The Social Neuroscience of Cooperation

Julian A. Wills
New York University

Leor Hackel
Stanford University

Oriel FeldmanHall
Brown University

Philip Pärnamets
New York University

Jay J. Van Bavel
New York University

Word count: 4,846 (body and figure caption)

Acknowledgments and author's note: This chapter was partially funded by a grant from the US National Science Foundation to JVB (award #1349089) and from the Swedish Research Council to PP (2016-06793).

Please direct correspondence to:
Jay J. Van Bavel
New York University
Department of Psychology
New York, NY, USA. 10012
jay.vanbavel@nyu.edu

The Social Neuroscience of Cooperation

Cooperation occurs at all stages of human life and is necessary for large-scale societies to emerge and thrive: when individuals prioritize themselves over their community, the consequences can damage social communities, scientific institutions, and our planet. Hence, understanding the psychological and neural underpinnings of cooperative behavior is an important goal for social and cognitive neuroscience. Yet, extensive research devoted to the mental processes underlying human pro-sociality have failed to produce a satisfying framework for understanding how the selfish and pro-social impulses unfold in the human brain.

For centuries, philosophers have debated whether prosocial tendencies are rooted in institutions that regulate our selfish impulses (Hobbes, 1650) or emerge through natural intuitions (Rousseau, 1754). These ancient philosophical debates about human nature remain unresolved. Contemporary scientists continue to grapple with the origins of human pro-sociality. One on hand, models of *prosocial restraint* assert that the better angels of our nature stem from deliberate restraint of selfish impulses (DeWall, Baumeister, Gailliot, & Maner, 2008; Kocher, Martinsson, Myrseth, & Wollbrant, 2012; Stevens & Hauser, 2004), whereas models of *prosocial intuition* argue that cooperation stems from intuition and is only corrupted by deliberate attempts to maximize self-interest (Rand, 2016; Rand, Greene, & Nowak, 2012). In this chapter, we bridge cognitive neuroscience, neuroeconomics, and social psychology to examine the issue of human pro-sociality and cooperation.

In the first section, we review literature from several fields to describe common experimental tasks used to measure human cooperation. In the second section, we review dominant theoretical models that have been used to characterize cooperative decision-making, as well as brain regions implicated in cooperation. Building on work in neuroeconomics, we suggest a value-based account may provide the most powerful understanding the psychology and neuroscience of group cooperation. In the third and fourth sections we review the role of individual differences and social context in shaping the mental processes that underlie cooperation. Finally, we consider gaps in the literature and offer directions for future research on the cognitive neuroscience of cooperation. We suggest that this multi-level approach provides a more comprehensive understanding of the mental and neural processes that underlie the decision to cooperate with others.

Measuring Cooperation

Cooperation involves any action where one individual incurs a cost in order to benefit others (Rand & Nowak, 2013). These cost and benefits can range from primary reinforcers (e.g., food, drugs, sex) to secondary reinforcers (e.g., wealth, status, publications). Critically, cooperative acts are not always selfless; sometimes we help others at a cost to obtain rewards in the future. When tipping a bartender, for instance, you may be motivated to not only reward their attentive service, but to continue receiving excellent service in the future. For this reason, some researchers distinguish between pure or *altruistic cooperation* (i.e., when current or future rewards are ignored) and *strategic cooperation* (i.e., when future rewards motivate the cooperative act) (Camerer & Fehr, 2004; Gintis, 2014). Cooperative acts can be pure, strategic,

or a mixture of both. As a result, researchers go to great lengths to disambiguate these motives (Camerer & Fehr, 2004). To better understand the motives that underlie cooperation and how they are studied, we briefly review four measures of cooperation.

Social dilemmas

The most common approach to studying cooperation involves the use of social dilemmas and perhaps the most widely used measure of cooperation is the *Prisoner's Dilemma* (PD) game[1]. In the PD, there are two players who are each given the choice to either defect (D) or cooperate (C). This game has been popularized on the British game show "Golden Balls" because it creates a tension in which the fates of two players are tied together. In the standard, symmetric version of the game, both players receive payoff R(eward) if both choose C, payoff P(unishment) if both choose D, and payoffs T(temptation) or S(ucker) if one defects and the other cooperates, respectively. Thus, the hierarchical payoff structure is $T > R > P > S$. As in the legal system, there is a strong temptation not to be a sucker.

In the *Prisoner's Dilemma*, each player can maximize their individual profit by choosing D, regardless of what the other player chooses. In other words, outcome DD is the unique Nash equilibrium of the game and the prediction for fully rational and selfish players. However, the the cooperative outcome, CC, maximizes their collective profit. This feature, that the players are always worse off if both of them defect compared to cooperate—but each is individually better off by defecting—is what makes the PD a social dilemma (Dawes, 1980; Van Lange, Joireman,

[1] Invented in 1950 by Merrill Flood and Melvin Dresher, while working at the Rand Corporation (no known relation to Dave Rand, who is cited throughout this chapter) as part of research investigating the use of game theory to inform nuclear strategy.

Parks, & Van Dijk, 2013). Pitting self-interest against collective interest captures the dynamic at play in countless real-world social decisions, from negotiating nuclear arms agreements to sharing research ideas.

Since decisions are typically made simultaneously, anonymous one-shot PDs (i.e. one round only) are used to measure pure cooperation in both players. In contrast, the iterated PD, in which players play multiple rounds with one another, measures strategic cooperation since players' decisions may impact expectations for subsequent choices. In addition, people cooperate strategically when their choices are made public and players can select partners known to be cooperative (Barclay & Willer, 2007; Feinberg et al., 2014). Despite understanding that defecting is in one's best self-interest, decades of evidence from both iterated and one-shot versions of the PD reveal that people willingly cooperate—even with complete strangers.

To understand cooperation in groups with more than two players, researchers employ the *Public Goods Game* (PGG). In this game, players choose between contributing their endowment to a collective pool (i.e., maximizing joint payoffs) or free-riding, in which they keep their own endowment while also reaping the benefits of others' contributions (i.e., maximizing individual payoffs in the short-term). The PGG has a similar incentive structure to the PD and is sometimes suggested to be a generalization of it (Rand & Nowak, 2013). The PGG inherits many properties of the PD (e.g., anonymous one-shot games index pure cooperation), since contributing and free-riding are group-based analogues of cooperating and defecting. Similar to the findings in the PD, evidence reveals that in typical variants of the PGG, people donate on average 60% of the trials. However, because the PGG also inherits properties of group psychology, important differences

can emerge (Dawes, 1980). For instance, contributions in iterated PGGs routinely diminish over time (Andreoni, 1988), where those in the PD do not. This may be due to the diffusion of responsibility or absence of direct reciprocity in the PGG, where punishing one free-rider equally penalizes the entire group. PGGs may also be particularly sensitive to other aspects of group psychology, such as norms concerning promise-keeping (Bicchieri, 2002) and social identity (Kramer & Brewer, 1984). Furthermore, the PGG likely provides superior ecological validity to the PD since the most pressing real-world cooperative dilemmas, like climate change or science reform, involve more than two people (Camerer, 2011).

Social dilemmas sometimes include additional dimensions, such as introducing reinforcement or punishment opportunities[2] (Fehr & Gächter, 2002; Kelley, 2003), reputational concerns (Milinski, Semmann, & Krambeck, 2002), or manipulating the framing of the game (Van Lange et al., 2013). For instance, framing a social dilemma as a "community game" can double rates of cooperation compared to when it is framed as a "Wall Street game", likely due to activating norms associated with those contexts (Liberman, Samuels & Ross, 2004). Moreover, introducing opportunities for reward and punishment almost always boosts contributions (Dreber, Rand, Fudenberg, & Nowak, 2008; Fehr & Gächter, 2002; Andreoni, Harbaugh, & Vesterlund, 2002). These factors appear to alter the value people place on the decision to cooperate.

Bargaining games

[2] This manipulation also provides an opportunity to observe costly punishment.

Another measure of cooperation comes from bargaining games where responsiveness to fairness norms can be assessed. In the *Ultimatum Game* (UG; Güth, Schmittberger & Schwarze, 1982), two players take the role of either proposer or responder. The Proposer is given some endowment *E* and must offer the responder some amount *O* (which may be zero). The Responder can either accept or reject the offer. If the offer is accepted, the responder receives *O* and the proposer keeps the remainder (*E minus O)*. If the offer is rejected, neither player receives anything. From an economically rational standpoint, responders should accept any offer, since some money is better than no money. However, it has been repeatedly observed across cultures that responders will reject offers that are considered unfair according to local norms (Camerer & Fehr, 2004; Henrich et al., 2005), which is typically anything below 20% of the endowment. By rejecting the offer, people are signaling their willingness to forgo their own profit to punish a transgressor who violated fairness norms—harming both parties. Thus, a degree of cooperation is normally required to ensure a fair offer is accepted.[3]

To capture pure pro-sociality, a modified UG is used in which the responder is not given the option to reject the proposer's offer (Kahneman, Knetsch & Thaler, 1986)—known as the *Dictator Game* (DG). In this game, the experimenter endows a sum of money to the dictator, who can then decide how much to give to the receiver. True to its name, the receiver has no bargaining power in the DG and has no choice but to accept the initial offer from the dictator. Surprisingly, dictators nevertheless give almost 30% of the pie in these one-sided games,

[3] This can be considered a departure from the strict definition of cooperation we introduced above. However, we include it here for completeness since this class of games is used to study prosociality.

revealing just how altruistic people can be (Engel, 2010). This is the case even when the experimenter ensures complete anonymity between the two players, providing a measure of pure pro-sociality for the dictator since there is no opportunity to reciprocate or punish an unfair split. These games provide some evidence for the tendency for humans to cooperate under a wide variety of conditions.

Models of Cooperation

Models of prosocial behavior make assumptions about the underlying mental computations that guide people towards self-interest or cooperation. In this following section, we contrast three such models of cooperation, the first two are based on a dual-process account that cast intuitive and deliberative processes as competing for control in cooperative behavior. The third offers a single-process framework from neuroeconomics that emphasizes the role of valuation circuits. We briefly review each approach and argue that social and cognitive neuroscience might prove fruitful for arbitrating between these different models.

Intuition vs. Deliberation

One of the most ubiquitous frameworks in psychology is the dual-process model, which posit that the mind can be carved into two core systems: *intuition* (i.e., fast, automatic, and unconscious processes) and *deliberation* (i.e., slow, controlled, and rational processes) (Chaiken & Trope, 1989; Evans & Stanovich, 2013; Kahneman, 2011). Research in social neuroscience has attempted to map neural systems onto intuition and deliberation (Cohen, 2005; Satpute & Lieberman, 2006). For instance, patients with ventromedial Prefrontal Cortex (vmPFC) or

amygdala damage presented with blunted affective processing (Bechara, 2000), whereas damage to the dorsolateral Prefrontal Cortex (dlPFC) impaired deliberative processes, like working memory, reasoning, and self-regulation (Barbey, Koenigs, & Grafman, 2013). The dissociations between these systems has been seen by several scholars as further evidence for dual process models. In psychology, these models have been used to explain a wide range of phenomena including phenomena, including stereotypes (Devine, 1989), persuasion (Chaiken, 1987), and moral judgment (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). More recently, competing dual process models of cooperation have proven reminiscent of old philosophical debates regarding humanity's intrinsic benevolence (Rousseau, 1754) versus the need for institutions to restrain our greedy impulses (Hobbes, 1650).

The most prominent dual process models of cooperation have argued that pro-social decisions stem primarily from intuition (Rand et al., 2014; Zaki & Mitchell, 2013). For instance, the *Social Heuristics Hypothesis* (Rand et al., 2014) makes three core assumptions: (1) rational self-interested agents should never cooperate in anonymous one-shot games, (2) cooperation stems from error-prone intuitions whereas self-interest stems from more corrective deliberation, and (3) experimentally boosting reliance on intuition (vs. deliberation) should only result in increased or static cooperation. In their words, "deliberation only ever reduces cooperation in social dilemmas…or has no effect...but never increases social-dilemma cooperation" (Bear, Kagan, & Rand, 2017). According to this view, cooperation is frequently rational—but people develop error-prone heuristics to cooperate even when it would be irrational.

Support for the Social Heuristics Hypothesis comes from a mix of behavioral and neural evidence. The most important behavioral evidence comes from experiments showing that people are slower to make self-interested choices compared to cooperative choices in both the one-shot PD and PGG (Rand et al., 2012). Moreover, putting people under time-pressure increases cooperation rates (Rand et al., 2012; Everett, Ingbretsen, Cushman, & Cikara, 2017). However, a recent international replication effort came up with mixed support for this key finding, suggesting that the behavioral evidence in support of the Social Heuristic Hypothesis may be weaker than previously thought (Bouwmeester et al., 2017; but see also Rand, 2017; Everett et al., 2017). Recent fMRI studies find that greater dlPFC activity was associated with decisions that prioritize selfish gain over another's pain (FeldmanHall et al., 2013), while reduced dlPFC functional activity and volume was associated with more generosity in a dictator game, which together suggest a link between deliberation and self-interest (Fermin et al., 2016; Yamagishi et al., 2016). Those findings are in line with dual-process models in general, and the Social Heuristic Hypothesis in particular.

This perspective has proven particularly provocative and controversial because it contrasts with more traditional *prosocial restraint* models, whereby cooperation primarily stems from deliberate restraint of our selfish impulses (Achtziger, Alós-Ferrer, & Wagner, 2015; Lohse, 2016; Martinsson, Myrseth, & Wollbrant, 2012). That is, some argue that humans' unique capacity for self-reflection (i.e., compared to other primates) provides a critical avenue to promote prosocial behavior (Stevens & Hauser, 2004). Moreover, prosocial restraint models are supported by evidence that depleting cognitive resources impairs helping behavior (DeWall,

Baumeister, Gailliot, & Maner, 2008) and amplifies dishonesty (Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009; but see Saraiva & Marshall, 2015). We recently found that patients with damage to the dlPFC showed impaired cooperation—and reductions in cooperation scaled with the scope of damage in this region (Wills et al., 2017). We found no such decrements for patients with damage to the vmPFC, amygdala or other brain damaged control patients. One limitation of this research area is that several pre-registered attempts to replicate ego-depletion effects have found null or very small effect sizes—calling many findings in this literature in question. As such, the evidence behind these models has proven unconvincing to opposing camps.

A value-based approach to cooperation

A central approach to neuroeconomics has examined how value is represented in the human brain and used to guide decision-making. Instead of conceptualizing cooperation as arising from distinct, competing psychological systems, we argue that cooperation, and social preferences in general, should be situated within such a value-based decision framework. Central to this framework is the assumption, found in most economic and psychological theories of choice, that prior to deciding between one or several alternatives, an organism determines the subjective value of each alternative. Subjective value allows comparisons between complex and qualitatively different alternatives by placing them on a common scale (Rangel, Camerer & Montague, 2008; Levy & Glimcher, 2012; Bartra, McGuire & Kable, 2013). Moreover, this approach allows for individual differences and contextual factors to shape the value of these

alternatives. We provide an overview of this perspective, the underlying neural system involved in value computations, and how this might be applied fruitfully to the study of cooperation.

The field of neuroeconomics has been focused on understanding how the brain computes the value of alternative actions during decisions, such as when they are forced to decide between engaging in self-interest or cooperation. A consistent finding in the decision-making literature across topics has been that brain activation in the orbitofrontal cortex or vmPFC, ventral striatum (VS), and posterior cingulate cortex increase with subjective value during choice tasks and while receiving monetary, primary, or social rewards (Levy & Glimcher, 2012; Bartra, McGuire & Kable, 2013). This has been taken as evidence that representations of value are computed in these regions and used as a common currency to make decisions between different options (Levy & Glimcher, 2012; Grabenhorst & Rolls, 2011).

Recent studies suggest that a value-based framework better explains human cooperation than either dual-process account mentioned above. Prosocial intuition models argue that intuitive responses are shorter than deliberative ones. But from the perspective of value-based frameworks, response times are a function of the discriminability of alternatives: people make faster choices when deciding between very different values as opposed to similar values (Krajbich, Armel & Rangel, 2010; Shadlen & Kiani, 2013). Thus, these models make competing predictions about cooperation. In one such experiment, participants played multiple PGGs with varying returns on money contributed (Krajbich, Bartling, Hare, & Fehr, 2015). In one condition, for each monetary unit contributed, each player would receive 50% back. In the other conditions

the multipliers were 30% (rewarding selfishness) and 90% (rewarding cooperation)[4]. Consistent with the value-based approach, the relationship between reaction time and cooperation was determined by the reward structure: cooperation was fast when it was rewarded, and selfishness was fast when it was rewarded. In other words, cooperation decisions were fastest when the reward structure made the alternatives clear. These findings also highlight why researchers should be cautious when interpreting reaction time differences as evidence for intuition or deliberation.

A growing body of work in cognitive neuroscience also supports the value-based account of cooperation. Specifically, several studies have found that vmPFC activation relates to value-based quantities during cooperative decisions (FeldmanHall, Dalgeish, Evans & Mobbs, 2015; Hutcherson, Bushong, & Rangel, 2015; Zaki, Lopez, & Mitchell, 2014). During altruistic decision-making, for instance, the brain forms an overall value signal as a weighted sum of two quantities: the payoffs available for oneself and to a recipient (Hutcherson, Bushong, & Rangel, 2015). Both quantities were associated with activation in vmPFC during people's choices, supporting the idea that vmPFC encodes the overall value of prosocial choices.

The notion that the vmPFC encodes the subjective value of cooperation is also supported by findings from a neuroimaging study conducted while people engaged in the PGG (Wills, Hackel & Van Bavel, 2018). We found that vmPFC activity was greater when participants made choices aligned with their overall social preferences (i.e., when cooperative players made the

[4] Recall that in a PGG a player is *always* better off keeping their money rather than cooperating, in other words the multiplier per monetary unit and player is always strictly less than 1.

decision to cooperate and selfish players made the decision to act selfishly). In contrast, dlPFC activity was associated with choices that went against players' social preferences. Moreover, there was increased connectivity between vmPFC and dlPFC when people made cooperative decisions that violated social norms. In these cases, the dlPFC may be needed to integrate value signals computed in the vmPFC (Domenech, Redoute, Koechling & Dreher, 2017), as value-related signals in dlPFC activate after those in vmPFC (Sokol-Hessner, Hutcherson, Hare, & Rangel, 2012). Clarifying the connectivity between regions will likely be key to further arbitrate between the value-based model and competing frameworks (see Figure 1).

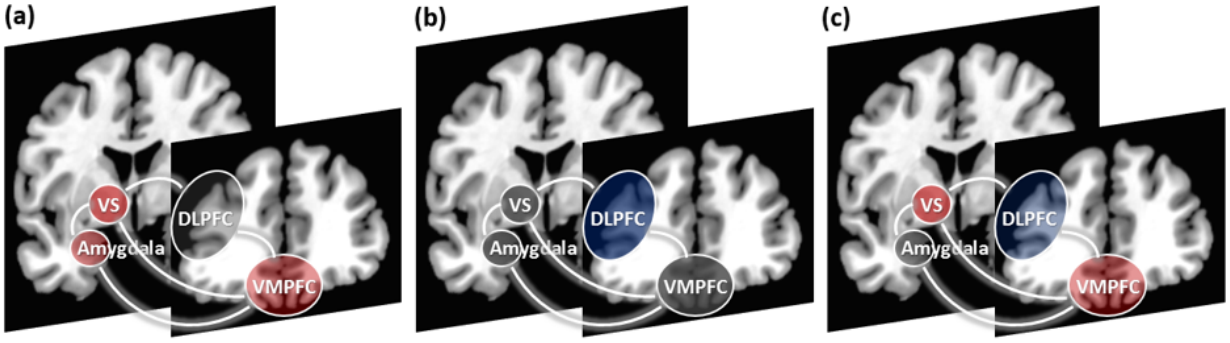

Figure 1. Candidate neural systems of cooperative decision-making. Dual-process models of prosocial behavior predict cooperation stems from either (a) neural regions involved in intuition (red) or (b) neural regions involved in deliberation (blue). On the other hand, (c) value-based models predict cooperation should stem from regions typically recruited during decision making (red), as well as heightened connectivity between dlPFC (blue) and vmPFC for decisions that require more effort. VS = ventral striatum; vmPFC = ventromedial prefrontal cortex; dlPFC = dorsolateral prefrontal cortex. Graphics adapted from (Phelps, Lempert, & Sokol-Hessner, 2014).

There is growing research into the various psychological factors that modulate (i.e., suppress or amplify) value. After all, when constructing interventions to promote cooperation, it is vital to understand *when* and *for whom* cooperation is valued. For instance, interventions

designed to block "deliberative self-interest" could fail—or even backfire—among those who do not intrinsically value cooperation and need to deliberate longer to fully consider the potential value of cooperation. Similarly, while efforts to deter "intuitive self-interest" could prevail under some circumstances, these same interventions might also reduce cooperation under contexts where cooperation is strongly valued. Here we review two broad classes of these potential value modulators: (1) contextual factors and (2) individual differences.

Contextual Factors

Several contextual factors can influence cooperative decision-making by shaping social value. For instance, group norms have been known to boost compliance in perceptual judgments (Asch, 1951) and prosocial behavior (Cialdini, Reno, & Kallgren, 1990; Nook, Ong, Morelli, Mitchell, & Zaki, 2016). Evidence for cognitive neuroscience suggests that group norms also modulate the neural substrates of subjective value (Nook & Zaki, 2015)(Wills et al., 2018) as well as systems implicated in conflict monitoring (Chang & Sanfey, 2013) and control (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Richeson et al., 2003). For example, disrupting the dlPFC has been shown to disrupt participants' ability to act in accordance with fairness norms and reject unfair offers in ultimatum games (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006). Notably, participants still reported accurate valuations of the offers, suggesting a role of the dlPFC in integrating the outputs from valuation circuits.

Social psychologists distinguish between descriptive norms (i.e., how do others typically behave?) and injunctive norms (i.e., how should others behave?). Since there is strong evidence that descriptive norms influence cooperation (Kopelman, Weber, & Messick, 2002), the same is

likely true for injunctive norms—especially since cooperation is often characterized as a moral imperative. Consider, for instance, an influential finding where framing the PGG as "The Community Game" boosts cooperation significantly more than when it is called "The Wall Street Game" (Liberman, Samuels, & Ross, 2004). Even when other players were expected to be selfish, those assigned to the community condition decided to cooperate nonetheless, suggesting that injunctive norms can bias moral behavior. In our view, norms may help account for the variance in intuitive cooperation observed across international samples (Bouwmeester et al., 2017)—contexts where cooperation is normative may increase the value placed on cooperation.

Social identity—a person's sense of who they are based on their group membership—is another core social psychological construct that drives cooperation and conflict (Tajfel & Turner, 2001). For instance, cooperative decisions can be influenced by existing intergroup conflicts, such as race relations (Kubota, Li, Bar-David, Banaji, & Phelps, 2013) and political partisanship (Iyengar & Westwood, 2015), as well as by artificially created identities (Marcus-Newhall, Miller, Holtz, & Brewer, 1993). Social identity may drive cooperation because it connotes interdependence: people assume in-group members will reciprocate with one another (Yamigishi, 1992). There is also reason to believe that identity can change the value people place on in-group members and their outcomes. For example, one study found greater activation in the ventral striatum when participants observed in-group members receive rewards compared to out-group members, but only for participants who heavily identified with the in-group (Hackel, Zaki & Van Bavel, 2017). Indeed, simply categorizing faces of in-group members activates neural circuitry associated with valuation, including the amygdala, orbitofrontal cortex and dorsal

striatum (Van Bavel, Packer & Cunningham, 2008). Thus, generating a shared group identity can induce cooperation by imbuing in-group members with value or increasing the expectations of future reward due to reciprocity.

Individual Differences

People differ in their tendency to cooperate, and these preferences tend to be stable over time (Volk, Thöni, & Ruigrok, 2012). Within PGGs, for instance, researchers have estimated that a substantial majority of people are (50 - 55%) of conditional cooperators (i.e., those who only cooperate when others cooperate), a sizable portion (23- 30%) are considered to be consistent free-riders (Fischbacher, Gächter, & Fehr, 2001), and only a small percentage (5 - 10%) fall into the category of consistent contributors who always cooperate (Weber & Murnighan, 2008). Some measures, such as The Social Value Orientation measure, are designed to capture these differences (see Van Lange, 1999; Van Lange et al., 1997). *Pro-selfs* are people who place a high value on their own rewards, whereas *pro-socials* are people who place a high value on collective rewards. Research in the past decade has consistently found that pro-socials are more inclined to cooperate in both one-shot and iterated games (Balliet et al., 2009). Thus, individual differences are robust predictors of cooperative (vs. selfish) behavior.

Critically, individual differences may determine which contextual factors steer cooperative decision-making. Take, for instance, consistent contributors, who are defined by their iconoclastic commitment to cooperating under any circumstance (i.e. even when everyone else in their group is free-riding). There is evidence that the mere presence of these consistent contributors can boost cooperation in others by activating moral identities (Gill, Packer, & Van Bavel, 2013). That is, consistent contributors may provide a contextual cue that predominantly boosts cooperation among individuals who consider generosity and fairness central features of their identity (Packer, Gill, Chu, & Van Bavel, 2018). In addition, there is evidence that experimentally invoking deliberation promotes cooperation, but only for people exhibiting prosocial tendencies (Mischkowski & Glöckner, 2015). Thus, individual differences can also predict which contextual factors are more or less likely to shape cooperative decision-making. More work should examine this interplay using neuroscience methods to better understand how individual differences and context are integrated in the brain during decision-making.

Future directions

Attention. A key element of dynamic value-based cognition is the role of attention. By measuring participants fixations during simple economic choices, researchers have shown that attention to certain options influences decisions (Krajbich et al., 2010). These findings have been shown to also hold for more complicated value-based choices, such as moral ones (Pärnamets, Balkenius & Richardson, 2014). By tracking participants' fixations and prompting them to make a choice only after sufficiently fixating on one option, researchers were even able to influence what choice participants made (Pärnamets et al., 2015). Moreover, one study found that value

signals in the striatum and vmPFC were modulated by the relative value of fixated versus non-fixated food options (Lim, O'Doherty & Rangel, 2011). Thus, visual attention influences valuation and alters pro-social behavior. In our view, integrating measures of attention and other sensory information into models of cooperative decision-making offers significant opportunities for understanding more about the underlying mental processes and potentially even designing effective interventions for increasing cooperation.

Learning. A key element of value-based models is that people learn the value of different actions over time, whether through personal experience (Daw et al., 2011, FeldmanHall, Otto & Phelps, 2018), social observation (Dunne, D'Souza, & O'Dohrety, 2016; Haaker et al 2017; Lindström et al), or instruction (Atlas et al., 2016, Behrens et al 2008). People may learn to value cooperation with specific partners, groups, and social contexts (Apps et al 2017). Understanding this process may offer new insights into how people choose to cooperate.

Canonical models of reciprocity suggest that people form impressions of others' generosity and tend to help those viewed as generous (Wedekind & Milinski, 2000). However, models of value learning in neuroscience suggest another route by which people may learn to cooperate with others. During cooperative interactions, people experience *reward value*—that is, the material benefits of the interaction. When receiving money from an interaction partner, people engage not only neural regions associated with forming social impressions, but also neural regions associated with reward learning (e.g., ventral striatum; Hackel, Doll, & Amodio, 2015). As a result, people learn to reciprocate not only with givers who frequently display generosity, but also with givers who have greater wealth and thus provide larger rewards (Hackel

& Zaki, 2018). Modeling how experience and feedback is integrated into value to guide future decisions is a key to fully understanding cooperation. Although the evidence is currently sparse, value learning likely plays a similar role in shaping whether people contribute to collective goods in social dilemmas.

Conclusion

Unlocking the secret to group cooperation is critical for solving social dilemmas ranging from climate change to public resource management to improving science. For this reason, the study on cooperation has attracted an enormous amount of attention in recent years. We believe that a value-based approach holds significant promise for understanding how different people in different contexts make cooperative decisions. This approach not only has explanatory power that can generate important directions in learning and attention, but it offers to bridge a number of literatures under a common multi-level framework. This has important implications since models consistent with neural architecture should be privileged over models that are not biologically described (van Ede & Maris, 2016), and theories that provide consistent evidence across multiple levels of analysis are most likely to provide a complete and enduring explanation of behavior (Wilson, 1998). If this approach can harness the collective intelligence of scientists and scholars from philosophy to neuroscience, it will allow them to cooperate on solving a longstanding scientific debate as well as some of the most pressing problems facing humanity.

References

Andreoni, J. (1988). Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics, 37*, 291-304.

Andreoni, J., Harbaugh, W. T, & Vesterlund, L. (2002). The carrot or the stick: Rewards, punishments and cooperation. *University of Oregon Department of Economics Working Paper*(2002-1).

Apps, M.A.J. & Sallet, J. (2017). Social learning in the medial prefrontal cortex. *Trends in Cognitive Science, 21,* 151-152.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men*, 222-236.

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex*, *49*, 1195-1205.

Barclay, P. & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London B: Biological Sciences, 274,* 749-753

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.

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

Bear, A., Kagan, A., & Rand, D. G. (2017). Co-evolution of cooperation and cognition: the impact of imperfect deliberation and context-sensitive intuition. *Proc. Biol Sci, 284*, 20162326.

Bechara, A. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*, 295-307.

Bicchieri, C. (2002). Covenants without swords: Group identity, norms, and communication in social dilemmas. *Rationality and Society*, *14*, 192-228.

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., ... & Evans, A. M. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*, 527-542.

Camerer, C. F., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Ginti (Eds.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* (pp. 55-95). Oxford University Press.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*, 277-284.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology.* New York, NY: Guilford Press.

Cohen, J. D. (2005). The vulcanization of the human brain: A neural perspective on interactions between cognition and emotion. *The Journal of Economic Perspectives, 19,* 3-24.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015-1026.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169-193.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5-18.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*, 348-351.

Domenech, P., Redouté, J., Koechlin, E., & Dreher, J. C. (2017). The neuro-computational architecture of value-based selection in the human brain. *Cerebral Cortex*, *28*, 585-601.

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

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*, 223-241.

Everett, J. A., Ingbretsen, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior—Even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology*, *73*, 76-81.

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

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive Affective Neuroscience, 7,* 743-751.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage*, *105*, 347-356.

FeldmanHall, O., Son, J., & Heffner, J. (2018). Norms and the Flexibility of Moral Action. *Personality Neuroscience, 1,* 1-14.

FeldmanHall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology: General, 147,* 1211-1224.

Fermin, A. S., Sakagami, M., Kiyonari, T., Li, Y., Matsumoto, Y., & Yamagishi, T. (2016). Representation of economic preferences in the structure and function of the amygdala and prefrontal cortex. *Scientific Reports*, *6*, 20982.

Feinberg, M., &Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25,* 656-664

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, *71*, 397-404.

Gill, M. J., Packer, D. J., & Van Bavel, J. (2013). More to morality than mutualism: Consistent contributors exist and they can inspire costly generosity in others. *Behavioral and Brain Sciences*, *36*, 90-90.

Gintis, H. (2014). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton University Press.

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, *15*, 56-67.

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.

Gu, X., Wang, X., Hula, A., Wang, S., Xu, S., Lohrenz, T. M., ... & Montague, P. R. (2015). Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: computational and lesion evidence in humans. *Journal of Neuroscience*, *35*, 467-473.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367-388.

Haaker, J., Yi, J., Petrovic, P., & Olsson, A. (2017). Endogenous opioids regulate social threat learning in humans. *Nature Communications, 8,* 15495.

Hackel, L. M., & Zaki, J. (2018). Propagation of economic inequality through reciprocity and reputation. *Psychological Science*, *29*, 604-613.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*, 1233-1235.

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

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Henrich, N. S. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795-815.

Hobbes, Thomas. (1650). Human Nature (1650). *Leviathan (Engl. 1651)*.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*, 451-462.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*, 690-707.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285-S300.

Kelley, H. H. (2003). *An atlas of interpersonal situations*: Cambridge University Press.

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.

Kopelman, S., Weber, J. M., & Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychological research. In E. Ostrom, T. Dietz, N. Dolsak, P. C. Stern, S. Stonich, & E. U. Weber (Eds.) *The drama of the commons*, (pp. 113-156) Wasington DC.: National Academy Press.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*, 1292-1298.

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.

Kramer, R. M., & Brewer, M. B. (1984). Effects of group identity on resource use in a simulated commons dilemma. *Journal of personality and social psychology*, *46*, 1044-1057.

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.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*, 1027-1038.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin*, *30*, 1175-1185.

Lindström, B., Haaker, J., & Olsson, A. (2018). A common neural network differentially mediates direct and social fear learning. *Neuroimage, 167, 121-129.*

Lim, S. L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *Journal of Neuroscience*, *31*, 13214-13223.

Marcus-Newhall, A., Miller, N., Holtz, R., & Brewer, M. B. (1993). Cross-cutting category membership with role assignment: A means of reducing intergroup bias. *British Journal of Social Psychology, 32*, 125-146.

Mischkowski, D., & Glöckner, A. (2016). Spontaneous cooperation for prosocials, but not for proselfs: Social value orientation moderates spontaneous cooperation behavior. *Scientific Reports*, *6*, 21555.

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

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature, 415,* 424-426.

Nook, E. C., & Zaki, J. (2015). Social norms shift behavioral and neural responses to foods. *Journal of Cognitive Neuroscience*, *27*, 1412-1426.

Packer, D. J., Gill, M. J., Chu, K., & Van Bavel, J. J. (2018). *How does a person like me behave? On how consistent contributors can inspire generous giving among people with prosocial values.* Unpublished manuscript.

Pärnamets, P., Balkenius, C. & Richardson, D. C. (2014). Modelling moral choice as a diffusion process dependent on visual fixations. In Bello, P., Guarini, M., McShane, M. & Scassellati, B. (eds.) *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX.

Pärnamets, P., Johansson, P., Balkenius, C., Hall, L., Spivey, M.J. & Richardson, D.C. (2015). Biasing moral choices by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*, 4170-4175.

Phelps, E. A., Lempert, K. M., & Sokol-Hessner, P. (2014). Emotion and decision-making: Multiple modulatory neural circuits. *Annual Review of Neuroscience, 37*, 263-287.

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. (2017) Reflections on the time-pressure cooperation Registered Replication Report. *Perspectives on Psychological Science, 12,* 543-547.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*, 413–425.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*, 427-430.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications* 5, 3677.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience, 9, 545-556.

Rousseau, Jean Jacques. (1754). A Discourse On A Subject Proposed By The Academy Of Dijon: What Is The Origin Of Inequality Among Men, And Is It Authorised By Natural

Law? *Retrieved 23 January 2009, from the Constitution Society in the Rousseau site http://www. constitution. org/jjr/ineq. htm*.

Saraiva, A. C., & Marshall, L. (2015). Dorsolateral–ventromedial prefrontal cortex interactions during value-guided choice: A function of context or difficulty? *Journal of Neuroscience, 35*, 5087-5088.

Satpute, A. B., & Lieberman, M. D. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research*, *1079*, 86-97.

Shadlen, M. N., & Kiani, R. (2013). Decision-making as a window on cognition. *Neuron*, *80*, 791-806.

Sokol-Hessner, P., Hutcherson, C., Hare, T., & Rangel, A.. (2012). Decision value computation in DLPFC and VMPFC adjusts to the available decision time. *European Journal of Neuroscience, 35*, 1065-1074.

Tajfel, H., & Turner, J. (2001). An integrative theory of intergroup conflict. In M. A. Hogg & D. Abrams (Eds.), *Key readings in social psychology. Intergroup relations: Essential readings* (pp. 94-109). New York, NY, US: Psychology Press.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science*, *19*, 1131-1139.

Van Lange, P. A.M., Joireman, J., Parks, C. D, & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes, 120,* 125-141.

Volk, S., Thöni, C., & Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization*, *81*, 664-676.

Weber, J. M., & Murnighan, J. K. (2008). Suckers or saviors? Consistent contributors in social dilemmas. *Journal of Personality and Social Psychology*, *95*, 1340-1353.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850-852.

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. *Unpublished manuscript.*

Wilson, E. O., (1998). *Consilience: The unity of knowledge*. Now York: Knopf.

Yamagishi, T. (1992). Group size and the provision of a sanctioning system in a social dilemma. In W. B. G. Liebrand, D. M. Messick, & H. A. M. Wilke (Eds.), *International series in experimental social psychology. Social dilemmas: Theoretical issues and research findings* (pp. 267-287). Elmsford, NY, US: Pergamon Press.

Yamagishi, T., Takagishi, H., Fermin, A. D. S. R., Kanai, R., Li, Y., & Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences*, *113*, 5582-5587.

Zaki, J., & Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, *22*, 466-470.