of Economic Perspectives*, 14, 137-158.
- Pacini, R., & Epstein, S. (1999). The relation of rational and experential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76, 972-987.
- Pavey, L., Greitemeyer, T., & Sparks, P. (2011). Highlighting relatedness promotes prosocial motives and behavior. *Personality and Social Psychology Bulletin*, *37*, 905-917.
- Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*, 163-17.
- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, 48, 341-348.
- Pennycook, G., De Neys, W., Evans, J. St. B. T., Stanovich, K. E., & Thompson, V. (in press) The mythical dual-process typology. *Trends in Cognitive Science*.
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80, 34-72.
- Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., & Szolnoki, A. (2017). Statistical physics of human cooperation. *Physics Reports*, 687, 1-51.
- Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a "cooperative phenotype" that is domain general and temporally stable. *Nature Communications*, *5*, 4939.
- Pfattheicher, S., Keller, J., Knezevic, G. (2017). Sadism, the intuitive system, and antisocial punishment in the public goods game. *Personality and Social Psychology Bulletin*, *43*, 337-346.
- Piovesan, M., & Wengström, E. (2009). Fast or fair? A study of response times. *Economics Letters*, 105, 193-196.
- Pitesa, M., Thau, S., & Pillutla, M. M. (2013). Cognitive control and socially desirable behavior: The role

- of interpersonal impact. Organizational Behavior and Human Decision Processes, 122, 232-243.
- Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, 1389-1391.
- Popper, K. (1959). The Logic of Scientific Discovery. London: Hutchison.
- Rahman, A., Reato, D., Arlotti, M., Gasca, F., Datta, A., Parra, L. C., et al. (2013). Cellular effects of acute direct stimulation: Somatic and synaptic terminal effects. *Journal of Physiology*, 591, 2563-2578.
- Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*, 1192-1206.
- Rand, D. G. (2019). Intuition, deliberation, and cooperarion: Further meta-analytic evidence from 91 experiments on pure cooperation. *Available at:*https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3390018
- Rand, D. G., Brescoll, V. L., Everett, J. A. C., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General*, 145, 389-396.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489, 427-430.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2013). Rand et al. reply. Nature, 498, E2.
- Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. Frontiers in Behavioral Neuroscience, 8, 300.
- Rand, D. G., Newman, G. E., & Wurzbacher, O. M. (2015). Social context and the dynamic of cooperative choices. *Journal of Behavioral Decision Making*, 28, 159-166.
- Rand, D. G., & Nowak, M. A. (2013). Human cooperation. Trends in Cognitive Sciences, 17, 413-425.
- Rand, D. G., Peysakhovich, A., Kraft-Todd, G., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*, 3677.

- Rantapuska, E., Freese, R., Jääskeläinen, I. P., & Hytönen, K. (2017) Does short-term hunger increase trust and trustworthiness in a high trust society? *Frontiers in Psychology*, *8*, 1944.
- Rawls, J. (2009). A theory of justice. Harvard University Press.
- Reber, A. S. (1993). Implicit Learning and Tacit Knowledge. Oxford, UK: Oxford Univ. Press.
- Recalde, M. P., Riedl, A., & Vesterlund, L. (2018). Error-prone inference from response time: The case of intuitive generosity in public-good games. *Journal of Public Economics*, 160, 132-147.
- Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, 62, 23-48.
- Roch, S. G., Lane, J. A. S., Samuelson, C. D., Allison, S. T., Dent, J. L. (2000). Cognitive load and the equality heuristic: A two-stage model of resource overconsumption in small groups. *Organizational Behavior and Human Decision Processes*, 83, 185-212.
- Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, 69, 44-58.
- Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A. (2009). Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical Neurophysiology*, 120, 2008-2039.
- Rubinstein, A. (2007). Istinctive and cognitive reasoning: A study of response times. *The Economic Journal*, 117, 1243-1259.
- Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, *342*, 482-484.
- Sacco, D. F., Brown, M., Lustgraaf, C. J., Hugenberg, K. (2016). The adaptive utility of deontology:

 Deontological decision-making fosters perceptions of trust and likeability. *Evolutionary*Psychological Science, 3, 125-132.
- Sæetrevik, B., & Sjåstad, H. (2019). A pre-registered attempt to replicate the mortality salience effect in traditional and novel measures. *Available at https://psyarxiv.com/dkg53/*

- Samson, K., & Kostyszyn, P. (2015). Effects of cognitive load on trusting behavior An experiment using the Trust Game. *PLoS ONE, 10*, e0127680.
- Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1, 7-59.
- Santa, J. C., Exadaktylos, F., & Soto-Faraco, S. (2018). Beliefs about others' intentions determine whether cooperation is the faster choice. *Scientific Reports*, 8, 7509.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing I: detection, search and attention. *Psychological Review*, 84, 1-66.
- 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.
- Sen, A. (2018). Collective choice and social welfare. Harvard University Press.
- Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications).

 *Psychological Science, 23, 1264–1270.
- Shenhav, A., Rand, D. G., Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. Journal of Experimental Psychology: General, 141, 423-428.
- 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, 278-292.
- Shiv, B., & Nowlis, S. (2004). The effect of distractions while tasting a food sample: The interplay of informational and affective components in subsequent choice. *Journal of Consumer Research*, 31, 599-608.
- Singer, P. (1979). Practical Ethics. Cambridge University Press, Cambridge, UK.
- Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
- Solnick, S. J. (2001). Gender differences in the ultimatum game. *Economic Inquiry*, 39, 189-200.
- Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions

- in the human brain. Nature Neuroscience, 11, 543-545.
- Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: Affect and deliberations in donation decisions. *Organizational Behavior and Human Decision Processes*, 102, 143-153.
- Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131.
- Spears, D., Fernández-Linsenbarth, I., Okan, Y., Ruz, M., & González, F. (2018). Disfluent fonts lead to more utilitarian decisions in moral dilemmas. *Psicológica Journal*, *39*, 41-63.
- Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Xheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, *12*, 2849-2853.
- Stanovich, K. E. (1999). Who is Rational? Studies of Individual Differences in Reasoning. Mahwah, NJ: Elrbaum.
- Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: Chicago Univ. Press.
- Stanovich, K. E. (2011). Rationality and the Reflective Mind. New York, NY: Oxford University Press.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality* and Social Psychology Review, 8, 220-247.
- Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. T. (2015). Be nice if you have to the neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, 10, 790-796.
- Strohminger, N., Lewis, R. L., & Meyer, D. E. (2011). Divergent effects of different positive emotions on moral judgment. *Cognition*, 119, 295-300.
- Strømland, E., Tjotta, S., & Torsvik, G. (2016). Cooperating, fast and slow: Testing the social heuristics hypothesis. *Available at SSRN:* https://ssrn.com/abstract=2780877
- Strømland, E., & Torsvik, G. (2019). Intuitive prosociality: Heterogeneous treatment effects or false

- positive? Available at https://osf.io/hrx2y/
- Stroop, J. R. (1935). Studies on interferences in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662.
- Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. Cognition, 119, 454-458.
- Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal*, 119, 47-60.
- Sutter, M., Kocher, M., & Strauß, S. (2003). Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, 81, 341-347
- Swann, W. B., Hixon, J. G., Stein-Seroussi, A., & Gilbert, D. T. (1990). The fleeting gleam of praise:

 Cognitive processes underlying behavioral reactions to self-relevant feedback. *Journal of Personality and Social Psychology*, *59*, 17-26.
- Tajfel, H. (1970). Experiments in intergroup discrimination. Scientific American, 223, 96-103.
- Tajfel, H. (1982). Social psychology of intergroup relations. Annual Review of Psychology, 33, 1-39.
- Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*, 149-178.
- Tappin, B. M., & Capraro, V. (2018). Doing good vs. avoiding bad in prosocial choice: A refined test and extension of the morality preference hypothesis. *Journal of Experimental Social Psychology*, 79, 64-70.
- Thoma, V., White, E., Panigrahi, A., Strowger, V., & Anderson, I. (2015). Good thinking or gut feeling? Cognitive reflection and intuition in traders, bankers and financial non-experts. *PLoS ONE, 10*, e0123202.
- Thompson, V. A., Turner, J. A. P., & Pennycook. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107-140.
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11, 99-113.
- Timmons, S., & Byrne, R. M. (2018). Moral fatigue: The effects of cognitive fatigue on moral reasoning.

- Quarterly Journal of Experimental Psychology, 174702181877204.
- Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Västfjäll, D., Kirchler,M., & Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498, E1.
- Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016).

 Intuition and moral decision-making The effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS ONE, 11,* e0164012.
- Toates, F. (2006). A model of the hierarchy of behaviour, cognition and consciousness. *Consciousness and Cognition*, 15, 75–118.
- Tomasello, M., Carpenter, M., Call, J, Behne, T., & Moll, H. (2005). In search of the uniquely human. Behavioral and Brain Sciences, 28, 721-727.
- Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, 40, 923-930.
- Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, 124, 379-384.
- Trivers, R. (1971). The evolution of reciprocal altruism. The Quarterly Journal of Biology, 46, 35-57.
- Trope, Y., & Alfieri, T. (1997). Effortfulness and flexibility of dispositional judgment processes. *Journal of Personality and Social Psychology*, 73, 662-674.
- Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12, 579-590.
- Urbig, D., Terjesen, S., Procher, V., Muehlfeld, K., & van Witteloostuijn, A. (2016). Come on and take a free ride: Contributing to public goods in native and foreign language settings. *Academy of Management Learning and Education*, 15, 268-286.
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment.

 *Psychological Science, 17, 476-477.
- Van der Cruyssen, I., D'hondt, J., Meijer, E., & Verschuere, B. (2019) Does honesty require time? Two preregistered replications of Experiment 2 of Shalvi, Eldar and Bereby-Meyer.

- https://www.researchgate.net/publication/333909777.
- Van Lange, P.A.M., Bekkers, R., Schuyt, T., & Van Vugt, M. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, 29, 375–384.
- van't Veer, A. E., Stel, M., & Van Beest, I. (2014). Limited capacity to lie: Cognitive load interferes with being dishonest. *Judgment and Decision Making*, *9*, 199–206.
- Verkoeijen, P. P. J. L., & Bouwmeester, S. (2014). Does intuition cause cooperation? *PLoS ONE*, 9, e96654.
- Verschuere, B., & Shalvi, S. (2014). The truth comes naturally! Does it? *Journal of Language and Social Psychology*, 33, 417-423.
- Vives, M. L-, Aparici, M., & Costa, A. (2018). The limits of the foreign language effect on decision-making: The case of outcome bias and the representativeness heuristic. *PLoS ONE*, *13*, e0203528.
- Vohs, K. D., Glass, B. D., Maddox, W. T., & Markman, A. B. (2011). Ego depletion is not just fatigue: Evidence from a total sleep deprivation experiment. *Social Psychological and Personality Science*, 2, 166-173.
- Volk, S., Köhler, T., & Pudelko, M. (2014). Brain drain: The cognitive neuroscience of foreign language processing in multinational corporations. *Journal of International Business Studies*, 45, 862-885.
- Walczyk, J. J., Mahoney, K. T., Doverspike, D., Griffith-Ross, D. A. (2009). Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business Psychology*, 24, 33-49.
- Walsh, V., & Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience*, 1, 73.
- Wang, C. S., Sivanathan, N., Narayanan, J., Ganegoda, D. B., Bauer, M., Bodenhausen, G. V., & Murnighan, K. (2011). Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organizational Behavior and Human Decision Processes*, 116, 46-54.
- Wang, Y., Zhang, X., Li, J., & Xie, X. (2019). Light in darkness: Low self-control promotes altruism in crises. *Basic and Applied Social Psychology*, 41, 201-213.

- Welsh, D. T., Ellis, A. P. J., Christian, M. S., & Mai, K. M. (2014). Building a self- regulatory model of sleep deprivation and deception: The role of caffeine and social influence. *The Journal of Applied Psychology*, 99(6), 1268–1277.
- Whitt, S., & Wilson, R. K. (2007). The dictator game, fairness and ethnicity in postwar Bosnia. *American Journal of Political Science*, *51*, 655-668.
- Wills, J. A., Hackel, L. M., & Van Bavel, J. J. (2018). Shifting prosocial intuitions: Neurocognitive evidence for a value based account of group-based cooperation. *Available at*https://psyarxiv.com/u736d/
- Wills, J. A., FeldmanHall, O., NYU Prospect Collaboration, Meager, M. R., Van Bavel, J. J. (2018).
 Dissociable contributions of the prefrontal cortex in group-based cooperation. *Social Cognitive Affect Neuroscience*, 13, 349-356.
- Wilson, T. D. (2002). Strangers to Ourselves. Cambridge, MA: Belknap.
- Yam, K. C. (2018). The effects of thought suppression on ethical decision making: Mental rebound versus ego depletion. *Journal of Business Ethics*, *147*, 65-79.
- Yam, K. C., Chen, X. P., & Reynolds, S. J. (2014). Ego depletion and its paradoxical effects on ethical decision making. *Organizational Behavior and Human Decision Processes*, 124, 204-214.
- Yamagishi, T., Matsumoto, Y., Kiyonari, T., Takagishi, H., Li, Y., Kanai, R., & Sakagami, M. (2017).

 Response time in economic games reflects different types of decision conflict for prosocial and proself individuals. *Proceedings of the National Academy of Sciences*, 114, 6394-6399.
- Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, 145, 1448-1459.
- Zhong, C.-B. (2011). The ethical dangers of deliberative decision making. *Administrative Science Quarterly*, *56*, 1–25.

Footnotes

¹ I wrote this review in occasion of a PhD course I was invited to give at the IMT School of Advanced Studies of Lucca, Italy. I would like to thank Ennio Bilancini for the invitation and all the course participants for providing very helpful comments. I would like to thank also the following scholars who have commented on a first draft of this review: Luis Artavia-Mora, Ofer Azar, Sachin Banker, Nick Byrd, Fadong Chen, Mina Cikara, Paul Conway, Wim De Neys, David Dickinson, Antonio Espín, Jim A. C. Everett, Andreas Glöckner, Felisa González, Jörg Gross, Guy Itzchakov, Michele Garagnani, Constantinos Hadjichristidis, Yam Kai Chi, Michal Wiktor Krawczyk, Johannes Lohse, Daniele Nosenzo, Sebastian Olschewski, Marco Piovesan, David G. Rand, Shaul Shalvi, Eirik Strømland, Shane Timmons, Bastien Tremoliere, Diemo Urbig, Roel van veldhuizen, and Julian Wills. Finally, I would like to thank the Integrated Activity Project "Prosociality, Cognition, and Peer Effects" (PRO.CO.PE.) that provided the financial support for my stay in Lucca.

² In this review, I will focus only on experimental manipulations, which allows to draw a causal link between the independent and the dependent variables, as opposed to correlational studies that can be driven by unobserved variables. For example, since Type 1 is typically fast and Type 2 is typically slow, one might think that looking at the response time in the control condition, rather than experimentally manipulating response time, would be enough to draw conclusions regarding which choice is typically driven by Type 1 and which is typically driven by Type 2. Following this line of reasonings, some works have used response times to study the cognitive basis of different types of social behaviours (Piovesan & Wengström, 2009; Cappelen, Nielsen, Tungodden, Tyran & Wengström, 2016; Mischkowski, Glöckner & Lewisch, 2018; Chen & Fischbacher, 2019). However, it has been noticed that response time predicts decision conflict and choice discriminability, rather than deliberation, at least in some contexts (Evans, Dillon & Rand, 2015; Krajbich, Bartling, Hare & Fehr, 2015; Evans & Rand, 2018). In general, all correlational tests suffer of the same conceptual problem: any change in the dependent variable need not be driven by the independent variable under consideration, but it might be driven by other unobserved variables. For the same reason, I decided to not include in the review those studies exploring the cognitive

basis of sociality by comparing patients with brain lesions in areas typically associated to Type 1 (e.g., the ventromedial prefrontal cortex) and/or with patients with brain lesions in areas typically associated to Type 2 (e.g., the dorsolateral prefrontal cortex), with healthy participants. See Mendez et al. (2005), Ciaramelli et al. (2007); Koenigs et al. (2007), Koenigs & Tranel (2007), Wills et al. (2018), Zhu et al. (2014). Although I acknowledge the existence and the importance of these works, I have decided to leave them out of this review, exactly because participants are not randomized between conditions (for obvious reasons) and therefore these studies do not allow to make causal inferences regarding the role of Type 1 and Type 2 on social behaviour.

³ There are also studies exploring how inducing a particular emotion affect a subsequent decision (Strohminger, Lewis & Meyer, 2011; Ugazio, Lamm & Singer, 2012; Motro, Ordóñez, Pittarello & Welsh, 2018). These works have shown that different emotions affect social decisions differently. For example, Strohminger et al. (2011) and Ugazio et al. (2012) found that anger tend to increase consequentialist judgments while disgust tend to increase deontological judgments. Motro et al. (2018) let participants write about a time in their life in which they felt angry vs guilt; they found that inducing anger made people more likely to lie, while inducing guilt made people more honest. This emotion-dependent effect of priming suggests that inducing a particular emotion does not simply promote Type 1 processing. For this reason, in this review I will include only studies inducing people to "rely on emotion", *in general*, and not those inducing a particular emotion.

⁴ Some authors have argued that sleep deprivation and ego depletion are different (Vohs, Glass, Maddox & Markman, 2011). However, a review found a positive effect of sleep deprivation on aggression, a behaviour typically driven by lack of self-control, suggesting that sleep deprivation successfully impairs self-control (Kamphuis, Meerlo, Koolhaas & Lancel, 2012).

⁵ For example, Danziger, Levav and Avnaim-Pesso (2011) found that judges' favourable rulings drop from nearly 65% to nearly 0% from the morning to lunch time, and then again from after lunch time to dinner time.

⁶ Alcohol intoxication is known to diminish the executive cognitive functions necessary for effortful and controlled processing (Moskowitz, Burns & Williams, 1985; Fillmore, Carscadden & Vogel-Sprott, 1998; Curtin & Fairchild, 2003).

⁷ Mortality salience is a cognitive manipulation based on Terror Management Theory (Becker, 1973; Greenberg, Pyszczynski & Solomon, 1986). According to this theory, people rely on their cognitive resources to avoid the anxiety provoked by death-related thoughts. Therefore, reminding people about their mortality has the effect of depleting their cognitive resources. Several studies supported this view (see, e.g., Arnds, Greenberg, Solomon, Pyszczynski, and Simon, 1997). However, mortality salience also has been recently criticized, as a pre-registered replication of the original study failed to find an effect (Sætrevik & Sjåstad, 2019).

⁸ The Need for Cognition is a scale introduced by Cacioppo and Petty (1982) to measure the extent to which a person is inclined towards effortful cognitive activities.

⁹ The Faith in Intuition is a scale introduce by Epstein, Pacini, Denes-Raj and Heier (1996) to measure the extent to which a person is inclined towards effortless, intuitive, activities.

¹⁰ Note, indeed, that the public goods game reduces to the prisoner's dilemma when there are only two players, each of whom can either contribute to the public good the full endowment or nothing.

¹¹ There is some work looking at the correlation between response time and cooperation in the traveller's dilemma (Rubinstein, 2007; Santa, Exadaktylos, & Soto-Faraco, 2018), but, to the best of my knowledge, no studies exploring the causal link.

¹² There are also two correlational studies that are roughly in line with this view. Mischkowski and Glöckner (2016) found that response time is positively correlated with cooperation among prosocials, but not among pro-selves, for whom there is no correlation. Yamagishi et al. (2017) found the same result for pro-socials; however, among pro-selves, they found response time to be negatively correlated with cooperation.

¹³ Researchers have also explored the effect of ego-depletion on compliance to a charitable request (Fennis, Janssen & Vohs, 2008). Although interesting, I have opted for not reporting this line of

research in the main text because the effect does not seem to be driven by increased altruistic tendencies, but by increased compliance to authority (Janssen, Fennis, Pruyn & Vohs, 2009).

¹⁴ Fromell et al. (2018) included in their meta-analysis also studies using the dictator game with varying cost to benefit ratio. However, in a private communication, one of the authors (Daniele Nosenzo) confirmed that the results remain qualitatively the same when restricting the analysis to the standard dictator games.

¹⁵ Some authors have also considered situations in which participants are not incentivized to lie. Several works have provided evidence that, in this case, honesty tends to be intuitive (Spence et al. 2001; Walczyk, Mahoney, Doverspike & Griffith-Ross, 2009; Debey, Verschuere & Crombez, 2012).

¹⁶ There is also a lot of work on the effect of self-control depletion on ethical behaviour in the workplace (see Lian, Yam, Ferris and Brown, 2017, for a review. In the present article, I decided to focus specifically on the conflict between lying and truth-telling, which can be easily measured in the laboratory using economic games.

¹⁷ Note that this is not inconsistent with the result mentioned in Footnote 11, that honesty is intuitive when participants are not incentivized to lie. As observed by Verschuere and Shalvi (2014), "truth telling may be the natural response absent clear motivations to lie […] and that lying may prevail as the automatic reaction when it brings about important self-profit.

¹⁸ Conway, Goldstein-Greenwood, Polacek & Greene (2018) recently used a process dissociation approach to argue that "sacrificial utilitarian judgments [do] reflect genuine moral concern". According to this work, "antisociality predicts reduced deontological inclinations, not increased utilitarian inclinations".

¹⁹ A recent correlational paper argues that different types of reflection may determine different moral tendencies, with arithmetic reflection being correlated with utilitarian judgments and logical reflection being correlated with both utilitarian and deontological judgments (Byrd & Conway, in press).

²⁰ Along these lines, indeed, several works have shown that people making deontological judgments are judged better than those making utilitarian judgments, along a number of dimensions

(Bostyn & Roets, 2017; Everett, Pizarro & Crockett, 2016; Rom, Weiss & Conway, 2017; Sacco et al. 2016).

²¹ The terminology "secondary dimensions" is more than just a metaphor. Indeed, the simplest example of a weak embedding is the standard projection of the n-dimensional real space, endowed with the lexicographic order, into its first k-components (with n > k).

²² The particular voting rule does not really matter. I use the unanimity rule because, in reality, it seems reasonable to assume that another interaction will be played only if all players want to do it. However, it is well possible that there are specific situations in which other rules might apply. The general framework does not depend on the particular voting rule, although the details of the formal arguments might do.

²³ Note however, that results are more mixed among studies using neurostimulation or patients with brain damage, that are not included in these meta-analyses. These studies suggest that activation of the DLPFC has a positive effect on cooperation. To the extent that activation of the DLPFC actually promoted Type 2 processing, this finding would not be reconcilable with the GSHH. Future work should clarify this point. I will elaborate on this point in the final section.

²⁴ Observe that, although Table 2 seems to suggest that ego-depletion studies trend towards a negative effect of ego-depletion on altruism, the overall effect is not significantly different from zero (Fromell et al. 2018).

²⁵ With the exception of neurostimulation studies, which are typically not included in the metaanalyses. This exclusion, however, is not done on a theoretical ground. Simply, neurostimulation has emerged only recently as a research area and there are only a handful of papers that use this technique and most of them came out after the corresponding meta-analyses.

Tables

Table 1
List of empirical studies exploring the cognitive basis of cooperation.

	Measure	Main result
Time constraint studies	•	
Rand et al. (2012)	PGG	Pressure <i>increases</i> cooperation, compared to Delay
Tinghög et al. (2013)	PD	No effect of Pressure vs Delay
Tinghög et al. (2013)	Binary PGG	No effect of Pressure vs Delay
Cone & Rand (2014)	Competitive PGG	Pressure <i>increases</i> cooperation, compared to Delay
Rand et al. (2014)	PGG	Pressure <i>increases</i> cooperation, compared to Delay, but only among inexperienced participants
Rand & Kraft-Todd (2014)	PGG	Pressure <i>increases</i> cooperation, but only among participants that are naïve and trusting
Verkoeijen & Bouwmeester (2014)	PGG	No effect of Pressure vs Delay on cooperation
Capraro & Cococcioni (2015)	PD	No effect of Pressure vs Delay on cooperation
Rand et al. (2015)	PGG with outgroups	Pressure <i>increases</i> cooperation, compared to Delay
Capraro & Cococcioni (2016)	PD	Pressure <i>decreases</i> cooperation
Lohse (2016)	PGG	Pressure, compared to Baseline, decreases cooperation, but only among subjects with high CRT
Strømland et al. (2016)	PGG	Pressure has <i>no effect</i> on cooperation, compared to Delay
Bouwmeester et al. (2017)	21 preregistered PGGs	No effect of Pressure vs Delay on the whole sample. Pressure <i>increases</i> cooperation, among

	Measure	Main result
	Measure	participants who respect the
		time constraint
Kessler et al. (2017)	PD with varying b/c	Pressure <i>increases</i>
	<i>J</i> 8	cooperation for small b/c
		and decreases cooperation
		for big b/c
Everett et al. (2017)	PD	Pressure increases
		cooperation, compared to
		delay, also towards
		competitive out-groups
Goeschl & Lohse (2018)	PGG	Pressure decreases
.11/ 5	P.C.C.	cooperation
Alós-Ferrer & Garagnani (2018)	PGG	Pressure increases
		cooperation among
		prosocial participants; Pressure has <i>no effect</i> on
		individualist participants;
		Pressure <i>decreases</i>
		cooperation among
		competitive participants
Bird et al. (2019)	PGG only males	Pressure <i>increases</i>
	·	cooperation among risk-
		averse males; Pressure
		decreases cooperation
		among risk-seeking males
Conceptual prime studies		
Rand et al. (2012)	PGG	Intuition <i>increases</i>
		cooperation, compared to
		Deliberation
Lotz (2015)	Asymmetric PD	Intuition increases
		cooperation, compared to
11.1.1 4 -1 (2016)	DCC	Deliberation
Urbig et al. (2016)	PGG	Second language decreases
		cooperation, compared to mother tongue
Gargalianou et al. (2017)	PGG	No effect of second
Gargananou et al. (2017)	100	language on cooperation,
		compared to mother tongue
Levine et al. (2018)	PD	Emotion <i>increases</i>
()		cooperation, compared to
		Reason
Camerer et al. (2018)	PGG	No effect of Intuition vs
		Deliberation
Ego depletion studies		

	Measure	Main result
Capraro & Cococcioni (2016)	PD	Ego depletion decreases cooperation
Clark & Dickinson (2017)	PGG	No effect of sleep restriction on cooperation
Rantapuska et al. (2017)	PGG	No effect of hunger on cooperation
Neurostimulation studies		
Li et al. (2018)	Asymmetric PGG	Positive activation of rDLPFC <i>increases</i> cooperation, compared to "sham" baseline; negative activation of rDLPFC <i>decreases</i> cooperation, relative to sham
2-response paradigm studies		
Bago et al. (2019)	PGG	Deliberation has <i>no effect</i> on cooperation: most cooperative choices under deliberation are already cooperative under intuition

Note: PGG stands for "public goods game"; PD stands for "prisoner's dilemma".

Table 2
List of empirical studies exploring the cognitive basis of altruism.

	Measure	Main result
Time constraint studies		
Carlson et al. (2016)	DG charity	No effect of Pressure compared to Delay
Jarke & Lohse (2016)	DG	No effect of Pressure compared to Delay
Tinghög et al. (2016)	DG charity	No effect of Pressure compared to Delay
Artavia-Mora et al. (2017)	Field experiment	Pressure <i>increases</i> helping
Kessler et al. (2017)	DG	Pressure has <i>no effect</i> compared to Baseline
Mrkva (2017)	DG charity	Delay <i>increases</i> giving, compared to Baseline
Artavia-Mora et al. (2018)	Field experiment	Pressure increases helping
Andersen et al. (2018)	DG	No effect of Delay compared to Baseline
Chuan et al. (2018)	DG charity	Delay decreases giving
Gärtner (2018)	DG	Pressure <i>decreases</i> giving, only when the dictator game is presented with no status-
Strømland & Torsvik (2019)	DG	quo Pressure has <i>no effect</i> on giving
Grolleau et al. (2018)	DG	Delay decreases giving
Cognitive load studies		
Roch et al. (2000)	Taking from common pool	Load increases taking
Cornelissen et al. (2011)	DG	Load <i>increases</i> giving among pro-socials; no effect among individualists
Benjamin et al. (2013)	DG	Load has <i>no effect</i> on giving
Schulz et al. (2014)	DG	Load <i>increases</i> giving
Kessler & Meier (2014)	DG charity	Load has <i>no effect</i> on giving
Hauge et al. (2016)	DG	Load has no effect on giving
Tinghög et al. (2016)	DG charity	Load has <i>no effect</i> on giving
Grossman & Van der Weele (2017)	DG charity	Load has <i>no effect</i> on giving
Cognitive prime studies		
Small et al. (2007)	DG charity	Intuition <i>increases</i> giving, but only when the target is identifiable and not when is statistical

	Measure	Main result
Zhong (2011)	DG charity	"Feel" increases giving,
Kinnunen & Windmann (2013)	DG charity	compared to "Decide" Intuition has <i>no effect</i> on giving, compared to Deliberation
Ego depletion studies		
Halali et al. (2013)	DG	Depletion <i>decreases</i> frequency of equal split
Achtziger et al. (2015)	DG	Depletion decreases giving
Ferrara et al. (2015)	DG	Sleep restriction <i>decreases</i> giving, but only among females
Banker et al. (2017)	DG	Depletion has <i>no effect</i> on giving, it only makes participants more likely to choose the status quo
Dickinson & McElroy (2017)	DG	Sleep restriction <i>decreases</i> giving
Rantapuska et al. (2017)	DG charity	Hunger has <i>no effect</i> on giving
Itzchakov et al. (2018)	DG charity	Depletion <i>decreases</i> giving, but only when giving is accompanied by a persuasion message
Wang et al. (2019)	Helping in high-crisis scenarios	Depletion <i>increases</i> helping in high-crisis scenarios
Holbein et al. (2019)	DG charity	Sleep restriction decreases giving
Neurostimulation studies		
Ruff et al. (2013)	DG	rLPFC stimulation decreases giving, relative to sham; rLPFC inhibition increases giving
Strang et al. (2015)	DG	DLPFC stimulation increases giving, relative to sham; DLPFC inhibition decreases giving
Gross et al. (2018)	DG	rLPFC inhibition has <i>no</i> effect on altruism or selfishness, but only makes people more likely to follow rules of behavior

²⁻response paradigm studies

	Measure	Main result
Bago et al. (2019)	DG	Deliberation has <i>no effect</i> on altruism: most altruistic choices under deliberation are already altruistic under intuition

Note: DG stands for the standard dictator game; DG charity stands for the DG when the recipient of the donation is a charitable organization.

Table 3

List of empirical studies exploring the cognitive basis of truth-telling.

	Measure	Main result
Time constraint studies	•	
Gunia et al. (2012)	SR	Delay increases honesty,
Shalvi et al. (2012)	DR	compared to Baseline Pressure <i>decreases</i> honesty, compared to Baseline
Foerster et al. (2013)	DR	Pressure <i>increases</i> honesty, compared to Baseline
Capraro (2017)	SR	Pressure <i>increases</i> honesty, compared to Delay
Andersen et al. (2018)	SSC	Delay has <i>no effect</i> on honesty, compared to Baseline
Lohse et al. (2018)	SSC	Pressure <i>increases</i> honesty, by making participants less aware of the opportunity to lie
Capraro et al. (2019)	SR	Pressure <i>increases</i> honesty, compared to Delay
Van der Cruyssen et al. (2019)	DR	Pressure has <i>no effect</i> on cheating
Cognitive load studies		
van't Veer et al. (2014)	DR	Load increases honesty
Conceptual primes studies		
Cappelen et al. (2013)	SR Pareto lies	Intuition <i>increases</i> honesty, compared to Deliberation, but only for males
Zhong (2011)	SR	Emotion <i>increases</i> honesty, compared to Deliberation
Bereby-Meyer et al. (in press)	DR	Second language <i>decreases</i> honesty
Ego depletion studies		
Muraven et al. (2006)	SSC	Ego depletion decreases honesty
Mead et al. (2009)	SSC	Ego depletion decreases honesty
Gotlib & Converse (2010)	SSC	Ego depletion <i>decreases</i> honesty
Gino et al. (2011)	SSC	Ego depletion <i>decreases</i> honesty

	Measure	Main result
Pitesa et al. (2013)	SSC	Ego depletion decreases
		honesty, but only when
		consequences on others are
		not salient
Welsh et al. (2014)	SR	Ego depletion has <i>no effect</i>
Y 1 (2014)		on honesty
Yam et al. (2014)	SSC	Ego depletion decreases
		honesty, but only when there is low social agreement that
		the dishonest action is
		immoral
Chiuo et al. (2017)	SSC only males	Ego depletion decreases
	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	honesty
Yam (2018)	SSC	Ego depletion decreases
		honesty
Neurostimulation studies		
Maréchal et al. (2017)	DR	DLPFC stimulation
		increases honesty, but only
		when lying would benefit
		the decision-maker and not
		when it would benefit a
G 1 (2212)	an-	third party
Gao et al. (2018)	SR	DLPFC stimulation
		increases honesty, but only
		for females

Note: SR stands for sender-receiver game; DR stands for die-rolling task; SSC stands for self-serving cheating task.

Table 4

List of empirical studies exploring the cognitive basis of negative reciprocity.

	Measure	Main result
	2P-punishment	
Time constraint studies	•	
Sutter et al. (2003)	UG responder	Pressure <i>increases</i> rejection of low offers, compared to Delay
Cappelletti et al. (2011)	UG responder	Pressure <i>increases</i> rejection of low offers, compared to Delay
Grimm & Mengel (2011)	UG responder	Delay <i>decreases</i> rejection of low offers, compared to Baseline
Wang et al. (2011)	UG responder	Delay <i>decreases</i> rejection of low offers, compared to Baseline
Neo et al. (2013)	UG responder	Delay <i>decreases</i> rejection of low offers, compared to Baseline
Ferguson et al. (2014)	UG responder	Delay, compared to baseline, <i>decreases</i> rejection rates of mildly unfair offers (60:40), leaving the rejection rates of unfairer offers unaffected
Oechssler et al. (2015)	UG responder	Delay, compared to baseline, <i>decreases</i> rejections rates when offers are made using lottery tickets and not when using cash
Balafoutas & Tarek (2018)	IG responder	Pressure <i>increases</i> rejection rates
Conceptual prime studies		
Hochman et al. (2015)	UG responder	Intuition <i>decreases</i> rejection rates, compared to Deliberation
Cognitive load studies		
Cappelletti et al. (2011)	UG responder	Load has <i>no effect</i> on rejection of low offers
Duffy & Smith (2014)	IPG	Load <i>decreases</i> speed of convergence to Nash equilibrium

	Measure	Main result
Olschewski et al. (2018)	UG responder	Load has <i>no effect</i> on
		rejection of low offers
Ego depletion studies		
Anderson & Dickinson (2010)	UG responder	Sleep restriction increases
		minimum acceptable offers
Crockett et al. (2008)	UG responder	Serotonin depletion
		increases rejection rates of
		unfair offers
Crockett et al. (2010a)	UG responder	Serotonin depletion
		increases rejection rates of
C 1 1 (20101)	IIC 1	unfair offers
Crockett et al. (2010b)	UG responder	Enhancing serotonin
		decreases rejection rates of unfair offers
Hololi et al. (2014)	IIC responder	
Halali et al. (2014)	UG responder	Ego depletion <i>increases</i> rejection of low offers
Morewedge et al. (2014)	UG responder	Alcohol intoxication
Morewedge et al. (2014)	OG responder	increases rejection of low
		offers
Liu et al. (2015)	UG responder	Ego depletion increases
Eta et al. (2013)	o d responder	rejection of low offers
Achtziger et al. (2016)	UG responder	Ego depletion decreases
	1	rejection of low offers
Clark & Dickinson (2017)	PGG with punishment	Sleep restriction has <i>no</i>
	-	effect on punishment
Achtziger et al. (2018)	UG responder	Ego depletion has no effect
		on rejection of low offers
Neurostimulation studies		
Knoch et al. (2006)	UG responder	Disruption of rDLPFC
		decreases rejection of low
		offers, compared to
		disruption of IDLPFC
Knoch et al. (2008)	UG responder	Disruption of rDLPFC
		decreases rejection of low
Devento and the angle of -1 (2011)	IIC 1 -	offers, compared to sham
Baumgartner et al. (2011)	UG responder	Disruption of rDLPFC
		decreases rejection of low
		offers, compared to
2 raspansa paradiam studias		disruption of lDLPFC
2-response paradigm studies Bago et al. (2019)	UG responder	Deliberation has no effect on
Dago et al. (2019)	OO responder	Deliberation has <i>no effect</i> on rejection rates: most
		rejection choices under
		deliberation are already
		acitocianon are aneady

	Measure	Main result
		rejection choices under
		intuition
	3P-punishment	
Time constraint studies		
Wang et al. (2011)	TG with punishment	Pressure <i>increases</i> punishment compared to Delay
Artavia-Mora et al. (2017)	Field experiment	Pressure has <i>no effect</i> on punishment, compared to Delay
Cognitive load studies		•
Yudkin et al. (2016)	SG with punishment	Load makes people punish outgroup members more severely than ingroup members
Ego depletion studies		
Liu et al. (2015)	UG with punishment	Ego depletion <i>increases</i> punishment of unfair offers
Neurostimulation studies		•
Buckholtz et al. (2015)	Punishment norm violators in vignettes	Disruption DLPFC decreases punishment, compared to sham
	Anti-social punishment	
Conceptual prime studies		
Pfattheicher et al. (2017)	PGG with punishment	Intuition <i>increases</i> antisocial punishment, compared to Baseline. Inhibiting Intuition <i>decreases</i> antisocial punishment, compared to Baseline.

Note: UG stands for ultimatum game; IG stands for impunity game; IPG stands for iterated prisoner's dilemma; SG stands for steal game.

Table 5

List of empirical studies exploring the cognitive basis of positive reciprocity.

	Measure	Main result
Time constraint studies		
Neo et al. (2013)	TG investee	Delay has <i>no effect</i> on return rates, compared to Baseline
Cabrales et al. (2017)	TG investee	Delay increases return rates
Conceptual primes studies		
Urbig et al. (2016)	Sequential PGG	Second language <i>decreases</i> reciprocation of high contributions
Ego depletion studies		
Halali et al. (2014)	TG investee	Ego depletion <i>increases</i> return rates
Dickinson & McElroy (2017)	TG investee	Sleep restriction <i>decreases</i> return rates
Rantapuska et al. (2017)	TG investee	Hunger has <i>no effect</i> on return rates
Neurostimulation studies		
Knoch et al. (2009)	TG investee	Disrupting rLPFC <i>decreases</i> return rates

Note: TG stands for trust game; PGG stands for public goods game.

Table 6

List of empirical studies exploring the cognitive basis of expectations of negative reciprocity.

	Measure	Main result
Time constraint studies		
Cappelletti et al. (2011)	UG proposer	High pressure (15s) increases offers, compared to Low pressure (180s)
Ego depletion studies		
Halali et al. (2013)	UG proposer	Depletion <i>increases</i> proportion of fair offers
Morewedge et al. (2014)	UG proposer	Alcohol intoxication has <i>no effect</i> on offers
Achtziger et al. (2016)	UG proposer	Depletion increases offers
Clark & Dickinson (2017)	PGG with punishment	Sleep restriction weakly increases contributions
Achtziger et al. (2018)	UG proposer	Depletion decreases offers
Neurostimulation studies		
Ruff et al. (2013)	DG with punishment	Disrupting LPFC decreases donations; Stimulating LPFC increases donations
Strang et al. (2015)	DG with punishment	Disrupting rDLPFC decreases donations, compared to sham.

Note: UG stands for ultimatum game; PGG stands for public goods game; DG stands for dictator game.

Table 7

List of empirical studies exploring the cognitive basis of expectations of positive reciprocity.

	Measure	Main result
Time constraint studies	•	
Neo et al. (2013)	TG investor	Delay has <i>no effect</i> on trust
Jaeger et al. (2019)	TG investor	Pressure <i>increases</i> trust
•		compared to Delay
Ego depletion studies		
Anderson & Dickinson (2010)	TG investor	Sleep deprivation decreases
		trust
Evans et al. (2011)	TG investor	Ego depletion has no effect
		on trust, but it makes people
		more likely to follow the
		default
Baumeister et al. (2014)	TG investor	Ego depletion decreases
		trust
Dickinson & McElroy (2017)	TG investor	Sleep deprivation decreases
D 1 (2015)	TO 1	trust
Rantapuska et al. (2017)	TG investor	Hunger has no effect on
		trust
Cognitive load studies		
Bonnefon et al. (2013)	TG investor	Load has <i>no effect</i> on the
		capacity to detect
		trustworthiness
Samson & Kostyszyn (2015)	TG investor	Load decreases trust
Conceptual prime studies		
Urbig et al. (2016)	First mover in	Second language has no
	sequential PGG	effect on trust

Note: TG stands for trust game; PGG stands for public goods game.

Table 8

List of empirical studies exploring the cognitive basis of deontology vs. act utilitarianism.

	Measure	Main result
Time constraint studies	<u> </u>	
Suter & Hertwig (2011)	SD	Pressure increases
Cummins & Cummins (2012)	SD	deontological choices Pressure <i>increases</i> deontological choices
Trémolière & Bonnefon (2014)	SD varying save/kill	Pressure <i>increases</i> deontological choices for save/kill = 5 but not save/kill = 500
Cognitive load studies		
Greene et al. (2008)	SD	Load increases response
Trémolière & Bonnefon (2014)	SD varying save/kill	time of utilitarian choices Load <i>increases</i> deontological choices for save/kill = 5 but not save/kill = 500
Conway and Gawronski (2013)	SD process dissociation	Load <i>reduces</i> utilitarian judgments, while leaving deontological judgments unaffected
Białek & De Neys (2017).	SD	Load reduces utilitarian
Hayakawa et al. (2017)	SD process dissociation	judgments Second language decreases deontological judgments, but has no effect on utilitarian judgments
Li et al. (2018)	SD process dissociation	Load <i>reduces</i> utilitarian judgments, while leaving deontological judgments unaffected
Muda et al. (2018)	SD process dissociation	Second language <i>decreases</i> both deontological and utilitarian judgments
Conceptual primes studies		
Valdesolo & DeSteno (2006)	SD	Affect induction increases
Paxton et al. (2012)	SD	deontological judgments Deliberation induction increases utilitarian
Costa et al. (2014)	SD	judgments Second language <i>increases</i> utilitarian judgments

	Measure	Main result
Kvaran et al. (2013)	SD	Emotional prime <i>increases</i> deontological judgments; analytical prime <i>increases</i> utilitarian judgments
Geipel et al. (2015a)	SD	Second language <i>increases</i> utilitarian judgments
Cipolletti et al. (2016)	SD	Second language <i>increases</i> utilitarian judgments
Corey et al. (2017)	SD	Second language <i>increases</i> utilitarian judgments
Spears et al. (2018)	SD	Attention prime <i>increases</i> utilitarian judgments
Capraro et al. (2019)	OUS	Intuition <i>increases</i> non- utilitarian judgments in the instrumental harm dimension; intuition has <i>no</i> <i>effect</i> on beneficence dimension
Ego depletion studies		
Trémolière et al. (2012)	SD	Depletion <i>increases</i> deontological judgments
Timmons & Byrne (2018)	SD	Depletion <i>increases</i> deontological judgments
2-response paradigm studies	SD	
Bago & De Neys (2018)		Most utilitarian deliberative judgments were already utilitarian under intuition

Note: SD stands for sacrificial dilemmas game; OUS stands for Oxford Utilitarianism Scale.

Table 9

Are the predictions of the General Social Heuristics Hypothesis (GSHH) consistent with the empirical data?

Social decision	Are the predictions of the GSHH consistent with the data?
Cooperation	Yes
Altruism	Yes
Honesty (when lying harms concrete others)	Yes
Honesty (when lying harms abstract others)	No
Negative reciprocity	Yes
Expectations of negative reciprocity	Yes
Positive reciprocity	Inconclusive – but there are only seven studies
Expectations of positive reciprocity	Inconclusive – but results trend in the opposite direction than predicted
Instrumental harm	Yes – but one has to add one more "realistic" assumption
Impartial beneficence	Data consistent with the predictions – but there is only one study