American College *of* Neuropsychopharmacology

www.nature.com/npp

REVIEW ARTICLE OPEN

The prefrontal cortex and (uniquely) human cooperation: a comparative perspective

Yoonseo Zoh[1], Steve W. C. Chang[1,2] and Molly J. Crockett[1,2]✉

Humans have an exceptional ability to cooperate relative to many other species. We review the neural mechanisms supporting human cooperation, focusing on the prefrontal cortex. One key feature of human social life is the prevalence of cooperative norms that guide social behavior and prescribe punishment for noncompliance. Taking a comparative approach, we consider shared and unique aspects of cooperative behaviors in humans relative to nonhuman primates, as well as divergences in brain structure that might support uniquely human aspects of cooperation. We highlight a medial prefrontal network common to nonhuman primates and humans supporting a foundational process in cooperative decision-making: valuing outcomes for oneself and others. This medial prefrontal network interacts with lateral prefrontal areas that are thought to represent cooperative norms and modulate value representations to guide behavior appropriate to the local social context. Finally, we propose that more recently evolved anterior regions of prefrontal cortex play a role in arbitrating between cooperative norms across social contexts, and suggest how future research might fruitfully examine the neural basis of norm arbitration.

Neuropsychopharmacology (2022) 47:119–133; https://doi.org/10.1038/s41386-021-01092-5

INTRODUCTION

Relative to other species, humans have an exceptional ability to cooperate—we are willing to incur personal costs to benefit others, including strangers, and people who we will never meet again [1–7] (see Glossary). These abilities are thought to arise from complex systems of shared moral intuitions about what is "right" or "good" that are culturally transmitted across space and time [8, 9]. Here, we examine the neurocognitive processes that contribute to uniquely human cooperation, focusing on the prefrontal cortex, which has dramatically expanded over the course of human evolution [10] (see Preuss & Wise, this issue).

To organize our review of the prefrontal cortex and human cooperation, we adopt a comparative approach, considering similarities and differences between humans and nonhuman primates in cooperative behavior and its neural underpinnings. In bridging these bodies of research, we identify gaps in our understanding of the neurobiology of cooperation and suggest directions for future research. A comparative behavioral approach allows us to consider how humans navigate social interactions relative to nonhuman primates, including which aspects of cooperative behavior are unique to humans. Likewise, a comparative neuroscience approach can help identify brain mechanisms that may be unique to humans vs. shared with other nonhuman primate species.

In considering the neural underpinnings of human cooperation, we build on prior hypotheses that the prefrontal cortex, especially its more recently evolved anterior components, supports advanced cognitive functions that are unique to humans [11, 12] (see Preuss & Wise, this issue). We note that human cooperation also draws heavily on brain structures outside the prefrontal cortex, as shown in Fig. 1. Among the brain regions implicated in human cooperation, some areas (e.g., the temporoparietal junction or TPJ) do not show structural correspondences between primates and humans. Also, the lack of systematic and controlled comparative studies poses a challenge in drawing conclusions about functional correspondences of certain brain regions between primates and humans. Our primary focus here is on identifying how prefrontal networks support uniquely human aspects of cooperation while also highlighting functional homologies between human and nonhuman primates whenever applicable.

One critical challenge in describing the neural basis of human cooperation concerns defining the functional boundaries between social cognition, and domain-general cognition. Indeed, whether there are neural mechanisms that are specifically "social" remains a topic of ongoing debate [13, 14]. Here, we adopt a view that there are unlikely to be stark categorical boundaries between so-called "social", and "non-social" cognition. For instance, there is evidence for overlapping circuits processing social and non-social rewards [15, 16], and that social decision-making in humans draws upon domain-general valuation processes [17–19]. However, there is also an alternative possibility that there exists a population of neurons specialized for social processing in the same neural circuitry. Nevertheless, the spatial resolution of fMRI precludes drawing firm conclusions about whether indeed the very same neurons are engaged in value computation during social and non-social decisions, leaving open the possibility for domain specificity at a higher spatial or temporal resolution. For the purposes of this review, we therefore focus on the neural mechanisms of domain-general cognitive processes that likely play a central role in social cognition, while remaining agnostic about the question of domain specificity at a neuronal level.

[1] Department of Psychology, Yale University, New Haven, USA. [2] These authors contributed equally: Steve W. C. Chang, Molly J. Crockett. ✉email: molly.crockett@yale.edu

Received: 13 April 2021 Revised: 3 June 2021 Accepted: 24 June 2021
Published online: 19 August 2021

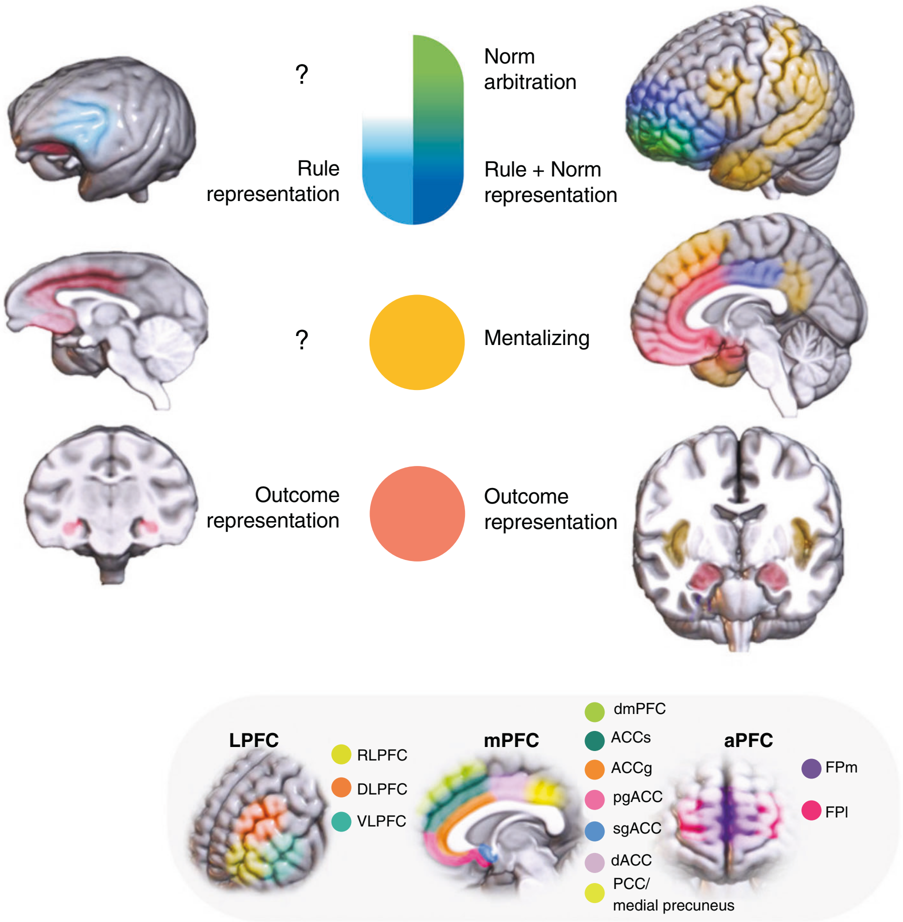

Fig. 1 Brain networks for human cooperation and functional homologies in primates. We highlight four brain networks that play complementary but distinct roles in human cooperation, with the functionally corresponding regions in the brain of nonhuman primate depicted where applicable. An outcome representation network encodes motivationally salient outcomes for self and other and encompasses ACC, dmPFC, amygdala, lOFC, and vmPFC. A second network is recruited during mentalizing, a collection of structures including ACC, dmPFC, pSTS, temporal sulcus, temporal pole, TPJ, medial precuneus, and PCC. A third network is activated when a norm is represented and coordinates the outcome values for self and other. This encompasses the regions of dlPFC, dACC, inferior frontal gyrus, and anterior insula. A fourth network is proposed to be engaged in cooperative norm arbitration and encompasses anterior prefrontal regions including FPC and alPFC. We note that some functional correspondences between humans and nonhuman primates have not been adequately explored to make a firm conclusion, which are denoted by question mark. The specific regions of prefrontal areas we focus on are labeled to clearly identify the distinct roles of subregions reported in previous works.

In this paper, we will first highlight a few key similarities and differences between nonhuman primates and humans in cooperative behaviors. Next, we will survey the prefrontal networks engaged in human cooperative behavior. We start by examining the role of medial prefrontal and striatal systems in the cognitive foundations of cooperative behavior that are present in both nonhuman primates and humans: valuing outcomes for oneself and others, and predicting others' behavior. We then turn to research on cooperative decision-making in humans, reviewing evidence that the lateral prefrontal cortex orchestrates the pursuit of cooperative goals by representing cooperative norms and modulating value representations to guide behavior appropriate to the local social context. Finally, we propose that more recently evolved anterior regions of the prefrontal cortex might play a role in arbitrating between cooperative norms across social contexts, and suggest how future research might fruitfully examine the neural basis of norm arbitration.

UNIQUELY HUMAN COOPERATION?

Comparative studies of cooperation across species suggest that human cooperation is remarkably sophisticated [20, 21]. Behavioral research on this topic occupies a vast literature (for comprehensive reviews, see [22–25]) and there remains debate about which aspects of cooperation (if any) are unique to humans. Many cooperative behaviors in humans and nonhuman species alike can be explained by kinship [26, 27], cooperative breeding [28, 29], and reciprocity [30–33]. However, these mechanisms cannot fully explain the richness and complexity of human cooperation, which encompasses not just cooperation with kin and the repayment of favors, but also cooperation in the absence of direct reciprocal benefits and a commitment to enforcing cooperative principles even when one is not directly affected by uncooperative behavior. In this section, we identify a few key cognitive processes that may undergird extensive cooperation in humans: self-regulation, metacognition, mentalizing, shared intentionality and norm representation. Our goal here is not to provide an exhaustive review, but rather to highlight a few differences across species to set the stage for our central question: how the prefrontal cortex supports human cooperation.

Self-regulation and metacognition

A critical feature of human cooperation is a capacity for self-regulation, i.e., adjusting one's inner states or behaviors according to personal goals, expectations, and standards [34]. Cooperative behavior requires (1) understanding that personal desires often conflict with those of specific others or societal welfare more broadly, (2) regulating those desires in order to behave appropriately, and (3) recognizing when one's behavior falls short of others' expectations in order to improve in the future. These abilities draw on the cognitive processes of metacognition and self-regulation, capacities that are substantially more advanced in humans than nonhuman primates.

In humans, there is evidence that cooperative abilities are related to self-regulation abilities, both in economic games and naturalistic settings [35–38]. Notably, self-regulation abilities (as indexed by reduced discounting of delayed rewards) are markedly more advanced in humans than nonhuman primate species, which may be related to cross-species differences in the scale and scope of cooperation [39].

One particular aspect of self-regulation that may be unique to humans is precommitment: the voluntary restriction of access to temptations[40–46]. For example, dieters might avoid purchasing unhealthy foods so as not to be tempted at home, or people looking to save up for a big purchase might lock funds away in accounts with high early withdrawal fees. In a cooperative setting, precommitment can take the form of a social contract, where all parties agree in advance to adhere to a set of mutually agreed-upon rules or expectations. Indeed, both formal and informal social contracts are a central feature of human moral life and may contribute to humans' extraordinary scope of cooperation over time and space.

Precommitment in humans relies on a metacognitive insight that one's own self-regulation is likely to fail in the absence of a binding contract [47]. Metacognition is the ability to monitor, assess and orchestrate one's own cognitive processes and their quality for the guidance of behavior [48–53]. There is evidence for metacognitive abilities in nonhuman primates [54–60], often operationalized as selective information-seeking behavior when more information needs to be collected to make an informed decision, or confidence judgements indexed by different amounts of wagers on the accuracy of one's performance. While the ability to assess one's prior experience might be shared between human and nonhuman primates, however, metacognition in humans with the use of language and narrative form implicates far more extensive and complex construction of mental models [61]. The advanced metacognitive ability in humans to introspect upon the effectiveness of one's performance also involves increased use of strategies and opportunities for improvement in the future [62–64]. The extensive metacognitive ability in humans prompts the reflection of one's quality as a cooperative partner and further social engagement [65–67].

Mentalizing and shared intentionality

In addition to monitoring and regulating our own cognitive processes, humans also monitor the cognitive processes of others through mentalizing, and communicate the contents of these thought processes with others [67, 68]. While our nonhuman primate relatives can represent what other agents see and know to make informed predictions about their behavior, there is limited evidence that they can represent others' beliefs, in particular false beliefs and ignorance [69]. On the other hand, humans develop early in life the ability to infer and understand the dynamic mental states of others that are distinct from one's own, even when those mental states deviate from reality [70–75]. Humans also spontaneously attribute mental states to others to make sense of others' behavior as arising from intentional stance [76]. Metacognition and mentalizing enhance cooperation by enabling people to share information about their reasons for acting, resulting in more accurate models of one's own and others' behavior [67, 68].

The capacity to attribute mental states to oneself and others offers significant advantages in building shared understanding of cooperative goals and actions to achieve them. Comparative and developmental research suggests humans can represent a concept of *shared intentionality*, which brings *us* to a common understanding that *we* are jointly committed to achieving mutual goals in collaborative interactions [77] (Fig. 2). This capacity for shared intentionality motivates us to engage in cooperative acts even with distant strangers and to regulate individual desires when they conflict with collective goals. It marks a significant departure from nonhuman primates for whom the ability to think and act interdependently is thought to be much weaker, resulting in a limited ability to collaborate with joint commitment to collective goals [78].

There are several possibilities as to why humans came to exhibit this significant departure from the capacity of nonhuman primates. Firstly, it is likely to be an evolutionary adaptation to better cope with the cognitive demands from expanding social group size and complexity [12]. The demands for interdependence and collaboration over collective goals might have emerged in tandem with an increase in group size for efficient distribution of environmental resources, resulting in the development of new sociocognitive skills to address them. Another possibility is that shared intentionality may have been acquired and transmitted through social interactions more effectively in humans than our primate relatives [79]. That is, through repeated social interactions,

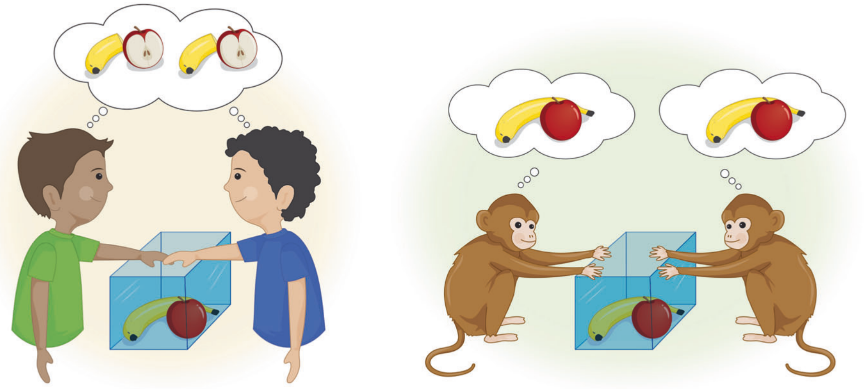

Fig. 2 Shared intentionality in humans. Comparative and developmental work shows that humans (but not our closest primate relatives) are able to represent shared intentions and goals in a way that facilitates cooperation. Human children are more likely to work toward a common goal and share the spoils (left panel), while nonhuman primates may work together to obtain rewards but do not show evidence of commitment to shared goals (right panel).

humans might have gradually acquired the better way to collaborate for mutual benefit and passed this cumulative knowledge to later generations through cultural learning. In both cases, it is evident that shared intentionality allowed humans to become richly involved in practices of cooperative culture and norms.

Evidence for shared intentionality comes from studies comparing cooperation in human children and nonhuman primates. For example, when presented with a choice between working alone and collaborating together to obtain food, from age 2 to 3, children preferred to collaborate, while chimpanzees preferred to work alone [80]. When interrupted during a joint activity, children aged between 21 and 27 months attempted to re-engage the partner, even when the partner was not necessarily needed, suggesting that children do not consider their partner merely as instrumental to fulfilling an individual goal, but rather as collaborative partners with whom they coordinate joint intentions [81]. Moreover, 3.5-year-old children in dyadic collaborative tasks continue to collaborate until both parties obtained rewards, rather than stopping once their own reward became available, supporting their understanding of mutual commitment for all involved parties [82]. Finally, when resources were gained through collaborative effort, children shared them more equally than when resources were windfalls or resulted from individual effort; chimpanzees, meanwhile, did not show such differentiation in sharing [83]. These findings together suggest a uniquely human capacity to establish joint goals and intentions from an early age.

Norm representation and enforcement

Shared intentionality enables humans to conceptualize *social norms:* "commonly known standards of behavior that are based on widely shared views about how individual group members ought to behave in a given situation" [84]. Research on social norms has identified a variety of different types of norms, from the mundane and conventional (e.g., norms for how to dress in different social settings) to norms that carry more moral weight (e.g., norms against lying, stealing, and cheating) [85]. Here, we focus on *cooperative norms*, which we define as social norms that facilitate cooperative interactions: i.e., norms that motivate individuals to cooperate when it is not in their economic self-interest. [84, 86, 87]

One common cooperative norm is a norm of fairness, i.e., that windfall resources should be distributed equally among group members [88]. Fairness concerns are observable across human cultures [5, 89] and emerge early in life [90]. Notably, children not only display aversion toward receiving less than others (disadvantageous inequity aversion), but also toward receiving more than others (advantageous inequity aversion) [90, 91]. Although nonhuman primates show evidence for disadvantageous inequity aversion (e.g., [92, 93]), it remains unclear if they also exhibit an aversion to advantageous inequity [94]. Other cooperative norms that appear to be unique to humans include norms of honesty, promise-keeping, and conditional cooperation [84].

When a cooperative norm is violated, humans will incur costs to enforce the norm by punishing the transgressor [95]. People are willing to enforce norms not just when they are the victim of the transgression ('second-party punishment'), but also when they are unaffected by the transgression ('third-party punishment') [96, 97]. Both second- and third-party punishment contribute to the stability of cooperation [95, 96, 98]. A taste for norm enforcement emerges at a young age in humans [99]. Costly third-party punishment can be observed in children as young as 3 [100]. More broadly, from age 3, children display an awareness of normative structures governing simple rule-based games, and will intervene to teach third parties who do not follow the rules [101]. This suggests children from an early age can acquire, implement, and enforce knowledge of normative principles. In contrast, it is unclear whether nonhuman primates are willing to engage in costly norm enforcement. Experimental work demonstrates that chimpanzees engage in second-party punishment but not third-party punishment [102]. Thus, a willingness to enforce cooperative norms beyond direct retaliation may be unique to humans.

Summary

Cooperation based on kinship and reciprocity is common to nonhuman primates and humans. However, humans are substantially more advanced in their ability to regulate individual desires towards the achievement of collective goals. Moreover, sophisticated metacognitive abilities of monitoring our own cognitive processes as well as others through mentalizing enables us to establish common understanding of joint goals and intentions. The ensuing shared intentionality, in turn, lays the foundation for uniquely human abilities to represent cooperative norms, comply with those norms oneself, and enforce norm compliance in others. These sets of advanced abilities in humans, building on one another and working in concert, may enable humans to conceptualize informal, temporally, and socially extended social contracts that play an important role in guiding cooperative behavior.

MEDIAL PFC AND SOCIAL VALUATION

When we make decisions, we must represent the set of available choice options, assign subjective values to the expected outcomes resulting from these options, select the most valuable option, evaluate the outcome, and update subjective value representations through learning [103]. For cooperative decisions, the computation of subjective value requires representing expected outcomes for oneself and others, and weighting those outcomes according to the anticipated responses of others and local cooperative norms [104–106]

For instance, imagine you are sitting in a park at lunchtime. Your stomach rumbles and so you begin unwrapping a sandwich you've just purchased at a local café. Before you can take a bite, a homeless person approaches and asks if you could spare some money for a meal. You have no cash on you. Do you offer them your sandwich?

Decisions like this engage a variety of cognitive processes. Your overall decision will depend not just on how much you value your own and others' outcomes for their own sake, but also the relative valuation of those outcomes and the valuation of normative principles like generosity more broadly [105]. So you will need to compute the value of the sandwich not just for yourself, but also for the homeless person, how guilty you will feel if you refuse, and how virtuous you will feel if you help. These latter computations will likely depend on local cooperative norms (e.g., are you in a country where it's expected to help strangers in need?), personal commitments (e.g., do you practice a religion which prioritizes generosity?), and whether others are watching (e.g., are you trying to impress a date?). In the next two sections, we briefly review the neural correlates of some of these computations engaged in cooperative decision-making, highlighting where relevant the neural mechanisms that appear to be common to both humans and nonhuman primates.

Representing outcomes for self and others

Rewarding and aversive outcomes for oneself are strong motivators of decision-making. Converging evidence in nonhuman primates and humans demonstrates that motivationally salient outcomes are encoded in a valuation circuitry encompassing multiple cortical and subcortical brain regions including ventromedial prefrontal cortex (vmPFC), orbitofrontal cortex (OFC), different subregions of anterior cingulate cortex (ACC), ventral striatum, and amygdala (for comprehensive reviews, see [107–110]). These regions are active during a series of mental operations involved in value-based decision-making, from representation, valuation, and action selection to outcome evaluation and updating value representations [98, 110–111] (see Shenhav; Murray & Adolphs, this issue). In particular, studies using single-unit recordings in nonhuman primates as well as model-based fMRI in humans revealed that the neuronal/BOLD activity in these regions is parametrically modulated according to the magnitude of the positive and negative outcomes expected or received [112, 113]. Moreover, these regions respond to diverse forms of reward and punishment (e.g., money, food, pain, social approval/disapproval), providing support for a ‘common neural currency’ schema for domain-general coding of value [114].

Cooperative decisions are guided not just by outcomes for oneself, but also outcomes for others. How does the brain represent *vicarious outcomes*—i.e., motivationally salient outcomes for others? Are others’ rewards and punishments encoded in similar or distinct neural circuits to one’s own rewards and punishments? A growing body of work implicates the medial prefrontal cortex, especially the ACC gyrus, in vicarious outcome processing [115]. Initial work on this question demonstrated that the brain’s valuation circuitry responds not just to rewards for oneself, but also vicarious rewards [116–118]. For example, viewing another person winning a game evoked increased neural responses in the ventral striatum, a region in the valuation network which was also active when participants themselves experienced winning. Notably, the effect of vicarious reward in the ventral striatum was modulated by the perceived similarity to the person observed, and this modulation was indexed by increased functional connectivity between the subgenual ACC and the ventral striatum [116]. A meta-analysis of 25 neuroimaging studies investigating neural correlates of vicarious reward identified overlapping activations between personal and vicarious reward in the valuation network, including the subgenual and dorsal ACC, rostral mPFC, vmPFC, and the amygdala [119]. Vicarious reward responses in the brain are stronger in individuals higher in empathy [120] and predict cooperative behavior [121–123].

Similar evidence for vicarious neural representations in the mPFC/ACC can be found in the aversive domain. Pain delivered to both oneself and others evokes activity in the mid-ACC and anterior insula (AI), regions implicated in the affective component of pain [124, 125]. Multivariate techniques suggest that first-hand and vicarious pain experiences share common neural representations in the AI and the mid-ACC [126, 127]. Vicarious pain responses in the brain, like vicarious reward responses, track with individual differences in empathy [128] and predict cooperative behavior [129]. Empathic brain responses in the mid-ACC and AI are also highly sensitive to context, and this contextual sensitivity is thought to arise from interactions with regions implicated in mentalizing, including the dorsomedial prefrontal cortex (dmPFC) and TPJ [130–132].

Alongside studies suggesting overlapping neural representations of outcomes for oneself and others [133] there is evidence that the mPFC, and in particular the gyrus of ACC, represents distinctive information about others’ outcomes [115, 134]. Single-cell recording in nonhuman primates, which enables finer-grained examination of neuronal encoding of vicarious outcomes, suggests separable encoding of outcomes for self and others. One study demonstrated that the functionally separate but anatomically intermingled populations of neurons in the dmPFC (pre-supplemental motor area and rostral area 9) encode reward information separately for self and other [93]. In another study, where monkeys made decisions about allocating juice rewards to themselves, a conspecific monkey, both self and other, or neither, many neurons in the rostral ACC gyrus selectively encoded others’ reward outcomes, while neurons in the OFC most prominently encoded one’s own reward outcomes [135]. Rostral ACC gyrus neurons in humans also encode others’ rewarding outcomes [136]. This suggests the rostral ACC and dmPFC may be necessary for learning which actions result in positive outcomes for others, an ability crucial for the development of cooperative behavior. Supporting this prediction, lesioning the ACC (encompassing both the ventral sulcus and the gyrus) prevented monkeys from learning which actions help others [137].

Relative and joint valuation of outcomes

Aside from representing the value of outcomes for oneself and others individually, medial prefrontal areas are also sensitive to the *relative values* of one’s own outcomes with respect to others and vice-versa. In one study, pairs of participants took part in a random drawing where one received $50 and the other received nothing. Following this, each was scanned while observing monetary transfers to self and other. For the participant who received nothing, vmPFC and ventral striatum responded more strongly to rewards for self than other. High-pay participants, meanwhile, showed the reverse pattern, with stronger responses to rewards for the disadvantaged partner than rewards for self [138]. Similarly, responses to money in the vmPFC and rostral ACC are higher when that money is offered as part of a fair split (e.g., $5 out of $10), relative to an unfair split (e.g., $5 out of $20; [139, 140]). Single-cell recordings in macaque monkeys provide convergent evidence for relative social valuation of rewards: the subjective value of rewards for self decreases in tandem with increasing reward allocations to others, and these relative values are encoded, respectively, in the dopaminergic midbrain and dmPFC [93].

In addition, there is evidence that medial prefrontal regions encode the *joint values* that arise from cooperative interactions where the whole is greater than the sum of its parts. Human fMRI studies of social exchange games like the prisoner’s dilemma show increased activity in vmPFC for cooperative relative to selfish decisions [141–143], and during repeated social interactions, activity in anterior and mid-cingulate cortex tracked the partner’s cooperative decisions [144]. Consistent with these findings, dorsal ACC neurons in macaque monkeys playing an iterative prisoner’s dilemma encoded the partner’s future decision to cooperate [145]. In the same study, a largely separate population of dorsal ACC neurons encoded the monkey’s own cooperative decisions, and disrupting dorsal ACC activity selectively inhibited mutual cooperation.

Integrative valuation in social decision-making

Social decision-making requires not only representing the value of decision outcomes for self and others, but also integrating those values into an “all-things-considered” subjective value, or “relative chosen value” corresponding to the value of the chosen course of action relative to the alternatives. The vmPFC and dmPFC (including the dorsal ACC and pre-SMA), respectively, are suggested to positively and negatively encode relative chosen value, both for decisions that impact only oneself [107, 146–148], only others [149, 150], or both oneself and others [19, 151–154]. A recent study suggests relative value encoding in dmPFC generalizes across tasks and across self- and other-related valuations, implicating this region as a node for computing relative subjective values for self and other [150].

Making decisions on behalf of others may be particularly difficult if the other person is a stranger or if their preferences are not well understood. To overcome this uncertainty, people may engage in mentalizing processes that activate a network encompassing superior temporal sulcus, TPJ, medial precuneus, and dmPFC [155–158]. Mentalizing regions may encode subjective values themselves [154, 159] or be functionally connected with medial prefrontal valuation regions during social decision-making [149, 152, 160, 161]. Together these data support an ‘extended common currency schema’ for social decision-making, whereby social cognitive information (represented in mentalizing areas) modulates the activity of a domain-general value-representation circuitry that computes subjective value for both social and non-social decisions [105].

Summary

Representing outcomes for oneself and others is an important first step during cooperative decision-making. Converging evidence from studies of humans and nonhuman primates demonstrates a role for medial prefrontal regions, including dorsal and ventral ACC, in encoding the value of outcomes for self and others. Medial prefrontal regions encode vicarious rewards and punishments, and individual differences in these responses predict individual differences in cooperation. These neurons are sensitive to relative and joint values in social contexts, which may explain cooperative behavior in both nonhuman primates and humans. During social decision-making, subjective values are computed in vmPFC and dmPFC, and in humans, there is evidence that these domain-general valuation regions receive input from other areas that represent social information (such as the mentalizing network). Although both humans and primates represent outcomes for self and others in medial prefrontal regions, humans go considerably further in incorporating social cognitive information into integrative values of action through actively engaging in mentalizing. In addition, human cooperative decision-making cannot be understood simply in terms of these basic outcome representations. In the next section, we'll examine how the subjective value of cooperative decisions depends not just on relative and joint valuation of social outcomes, but also on shared beliefs about what is normatively right or wrong.

LATERAL PFC AND NORMATIVE BEHAVIOR

When people make cooperative decisions, they consider not just the possible outcomes for themselves and others, but also the *meaning* of their actions and resulting outcomes in the context of local cooperative norms (see Glossary). As we suggested in Section 1, the ability to represent cooperative norms and use those norms to guide one's own and others' behavior may be unique to humans. In this section, we review work on the neural mechanisms of norm compliance and enforcement. As cooperative norms are rules that guide cooperative interactions, we begin by considering work in both nonhuman primates and humans implicating lateral PFC (lPFC; including dorsal and ventral components) in representing rules. Next, we turn to work on norm compliance and enforcement in humans, highlighting the role of lPFC and its interactions with the brain's valuation circuitry [106, 162]. This work suggests that representations of cooperative norms in lPFC modulate the processing of outcomes for self and other in mPFC broadly and subcortical areas, enabling individuals to prioritize norm compliance and enforcement over selfish interests.

Rule representation

The ability to represent and use rules to guide appropriate actions is a core aspect of goal-directed behavior [163–166]. Single-neuron recording studies in nonhuman primates have identified several brain regions involved in rule representation. When rhesus monkeys performed a task of switching between two rules, the activity of neurons distributed throughout the PFC, including dlPFC, vlPFC, and OFC, flexibly encoded the rule being applied [167]. Importantly, many of these neurons maintained the rule information over a sustained period of time and showed signatures of mapping the rules to a new set of stimuli, suggesting that multiple PFC areas support abstract rule representation that permits flexible applications of learned rules to new circumstances [168, 169]. Moreover, dlPFC lesions were specifically associated with the impairment in monkeys' ability to shift between rules [169], whereas OFC lesions were selectively linked to the deficit in the capacity to reverse stimulus-reward associations, suggesting that representations of rule and reward value can be dissociated in the PFC [169].

Recent evidence shows that human lPFC computes higher-order goals based on the associations between rules and expected outcomes to ultimately guide action selection in the striatum by modulating choice-related value signals [106, 170]. More specifically, lPFC is thought to mediate a contextual modulation of subjective value by adjusting the weights assigned to multiple environmental and behavioral attributes that are integrated and mapped onto prospective outcome values (see Shenhav, this issue). This view has broadened the scope of lPFC operations to all aspects of goal-directed behavioral control. A hierarchical organization of lPFC, where contextual information propagates through the rostral-caudal gradients of abstraction, as well as its robust projections to the striatum and vmPFC further support this view [171–173]. In addition, evidence for the higher-order modulatory role of lPFC in value-based decision-making has been continuously reported in the self-control literature, establishing a potential link between lPFC and the temporal regulation of value [174–176].

Such evidence indicates that both human and nonhuman primate lPFC is centrally involved in rule representation and flexible rule-guided behavior. It is possible that humans' exceptional ability to comply with and enforce cooperative norms may be realized by some aspects of the lPFC function that diverged between humans and nonhuman primates through evolutionary elaboration. Next, we discuss how lateral prefrontal mechanisms, in parallel with its domain general functions, contribute to norm compliance and enforcement. We note that there are no published studies of the neural basis of norm compliance and enforcement in nonhuman primates; therefore, the remainder of this section will focus on research in humans.

Cooperative norm compliance

In line with its domain-general support for rule-guided behavior, lPFC is reliably engaged in neuroimaging studies where human participants decide whether to comply with cooperative norms. For example, in settings where fairness norms are highly salient, several studies reported increased right dlPFC activity when participants make decisions to fairly distribute money with interaction partners [177, 178]. Disrupting activity in the right dlFPC reduces the fairness of decisions in repeated interactions without affecting explicit beliefs about what is fair [179–182]. When fairness norms are less salient, dlPFC activity has been associated with less fair behavior [180, 183–185], suggesting that dlPFC may implement fairness norms in a context-specific manner.

Other work has implicated lPFC in complying with a norm of honesty, using tasks where participants are tempted to earn more money by cheating or lying. Right dlPFC and vlPFC are more active when participants respond honestly in these tasks [186–188]. Enhancing right dlPFC activity increases honest behavior [189] and patients with dlPFC lesions make fewer honest choices [190], suggesting a causal role for dlPFC in upholding norms of honesty.

Similarly, dlPFC is more active when participants uphold a norm of conditional cooperation, i.e., reciprocating others' trusting or cooperative decisions [191–193]. Relatedly, individuals who are more prone to conditional cooperation show higher baseline tone in left dlPFC [194], and patients with damage to dlPFC are less likely to cooperate in social dilemmas [195]. Together, these findings demonstrate the engagement of lPFC regions, most commonly dlPFC, in complying with a variety of cooperative norms.

Initial theorizing on the role of lPFC in cooperative norm compliance proposed that this region implements a top-down inhibition of prepotent selfish impulses [196], drawing on classical accounts of lPFC in response inhibition [197]. More recently, lPFC's role in norm compliance has been reinterpreted through the lens of domain-general theories of value-based decision-making [123, 180, 198, 199]. By these accounts, lPFC integrates goal-relevant information, such as norms, beliefs, and the mental states of others into value computations in a goal-directed manner [104]. Rather than inhibiting prepotent selfish responses, lPFC modulates the subjective value of behaviors that maximize selfish outcomes at the expense of norm compliance. Put simply, we do not comply with norms by overcoming a temptation to deviate, but because norm-deviant behaviors are less tempting in the first place. This mechanism is thought to operate across a variety of domains, including abstract rule-based decision-making [200] and dietary self-control [174].

LPFC modulation of subjective value is hypothesized to operate via functional interactions with valuation circuitry including vmPFC and subcortical areas [201–205]. Several studies have reported dlPFC activity when participants have to trade off benefits to oneself against cooperative norm compliance [151, 188, 206, 207]. For example, in a study where participants had the opportunity to earn money by delivering painful electric shocks to either themselves or another person, participants with stronger cooperative preferences showed decreased responses to money earned by harming others (relative to oneself) in a network of value-encoding regions including dorsal striatum and vmPFC. dlPFC tracked anticipated blameworthiness for harmful choices and showed negative functional connectivity with dorsal striatum when participants chose to forego the ill-gotten gains [19]. Another study probed the neural representation of cooperative norms by explicitly instructing participants to either focus on the ethical implications of their choices, the impact on others, or respond naturally when deciding how to allocate money between themselves and another person. Participants made more generous choices when focusing on cooperative norms and social consequences, and showed goal-sensitive encoding of choice attributes in dlPFC, such that fairness and outcomes for others were weighted more strongly when participants focused on complying with cooperative norms [208].

What motivates people to comply with cooperative norms when no one is watching? Normative and descriptive theories have highlighted a role for moral emotions like guilt in guiding compliant behavior even in the absence of external punishments. Studies using formal models of guilt aversion reveal activation in dlPFC during guilt-averse decisions [177, 191]. One study disentangling guilt and inequity aversion during a modified trust game showed that guilt aversion was associated with right dlPFC activity, while inequity aversion was reflected in the activity of the ventral striatum and amygdala, and enhancing dlPFC excitability with tDCS increased reliance on guilt-aversion [177]. When individuals can choose freely between guilt-averse and inequity-averse strategies, preference for guilt-averse strategy corresponded with multivariate patterns of activity in a network including left dlPFC, AI, mPFC, and putamen [209]. Given the relationship between guilt and anticipated blame [210], these findings dovetail with the observation of blame representation in dlPFC during cooperative decision-making [19].

Cooperative norm enforcement

Previous theorizing has suggested norm compliance and enforcement might rely on common psychological and neural mechanisms [97, 162, 211–213]. Accordingly, lPFC (in particular dlPFC) has also been implicated in cooperative norm enforcement across diverse scenarios and tasks. Studies of norm enforcement behavior via second-party punishment in economic games show the engagement of dlPFC when participants punish others for treating them unfairly [214–216]. Disrupting right dlPFC activity with TMS reduces costly second-party punishment without affecting fairness judgments [215, 217] through functional interactions with vmPFC, which shows a reduced response during punishment decisions when right dlPFC is deactivated [215].

Norm enforcement via third-party punishment shares neural substrates with second-party punishment [218, 219]. Studies measuring both types of punishment in the same participants during economic games report common activation in dlPFC and bilateral AI [216, 220–222], findings also confirmed by meta-analysis of second- and third-party punishment studies [223, 224]. Relative to second-party punishment, third-party punishment is more likely to engage anterior vlPFC and TPJ more strongly, suggesting a greater involvement of mentalizing processes when punishing on behalf of others [224]. Consistent with these findings, patients with lesions to dlPFC and mentalizing network demonstrate atypical third-party punishment behavior [225].

During third-party punishment, dlPFC is hypothesized to integrate multiple streams of information, including the amount of harm and the intentions of the transgressor [97]. Supporting this view, in studies probing punishment decisions of criminal scenarios, dlPFC shows stronger activity for culpable acts [226] and increased functional connectivity with regions encoding mental states and harm to others [227–229]. Accordingly, disrupting activity in dlPFC with TMS interferes with the integration of information about mental states and harm to others [199]. Together these findings highlight a causal role for dlPFC in the representational integration of multiple attributes that contribute to punishment decisions.

Summary

Cooperative norms play a central role in guiding the large-scale cooperative interactions that characterize human social life. Humans are willing to incur considerable personal costs to comply with cooperative norms and enforce those norms in others. Converging evidence implicates the lateral PFC, most commonly dlPFC, in representing cooperative norms and integrating those representations with other streams of information to guide normative behavior. Functional interactions between lateral and medial prefrontal regions suggest that the former modulates value representations in the latter in a goal-directed manner, consistent with other work highlighting a domain-general role for dlPFC in rule representation that is common to both humans and nonhuman primates.

FUTURE DIRECTIONS: ANTERIOR PFC AND NORM ARBITRATION

Certain cooperative norms, such as a prohibition against physically harming an innocent person for pure personal gain, apply widely across a vast range of social and cultural settings [230]. However, other cooperative norms, such as an expectation to tip restaurant servers, are specific to particular cultures, social contexts, or social relationships. We therefore must be able to flexibly select among different cooperative norms to guide our behavior in a context-appropriate way. In this section, we identify future directions for research on the prefrontal cortex and human cooperation: how do we arbitrate between conflicting cooperative norms in situations that are ambiguous in terms of which norm to follow?

Take, for example, a set of cooperative norms around dining in a Western cultural setting. At a restaurant, a diner is expected to pay for their meal (indeed, this norm is codified into law). However, offering to pay for a home-cooked meal at a friend's house would be seen as socially awkward or outright rude. This discrepancy can be explained by the fact that different norms operate in these different settings: the restaurant meal is governed by an (economic) *exchange norm*, where benefits are provided with the expectation of receiving a comparable benefit or payment in return, while the friend's dinner is governed by a *communal norm*, where benefits are given without any expectation of compensation [231]. Oftentimes the relevant norm will be clear, perhaps via commonly understood "social scripts" or salient cues that indicate what is appropriate in the present situation. But what about situations where there is no clearly defined norm, such as dining out with a friend in a restaurant? Here, it may be more or less appropriate to pay for one's own meal, depending on the size of the bill, the nature of the relationship between the friends, and the occasion of the meal—an exchange norm might be more appropriate for a business lunch between acquaintances, while a communal norm might be more appropriate for a birthday meal between a child and a parent. How do people make decisions in normatively ambiguous situations like these?

To address this question, we build on studies of how decision-makers reflect on their own choices and arbitrate between different decision strategies. This work suggests that the anterior PFC, including anterior lPFC as well as frontopolar cortex (FPC, Brodmann area 10), plays a key role in metacognition, counterfactual processing, and arbitrating between valuation systems. We suggest that *norm arbitration* might draw on similar neural processes.

Notably, FPC is unique to anthropoid primates [232], and is the largest area in human PFC. Comparative work suggests lateral FPC (lFPC) is unique to humans [233, 234], suggesting this region may support uniquely human cognition for cooperation. Moreover, existing literature supports the presence of an anterior-posterior anatomical gradient in hierarchical processing in the primate PFC, with the anterior PFC being more specialized for higher-order and metacognitive functions than the posterior PFC [11]. In this section, we first briefly review work on the neural basis of metacognition and value arbitration, and then propose how the anterior PFC (and in particular the FPC) might guide cooperative norm arbitration in humans.

Neural basis of metacognition

Successful cooperation requires an ability to reflect on one's own thought processes and behaviors. Numerous studies have implicated a frontoparietal network that includes FPC and anterior lPFC (alPFC; BA 47) in metacognition (for reviews see [235–238]). As described in Section 1, metacognitive abilities appear to be more extensive in humans than nonhuman primates. Nevertheless, there is some evidence that FPC, the most anterior region in the primate PFC, plays a role in monitoring decision strategies in nonhuman primates. One study found that FPC neurons retrospectively encode chosen goals as feedback approaches [239]. Such signals could be used to assess the reliability of decision strategies and monitor self-generated goals [240]. Moreover, inactivation of FPC (area 10) selectively interfered with awareness of non-experienced events, but not with experienced events, suggesting this region is causally implicated in the evaluation of one's ignorance [241].

In humans, FPC is dramatically expanded compared with nonhuman primates [242], which may explain cross-species differences in the complexity of metacognitive processes [234, 243]. Individual variability in metacognitive accuracy (i.e., the ability to accurately judge the success of cognitive processes) is correlated with gray matter volume in mFPC [244], and patients with FPC lesions show reduced metacognitive accuracy [245], suggesting a causal role for this region in metacognitive performance. During value-based decision-making (in a task where hungry participants chose between different snack foods), alPFC encoded subjective confidence in choices, and showed functional connectivity with value-encoding regions of vmPFC. This functional connectivity, in turn, predicted individual variability in the relationship between confidence and choice accuracy [246].

One important aspect of evaluating the quality of one's decisions is keeping track of counterfactual outcomes: is the grass greener on the other side of the fence? Multiple studies have shown that FPC prospectively tracks counterfactual evidence, including the reward value of unchosen options [247] and counterfactual prediction errors [248]. When choices lead to unexpected outcomes, alPFC activity mediates the impact of postdecision evidence on choice confidence, suggesting it may guide changing one's mind on the basis of new evidence [249].

There is also evidence that FPC prospectively tracks internal variables that bear on the future success of decisions. During perceptual decision-making, people prospectively estimate an internal probability of making a correct choice, and these estimations are tracked in mFPC and alPFC [250]. A similar process may take place during self-regulation by precommitment. When people limit their access to tempting small immediate rewards, lFPC is more active, and connectivity between lFPC and the frontoparietal control network is stronger in people who stand to benefit more from precommitment [47]. A subsequent study showed that enhancing lFPC activity with anodal transcranial direct current stimulation selectively increased decisions to precommit [251]. These findings suggest that FPC and alPFC may orchestrate self-regulation in part through accessing internal signals that convey the likely success of different decision strategies.

Together these studies suggest that FPC, in particular its more lateral aspects that underwent a dramatic expansion over the course of human evolution, plays a fundamental role in enabling people to evaluate the quality of their decisions both prospectively and retrospectively—an important aspect of adjusting cooperative behavior across social contexts. However, it is important to note that the neural evidence of metacognition is not limited to the FPC. In fact, multiple regions in the FPC network, encompassing ACC subregions and mPFC subareas, will likely synergistically contribute to the metacognitive process. For instance, confidence-related signals have been observed in the perigenual ACC [252, 253] and dmPFC [253] for making decisions relevant for self and other. Moreover, the highly integral functions of ACC in flexibly implementing behaviors should play an important role in the FPC network with respect to metacognition and other cognitive processes underlying cooperative decision-making (see Monosov & Rushworth, this issue).

Arbitrating between valuation systems

Research on the neurocomputational mechanisms of value-based learning and decision-making reveals that there are multiple valuation systems in the brain that employ different algorithms for learning the expected value of actions and outcomes. In particular, an important distinction has been made between 'goal-directed' and 'habitual' systems, which are proposed to rely on model-based and model-free reinforcement learning algorithms, respectively [254, 255]. The computationally expensive model-based algorithm learns contingencies between actions and outcomes to build a model of the world, and selects actions by prospectively searching through the model and selecting a course of action that serves current goals. Meanwhile, the computationally efficient model-free algorithm assigns values to actions through trial and error, and habitually selects actions with the highest cached value. Oftentimes the goal-directed and habitual systems agree on what is the best choice, but sometimes they give conflicting answers.

Recent work has examined the neurocomputational mechanisms supporting arbitration between model-based and model-free control over decision-making. An early computational account suggested an uncertainty-based arbitration, whereby control is exerted by the system with the lowest uncertainty in its value predictions [256]. Building on this account, the "mixture of experts" framework proposes that the brain monitors the reliabilities of the predictions of different valuation systems (the 'experts'), and uses those reliabilities to allocate control over behavior [257] (see O'Doherty & Averbeck, this issue). This arbitration mechanism, proposed to rely on anterior PFC (including FPC and vlPFC), assigns a weight to each expert on the basis of its reliability, gating the extent to which that expert's recommended policy contributes to action selection, and transmits this information to the vmPFC, which serves as the system's output channel, encoding an integrated subjective value. The final policy arises from combining across the opinions of the individual experts, weighted by their relative confidence. Therefore, rather

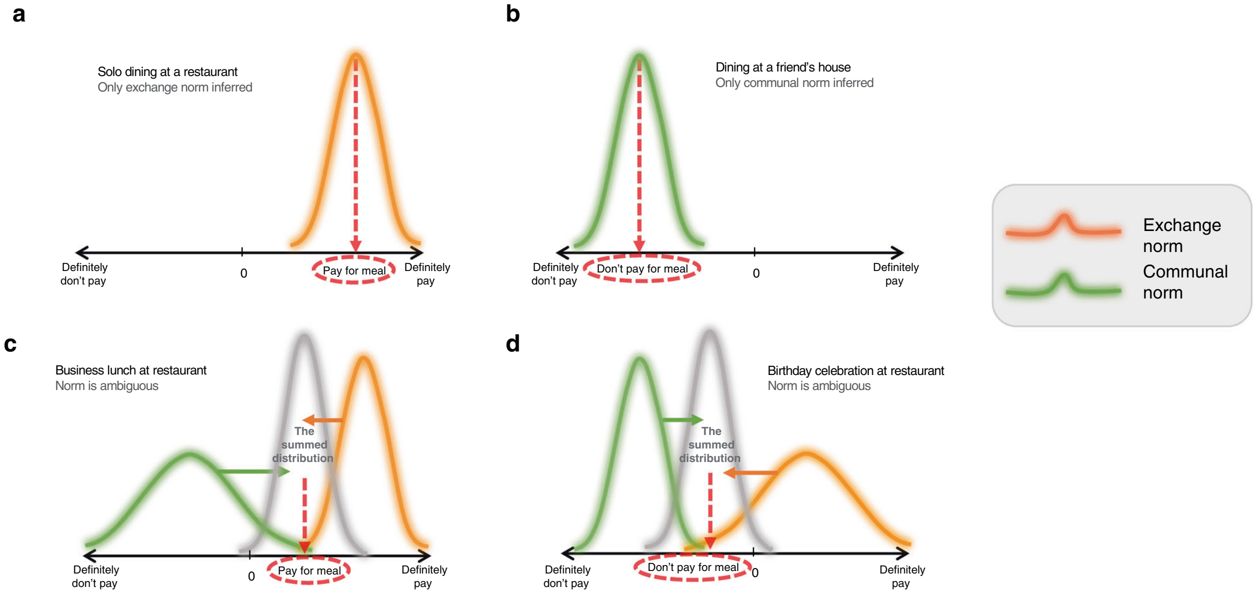

Fig. 3 Cue-guided cooperative norm arbitration. Graphical depiction of arbitration between communal and exchange norms across various social settings. **a** When dining alone at a restaurant, salient cues indicating an exchange norm is appropriate result in a more confident prediction about the expected value of paying for one's meal. **b** When dining at a friend's house, salient cues indicating a communal norm is appropriate result in a more confident prediction about the expected value of not paying for one's meal. **c**, **d** When you are having a business lunch or having a birthday celebration at a restaurant, there is no obvious norm. Therefore, cues indicating communal or exchange norms increase the precision of the respective predictions. The downstream decoder in anterior PFC allocates weights over two norms based on their relative reliabilities, which are determined as a function of each norm's expected value and (inverse of) uncertainty [273]. Two weighted distributions are linearly summed to evaluate the confidence in choosing the policy of communal norm over exchange norm, based on a mechanism informed by the multi-sensory cue integration literature [274]. This way, the decisions are adjusted flexibly based on the linear summation of the relative reliability of two norms. Even when the decision is expressed as binary choice between two presented options, because the final decision is made on the basis of weighted average, the relative reliability of the relevant norms alter the degree of confidence about the decision.

than implementing only one strategy at a time dictated by the dominating expert system, the brain can efficiently and flexibly utilize collective expertise of different systems. This model bears some resemblance to the optimal integration model of sensory perception, whereby the combination of multiple sensory cues is achieved by linear summation of population activity generated by each sensory cue [258–260]. In both models, the final policy (or percept) is influenced more strongly by the more confident "voice".

Evidence for the mixture of experts model comes from fMRI studies of human subjects in learning environments where different learning strategies are variably successful over time. In a study of arbitration between model-based and model-free systems, reliabilities for both systems were encoded in vlPFC and FPC [261]. Enhancing vlPFC activity with anodal tDCS increased model-based control, while inhibiting vlPFC activity with cathodal tDCS had the opposite effect, suggesting vlPFC gates the extent of default model-free control over behavior, amplifying model-based control when advantageous [262]. Another study showed that vlPFC tracks the reliabilities of emulative and imitative strategies during observational learning, with imitative strategy employed as a default [263]. Finally, a study of arbitration between individual experience and social advice also revealed reliability signals of each strategy encoded in right FPC and vlPFC [264]. Together these findings provide initial support for a domain-general arbitration process in anterior PFC that polls the reliabilities of different learning and decision-making systems to allocate control over behavior.

Arbitrating between cooperative norms

We suggest that a process similar to the mixture of experts model is likely to guide arbitration between different cooperative norms in guiding context-appropriate social behavior. That is, norm arbitration might involve allocating control over behavior by weighting the reliabilities of value predictions generated by different cooperative norms. We propose that anterior PFC computes the reliabilities of the expected values of different behavioral policies prescribed by any cooperative norms being considered. These predictions may be generated based on the presence of contextual cues that indicate whether a particular norm is relevant for the current context. The posterior over the behavioral policies can therefore be obtained as a weighted sum over the predicted values of cues indicating different norms, producing graded levels of confidence over the chosen policy.

Consider the example that opened this section: how do you decide whether to pay for your meal when dining? There are two potentially relevant cooperative norms that make opposite behavior policy recommendations: an exchange norm that dictates you should pay for your meal, and a communal norm that suggests you should not. When you are dining solo at a restaurant (Fig. 3a), the exchange norm is strongly confident in its recommendation that paying your bill is the best option because there are only exchange cues present; when invited for dinner at a friend's house, the communal norm makes a robust prediction that you should avoid pulling out your wallet when dessert is served, because there are only communal cues present (Fig. 3b). But what should you do when dining at a restaurant with a friend (Fig. 3c, d)? This will depend on the relative numbers and strengths of surrounding communal and exchange cues. For instance, if you and your friend have met in the cafeteria of your office building and have brought some work materials to discuss (an exchange cue), the exchange norm will make a more reliable prediction and you will feel more confident about asking for separate checks (Fig. 3c). Whereas if you are celebrating your friend's birthday and have brought them a birthday card (a communal cue), the communal norm will make a more reliable prediction and you will feel more confident about treating them to their meal (Fig. 3d).

Empirical evidence for this hypothesis is so far scarce. One possible explanation for this missing evidence is that most studies of social decision-making in humans consider only a single social context or norm, applied over a very brief timescale. For instance, neuroeconomic paradigms typically consider how individuals implement fairness norms in one-shot interactions. Such paradigms generally do not require participants to arbitrate between behavioral policies, but instead to consider implementing a single policy (such as a norm of fairness or reciprocity).

However, preliminary support for norm arbitration in FPC comes from neuroimaging studies of social decision-making that require participants to consider multiple decision contexts and strategies. One early study examined the neural basis of compliance with a fairness norm in two distinctive contexts: one where punishment was possible, and another where it was not. The contrast between the two conditions revealed increased activation in alPFC and vlPFC as well as dlPFC [178]. Another study measured the brain activity of individuals playing a repeated public goods game where they faced a series of decisions about whether to contribute some money to benefit their entire group. In this setting, individuals must trade off selfish concerns against long-term group benefits. Because participants interacted repeatedly within the same group, it was possible to dissociate brain signals encoding how much an individual stood to benefit from the current decision (individual utility) and how much the group could benefit from the remaining interactions (group utility). While individual utility was encoded in the vmPFC, group utility was encoded in the lFPC, and changes in individual choice strategies were mediated by functional interactions between lFPC, dorsal ACC, and vlPFC [265]. lFPC has also been implicated in adaptively choosing how to publicly communicate private mental states. In a task where participants privately assess their decision confidence and adapt their confidence reports to different social partners in order to maximize rewards, signals in lFPC tracked with social contexts requiring higher adjustments of confidence reports, and multivariate activity patterns in lFPC represented distinguished between these social contexts [266]. Together these findings converge on a role for anterior PFC in adaptively adjusting decision-making across different social contexts.

Finally, the anterior PFC undergoes a protracted period of development, rapidly increasing its volume and dendritic complexity throughout late childhood and adolescence [267–269] (see Kolk & Rakic, this issue). Intriguingly, developmental studies also show that children increasingly adjust their moral judgments to social-relational context as they get older. For example, while older children (aged 6–7) and adults believe that authority figures are more obligated than ordinary individuals to punish wrongdoing, but ordinary individuals are not, younger children (aged 4–5) believe that everyone is obligated to punish [270]. Relatedly, older children and adults believe that friends are more obligated to help one another than strangers, while younger children believe that everyone is equally obligated to help one another [271]. Whether the increasing sensitivity of moral judgments to relational context depends on the development of anterior PFC is an intriguing question for future study.

Of course, there may be other possible mechanisms that support arbitration between different cooperative norms. One alternative possibility is a “winner-take-all” mechanism whereby detection of a cue signaling one norm over another prompts a categorical dominance of the relevant norm, rather than the weighted averaging approach described above. Such a mechanism would be akin to sensory cue-separation where one of two channels dominates, rather than aggregating across multiple channels as in sensory cue integration [258, 259, 272]. Another possibility is that certain cooperative norms, like a prohibition against physically harming others, are so deeply ingrained and apply so universally that they dominate across contexts operating more like a Pavlovian reflex or habit [17, 273, 274]. Future work could fruitfully adjudicate between these possibilities.

Summary

The anterior PFC, in particular the FPC, has dramatically expanded over the course of human evolution and is implicated in a variety of cognitive processes that may be unique to humans, including advanced metacognitive abilities, the capacity to represent and learn from complex counterfactuals, and arbitrating between different strategies for learning and decision-making. Building on this work, we propose that the anterior PFC supports norm arbitration: determining which cooperative norm(s) to apply in a particular social context. Preliminary evidence for this hypothesis comes from neuroimaging studies of adaptive social decision-making, which show that the anterior PFC (including anterior vlPFC and lFPC) encodes variables that are necessary for maximizing rewards across diverse social contexts. We further hypothesize that there is a possible anterior-posterior gradient of norm processing such that norm arbitration is carried out by more anterior aspects of the PFC, whereas norm representation is mediated by more posterior aspects of the PFC. Future work can extend these findings by probing the engagement of anterior PFC in tasks where participants must use cooperative norms to guide decisions across different cultural or relational contexts.

CONCLUSION

The prefrontal cortex, in particular its more anterior regions, has expanded dramatically over the course of human evolution. In tandem, the scale and scope of human cooperation has dramatically outpaced its counterparts in nonhuman primate species, manifesting as complex systems of moral codes that guide normative behaviors even in the absence of punishment or repeated interactions. Here, we provided a selective review of the neural basis of human cooperation, taking a comparative approach to identify the brain systems and social behaviors that are thought to be unique to humans. Humans and nonhuman primates alike cooperate on the basis of kinship and reciprocity, but humans are unique in their abilities to represent shared goals and self-regulate to comply with and enforce cooperative norms on a broad scale. We highlight three prefrontal networks that contribute to cooperative behavior in humans: a medial prefrontal network, common to humans and nonhuman primates, that values outcomes for self and others; a lateral prefrontal network that guides cooperative goal pursuit by modulating value representations in the context of local norms; and an anterior prefrontal network that we propose serves uniquely human abilities to reflect on one’s own behavior, commit to shared social contracts, and arbitrate between cooperative norms across diverse social contexts. We suggest future avenues for investigating cooperative norm arbitration and how it is implemented in prefrontal networks.

REFERENCES

1. Bowles S, Gintis H. A cooperative species. Princeton, NJ: Princeton University press; 2011.
2. Camerer CF. Behavioral game theory: experiments in strategic interaction. Princeton, NJ: Princeton University press; 2011.
3. Dawes RM, Thaler RH. Anomalies: cooperation. J Econ Perspect. 1988;2:187–97.
4. Fehr E, Fischbacher U, Gächter S. Strong reciprocity, human cooperation, and the enforcement of social norms. Hum Nat. 2002;13:1–25.
5. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, et al. “Economic man” in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behav Brain Sci. 2005;28:795–815.
6. McCabe KA, Rigdon ML, Smith VL. Positive reciprocity and intentions in trust games. J Econ Behav Organ. 2003;52:267–75.

7. Henrich J, Fehr E. Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. Genetic and cultural evolution of cooperation, Cambridge, MA: MIT Press; 2003. p. 55–82.
8. Rand DG, Nowak MA. Human cooperation. Trends Cogn Sci. 2013;17:413–25.
9. Tomasello M. Why we cooperate. Cambridge, MA: MIT press; 2009.
10. Smaers JB, Gómez-Robles A, Parks AN, Sherwood CC. Exceptional evolutionary expansion of prefrontal cortex in great apes and humans. Curr Biol. 2017;27:1549.
11. Passingham RE, Wise SP. The neurobiology of the prefrontal cortex: anatomy, evolution, and the origin of insight. Oxford, UK: Oxford University Press; 2012.
12. Dunbar RIM. The social brain hypothesis. Evol Anthropol Issues N. Rev. 1998;6:178–90.
13. Lockwood PL, Apps MAJ, Chang SWC. Is there a 'Social' Brain? Implementations and algorithms. Trends Cogn Sci. 2020;24:802–13.
14. Young L, Dungan J. Where in the brain is morality? Everywhere and maybe nowhere. Soc Neurosci. 2012;7:1–10.
15. Izuma K, Saito DN, Sadato N. Processing of social and monetary rewards in the human striatum. Neuron 2008;58:284–94.
16. Lin A, Adolphs R, Rangel A. Social and monetary reward learning engage overlapping neural substrates. Soc Cogn Affect Neur. 2011;7:274–81.
17. Lockwood PL, Klein-Flügge MC, Abdurahman A, Crockett MJ. Model-free decision making is prioritized when learning to avoid harming others. Proc Natl Acad Sci. 2020;117:27719–30.
18. Shenhav A, Greene JD. Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron 2010;67:667–77.
19. Crockett MJ, Siegel JZ, Kurth-Nelson Z, Dayan P, Dolan RJ. Moral transgressions corrupt neural representations of value. Nat Neurosci. 2017;20:879–85.
20. Melis AP, Semmann D. How is human cooperation different? Philos Trans R Soc B Biol Sci. 2010;365:2663–74.
21. Enke B. Kinship, cooperation, and the evolution of moral systems. Q J Econ. 2019;134:953–1019.
22. Brosnan SF. Justice- and fairness-related behaviors in nonhuman primates. Proc Natl Acad Sci. 2013;110:10416–23.
23. Cronin KA, Hopper LM, Sommerville JA, Decety J. A comparative perspective on helping and fairness. Social cognition: development across the life span. New York, NY: Routledge; 2016. p. 26–45.
24. Smuts BB, Cheney DL, Seyfarth RM, Wrangham RW. Primate societies. Chicago, IL: University of Chicago Press; 2008.
25. Mitani JC, Call J, Kappeler PM, Palombit RA, Silk JB. The evolution of primate societies. Chicago, IL: University of Chicago Press; 2012.
26. Cheney DL. Extent and limits of cooperation in animals. Proc Natl Acad Sci. 2011;108:10902–9.
27. Clutton-Brock T. Breeding together: kin selection and mutualism in cooperative vertebrates. Science. 2002;296:69–72.
28. Burkart JM, Allon O, Amici F, Fichtel C, Finkenwirth C, Heschl A, et al. The evolutionary origin of human hyper-cooperation. Nat Commun. 2014;5:4747.
29. Burkart JM, Hrdy SB, Schaik CPV. Cooperative breeding and human cognitive evolution. Evol Anthropol Issues N. Rev. 2009;18:175–86.
30. Silk JB. Nepotistic cooperation in non-human primate groups. Philos Trans R Soc B Biol Sci. 2009;364:3243–54.
31. Axelrod R, Hamilton W. The evolution of cooperation. Science. 1981;211:1390–6.
32. Fischbacher U, Gächter S, Fehr E. Are people conditionally cooperative? Evidence from a public goods experiment. Econ Lett. 2001;71:397–404.
33. Schmelz M, Grueneisen S, Kabalak A, Jost J, Tomasello M. Chimpanzees return favors at a personal cost. Proc Natl Acad Sci. 2017;114:7462–7.
34. Vohs KD, Baumeister RF. Understanding self-regulation. Handbook of self-regulation: research, theory, and applications. New York, NY: Guilford Publications; 2004.
35. Curry OS, Price ME, Price JG. Patience is a virtue: cooperative people have lower discount rates. Pers Indiv Differ. 2008;44:780–5.
36. Harris AC, Madden GJ. Delay discounting and performance on the Prisoner's dilemma game. Psychological Rec. 2002;52:429–40.
37. Fehr E, Leibbrandt A. A field study on cooperativeness and impatience in the tragedy of the commons. J Public Econ. 2011;95:1144–55.
38. Yi R, Johnson MW, Bickel WK. Relationship between cooperation in an iterated prisoner's dilemma game and the discounting of hypothetical outcomes. Learn Behav. 2005;33:324–36.
39. Stevens JR, Cushman FA, Hauser MD. Evolving the psychological mechanisms for cooperation. Annu Rev Ecol Evol Syst. 2005;36:499–518.
40. Rachlin H, Green L. Commitment, choice and self-control1. J Exp Anal Behav. 1972;17:15–22.
41. Ainslie GW. Impulse control in pigeons1. J Exp Anal Behav. 1974;21:485–9.
42. Wertenbroch K. Consumption self-control by rationing purchase quantities of virtue and vice. Mark Sci. 1998;17:317–37.
43. Ariely D, Wertenbroch K. Procrastination, deadlines, and performance: self-control by precommitment. Psychol Sci. 2001;13:219–24.
44. Kalenscher T, Pennartz CMA. Is a bird in the hand worth two in the future? The neuroeconomics of intertemporal decision-making. Prog Neurobiol. 2008;84:284–315.
45. Fujita K. On conceptualizing self-control as more than the effortful inhibition of impulses. Pers Soc Psychol Rev. 2011;15:352–66.
46. Elster J, Jon E. Ulysses unbound: studies in rationality, precommitment, and constraints. New York, NY: Cambridge University Press; 2000.
47. Crockett MJ, Braams BR, Clark L, Tobler PN, Robbins TW, Kalenscher T. Restricting temptations: neural mechanisms of precommitment. Neuron. 2013;79:391–401.
48. John D, Janet M. Metacognition. Thousand Oaks, CA: Sage Publications; 2008.
49. Nelson TO, Narens, L. Metamemory: A theoretical framework and some new findings. The Psychology of Learning and Motivation. Vol. 26. New York, NY: Academic Press; 1990. p. 125–73.
50. Beran MJ, Brandl JL, Perner J, Proust J. Foundations of metacognition. Oxford, UK: Oxford University Press; 2012.
51. Norman DA, Shallice T. Attention to action. Consciousness and self-regulation. Boston, MA: Springer; 1986. p. 1–18.
52. Metcalfe J, Shimamura AP. Metacognition: Knowing about knowing. Cambridge, MA: MIT press; 1994.
53. Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD. Supra-personal cognitive control and metacognition. Trends Cogn Sci. 2014;18:186–93.
54. Rosati AG, Santos LR. Spontaneous metacognition in Rhesus monkeys. Psychol Sci. 2016;27:1181–91.
55. Smith JD, Shields WE, Washburn DA. The comparative psychology of uncertainty monitoring and metacognition. Behav Brain Sci. 2003;26:317–39.
56. Kornell N. Where Is the "Meta" in animal metacognition? J Comp Psychol. 2014;128:143–9.
57. Smith JD. The study of animal metacognition. Trends Cogn Sci. 2009;13:389–96.
58. Terrace HS, Son LK. Comparative metacognition. Curr Opin Neurobiol. 2009;19:67–74.
59. Crystal JD. Comparative approaches to metacognition: prospects, problems, and the future. Animal Behav Cogn. 2019;6:254–61.
60. Kornell N. Metacognition in humans and animals. Curr Dir Psychol Sci. 2009;18:11–15.
61. Johnson-Laird PN. Mental models and human reasoning. Proc Natl Acad Sci. 2010;107:18243–50.
62. Son LK, Metcalfe J. Metacognitive and control strategies in study-time allocation. J Exp Psychol Learn Mem Cogn. 2000;26:204–21.
63. Metcalfe J. Metacognitive judgments and control of study. Curr Dir Psychol Sci. 2009;18:159–63.
64. Nelson TO, Dunlosky J, Graf A, Narens L. Utilization of metacognitive judgments in the allocation of study during multitrial learning. Psychol Sci. 1994;5:207–13.
65. Hasson-Ohayon I, Gumley A, McLeod H, Lysaker PH. Metacognition and intersubjectivity: reconsidering their relationship following advances from the study of persons with psychosis. Front Psychol. 2020;11:567.
66. Fischer MW, Dimaggio G, Hochheiser J, Vohs JL, Phalen P, Lysaker PH. Metacognitive capacity is related to self-reported social functioning and may moderate the effects of symptoms on interpersonal behavior. J Nerv Ment Dis. 2019;208:138–42.
67. Frith CD. The role of metacognition in human social interactions. Philos Trans R Soc B Biol Sci. 2012;367:2213–23.
68. Frith CD, Frith U. Mechanisms of social cognition. Psychology. 2012;63:287–313.
69. Arre AM, Santos LR. Mentalizing in non-human primates. The neural basis of mentalizing, New York, NY: Springer Press; 2021. p. 131–47
70. Buttelmann D, Carpenter M, Tomasello M. Eighteen-month-old infants show false belief understanding in an active helping paradigm. Cognition. 2009;112:337–42.
71. Helming KA, Strickland B, Jacob P. Making sense of early false-belief understanding. Trends Cogn Sci. 2014;18:167–70.
72. Luo Y. Do 10-month-old infants understand others' false beliefs? Cognition. 2011;121:289–98.
73. Onishi KH, Baillargeon R. Do 15-month-old infants understand false beliefs? Science. 2005;308:255–8.
74. Scott RM, Baillargeon R. Early false-belief understanding. Trends Cogn Sci. 2017;21:237–49.
75. Sodian B. Theory of mind in infancy. Child Dev Perspect. 2011;5:39–43.
76. Dennett DC. The intentional stance. Cambridge, MA: MIT press; 1989.
77. Tomasello M. The ultra-social animal. Eur J Soc Psychol. 2014;44:187–94.
78. Tomasello M, Herrmann E. Ape and human cognition. Curr Dir Psychol Sci. 2010;19:3–8.

79. Cecilia H. Cognitive gadgets: the cultural evolution of thinking. Cambridge, MA: Harvard University press; 2018.
80. Rekers Y, Haun DBM, Tomasello M. Children, but not chimpanzees, prefer to collaborate. Curr Biol. 2011;21:1756–8.
81. Warneken F, Gräfenhain M, Tomasello M. Collaborative partner or social tool? New evidence for young children's understanding of joint intentions in collaborative activities. Dev Sci. 2012;15:54–61.
82. Hamann K, Warneken F, Tomasello M. Children's developing commitments to joint goals. Child Dev. 2012;83:137–45.
83. Hamann K, Warneken F, Greenberg JR, Tomasello M. Collaboration encourages equal sharing in children but not in chimpanzees. Nature. 2011;476:328–31.
84. Fehr E, Schurtenberger I. Normative foundations of human cooperation. Nat Hum Behav. 2018;2:458–68.
85. Turiel E. The development of social knowledge: morality and convention. Cambridge, UK: Cambridge University Press; 1983.
86. Fehr E, Fischbacher U. Social norms and human cooperation. Trends Cogn Sci. 2004;8:185–90.
87. Peysakhovich A, Rand DG. Habits of virtue: creating norms of cooperation and defection in the laboratory. Manag Sci. 2016;62:631–47.
88. Fehr E, Schmidt KM. A theory of fairness, competition, and cooperation. Q J Econ. 1999;114:817–68.
89. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, et al. Costly punishment across human societies. Science. 2006;312:1767–70.
90. McAuliffe K, Blake PR, Steinbeis N, Warneken F. The developmental foundations of human fairness. Nat Hum Behav. 2017;1:0042.
91. Ulber J, Hamann K, Tomasello M. Young children, but not chimpanzees, are averse to disadvantageous and advantageous inequities. J Exp Child Psychol. 2017;155:48–66.
92. Báez-Mendoza R, Coeverden CR, van, Schultz W. A neuronal reward inequity signal in primate striatum. J Neurophysiol. 2016;115:68–79.
93. Noritake A, Ninomiya T, Isoda M. Social reward monitoring and valuation in the macaque brain. Nat Neurosci. 2018;21:1452–62.
94. Brosnan SF, Waal FBMde. Evolution of responses to (un)fairness. Science. 2014;346:1251776.
95. Fehr E, Gächter S. Altruistic punishment in humans. Nature. 2002;415:137–40.
96. Fehr E, Fischbacher U. Third-party punishment and social norms. Evol Hum Behav. 2004;25:63–87.
97. Buckholtz JW, Marois R. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat Neurosci. 2012;15:655–61.
98. Mathew S, Boyd R. Punishment sustains large-scale cooperation in prestate warfare. Proc Natl Acad Sci. 2011;108:11375–80.
99. McAuliffe K, Jordan JJ, Warneken F. Costly third-party punishment in young children. Cognition. 2015;134:1–10.
100. Vaish A, Missana M, Tomasello M. Three-year-old children intervene in third-party moral transgressions. Brit J Dev Psychol. 2011;29:124–30.
101. Rakoczy H, Warneken F, Tomasello M. The sources of normativity: young children's awareness of the normative structure of games. Dev Psychol. 2008;44:875–81.
102. Riedl K, Jensen K, Call J, Tomasello M. No third-party punishment in chimpanzees. Proc Natl Acad Sci. 2012;109:14824–9.
103. Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci. 2008;9:545–56.
104. Pärnamets P, Shuster A, Reinero DA, Bavel JJV. A value-based framework for understanding cooperation. Curr Dir Psychol Sci. 2020;29:227–34.
105. Ruff CC, Fehr E. The neurobiology of rewards and values in social decision making. Nat Rev Neurosci. 2014;15:549–62.
106. Buckholtz JW. Social norms, self-control, and the value of antisocial behavior. Curr Opin Behav Sci. 2015;3:122–9.
107. Bartra O, McGuire JT, Kable JW. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage. 2013;76:412–27.
108. Liu X, Hairston J, Schrier M, Fan J. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. Neurosci Biobehav Rev. 2011;35:1219–36.
109. Clithero JA, Rangel A. Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neur. 2014;9:1289–302.
110. O'Doherty JP. Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol. 2004;14:769–76.
111. Monte OD, Chu CCJ, Fagan NA, Chang SWC. Specialized medial prefrontal–amygdala coordination in other-regarding decision preference. Nat Neurosci. 2020;23:565–74.
112. O'Doherty JP, Hampton A, Kim H. Model-Based fMRI and its application to reward learning and decision making. Ann Ny Acad Sci. 2007;1104:35–53.
113. Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature 2009;459:837–41.
114. Levy DJ, Glimcher PW. The root of all value: a neural common currency for choice. Curr Opin Neurobiol. 2012;22:1027–38.
115. Apps MAJ, Rushworth MFS, Chang SWC. The anterior cingulate gyrus and social cognition: tracking the motivation of others. Neuron. 2016;90:692–707.
116. Mobbs D, Yu R, Meyer M, Passamonti L, Seymour B, Calder AJ, et al. A key role for similarity in vicarious reward. Science. 2009;324:900.
117. Braams BR, Peters S, Peper JS, Güroğlu B, Crone EA. Gambling for self, friends, and antagonists: differential contributions of affective and social brain regions on adolescent reward processing. Neuroimage. 2014;100:281–9.
118. Fareri DS, Niznikiewicz MA, Lee VK, Delgado MR. Social network modulation of reward-related signals. J Neurosci. 2012;32:9045–52.
119. Morelli SA, Sacchet MD, Zaki J. Common and distinct neural correlates of personal and vicarious reward: a quantitative meta-analysis. Neuroimage. 2015;112:244–53.
120. Lockwood PL, Apps MAJ, Valton V, Viding E, Roiser JP. Neurocomputational mechanisms of prosocial learning and links to empathy. Proc Natl Acad Sci. 2016;113:9763–8.
121. Moll J, Krueger F, Zahn R, Pardini M, Oliveira-Souza R, de, Grafman J. Human fronto–mesolimbic networks guide decisions about charitable donation. Proc Natl Acad Sci. 2006;103:15623–8.
122. Harbaugh WT, Mayr U, Burghart DR. Neural responses to taxation and voluntary giving reveal motives for charitable donations. Science. 2007;316:1622–5.
123. Sul S, Tobler PN, Hein G, Leiberg S, Jung D, Fehr E, et al. Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proc Natl Acad Sci. 2015;112:7851–6.
124. Lamm C, Decety J, Singer T. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage. 2011;54:2492–502.
125. Bastiaansen JACJ, Thioux M, Keysers C. Evidence for mirror systems in emotions. Philos Trans R Soc B Biol Sci. 2009;364:2391–404.
126. Corradi-Dell'Acqua C, Hofstetter C, Vuilleumier P. Felt and seen pain evoke the same local patterns of cortical activity in insular and cingulate cortex. J Neurosci. 2011;31:17996–8006.
127. Corradi-Dell'Acqua C, Tusche A, Vuilleumier P, Singer T. Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. Nat Commun. 2016;7:10904.
128. Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD. Empathy for pain involves the affective but not sensory components of pain. Science. 2004;303:1157–62.
129. Hein G, Silani G, Preuschoff K, Batson CD, Singer T. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron. 2010;68:149–60.
130. Bruneau EG, Pluta A, Saxe R. Distinct roles of the 'Shared Pain' and 'Theory of Mind' networks in processing others' emotional suffering. Neuropsychologia. 2012;50:219–31.
131. Danziger N, Faillenot I, Peyron R. Can we share a pain we never felt? Neural correlates of empathy in patients with congenital insensitivity to pain. Neuron. 2009;61:203–12.
132. Zaki J, Weber J, Bolger N, Ochsner K. The neural bases of empathic accuracy. Proc Natl Acad Sci. 2009;106:11382–7.
133. Lamm C, Bukowski H, Silani G. From shared to distinct self–other representations in empathy: evidence from neurotypical function and socio-cognitive disorders. Philos Trans R Soc B Biol Sci. 2016;371:20150083.
134. Lockwood PL. The anatomy of empathy: vicarious experience and disorders of social cognition. Behav Brain Res. 2016;311:255–66.
135. Chang SWC, Gariépy J-F, Platt ML. Neuronal reference frames for social decisions in primate frontal cortex. Nat Neurosci. 2013;16:243–50.
136. Hill MR, Boorman ED, Fried I. Observational learning computations in neurons of the human anterior cingulate cortex. Nat Commun. 2016;7:12722.
137. Basile BM, Schafroth JL, Karaskiewicz CL, Chang SWC, Murray EA. The anterior cingulate cortex is necessary for forming prosocial preferences from vicarious reinforcement in monkeys. Plos Biol. 2020;18:e3000677.
138. Tricomi E, Rangel A, Camerer CF, O'Doherty JP. Neural evidence for inequality-averse social preferences. Nature. 2010;463:1089–91.
139. Tabibnia G, Satpute AB, Lieberman MD. The sunny side of fairness. Psychol Sci. 2007;19:339–47.
140. Feng C, Luo Y, Krueger F. Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. Hum Brain Mapp. 2015;36:591–602.
141. Decety J, Jackson PL, Sommerville JA, Chaminade T, Meltzoff AN. The neural bases of cooperation and competition: an fMRI investigation. Neuroimage. 2004;23:744–51.

142. Rilling JK, Gutman DA, Zeh TR, Pagnoni G, Berns GS, Kilts CD. A neural basis for social cooperation. Neuron. 2002;35:395–405.

143. Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. Neuroreport. 2004;15:2539–2243.

144. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. Getting to know you: reputation and trust in a two-person economic exchange. Science. 2005;308:78–83.

145. Haroush K, Williams ZM. Neuronal prediction of opponent's behavior during cooperative social interchange in primates. Cell. 2015;160:1233–45.

146. Kable JW, Glimcher PW. The neural correlates of subjective value during intertemporal choice. Nat Neurosci. 2007;10:1625–33.

147. Kolling N, Wittmann MK, Behrens TEJ, Boorman ED, Mars RB, Rushworth MFS. Value, search, persistence and model updating in anterior cingulate cortex. Nat Neurosci. 2016;19:1280–5.

148. Kolling N, Behrens TEJ, Mars RB, Rushworth MFS. Neural mechanisms of foraging. Science. 2012;336:95–98.

149. Janowski V, Camerer C, Rangel A. Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Soc Cogn Affect Neur. 2013;8:201–8.

150. Piva M, Velnoskey K, Jia R, Nair A, Levy I, Chang SW. The dorsomedial prefrontal cortex computes task-invariant relative subjective value for self and other. Elife. 2019;8:e44939.

151. Qu C, Hu Y, Tang Z, Derrington E, Dreher J-C. Neurocomputational mechanisms underlying immoral decisions benefiting self or others. Soc Cogn Affect Neur. 2020;15:135-49.

152. Hare TA, Camerer CF, Knoepfle DT, O'Doherty JP, Rangel A. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. J Neurosci. 2010;30:583–90.

153. Zaki J, Mitchell JP. Equitable decision making is associated with neural markers of intrinsic value. Proc Natl Acad Sci. 2011;108:19761–6.

154. Hutcherson CA, Bushong B, Rangel A. A neurocomputational model of altruistic choice and its implications. Neuron 2015;87:451–62.

155. Koster-Hale J, Saxe R. Theory of mind: a neural prediction problem. Neuron. 2013;79:836–48.

156. Jamali M, Grannan BL, Fedorenko E, Saxe R, Báez-Mendoza R, Williams ZM. Single-neuronal predictions of others' beliefs in humans. Nature. 2021;591:610–4.

157. Schurz M, Radua J, Aichhorn M, Richlan F, Perner J. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci Biobehav Rev. 2014;42:9–34.

158. Schurz M, Radua J, Tholen MG, Maliske L, Margulies DS, Mars RB, et al. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol Bull. 2021;147:293–327.

159. Fukuda H, Ma N, Suzuki S, Harasawa N, Ueno K, Gardner JL, et al. Computing social value conversion in the human brain. J Neurosci. 2019;39:3117–8.

160. Strombach T, Weber B, Hangebrauk Z, Kenning P, Karipidis II, Tobler PN, et al. Social discounting involves modulation of neural value signals by temporoparietal junction. Proc Natl Acad Sci. 2015;112:1619–24.

161. Kameda T, Inukai K, Higuchi S, Ogawa A, Kim H, Matsuda T, et al. Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. Proc Natl Acad Sci. 2016;113:11817–22.

162. Carlson RW, Crockett MJ. The lateral prefrontal cortex and moral goal pursuit. Curr Opin Psychol. 2018;24:77–82.

163. Ramnerö J, Törneke N. On having a goal: goals as representations or behavior. Psychological Rec. 2015;65:89–99.

164. O'hora D, Maglieri KA. Goal statements and goal-directed behavior. J Organ Behav Manag. 2006;26:131–70.

165. Fellner DJ, Sulzer-Azaroff B. A behavioral analysis of goal setting. J Organ Behav Manag. 2008;6:33–51.

166. Malott RW. A theory of rule-governed behavior and organizational behavior management. J Organ Behav Manag. 1993;12:45–65.

167. Wallis JD, Anderson KC, Miller EK. Single neurons in prefrontal cortex encode abstract rules. Nature. 2001;411:953–6.

168. Miller EK, Freedman DJ, Wallis JD. The prefrontal cortex: categories, concepts and cognition. Philos Trans R Soc Lond Ser B Biol Sci. 2002;357:1123–36.

169. Dias R, Robbins TW, Roberts AC. Dissociation in prefrontal cortex of affective and attentional shifts. Nature. 1996;380:69–72.

170. Dixon ML, Christoff K. The lateral prefrontal cortex and complex value-based learning and decision making. Neurosci Biobehav Rev. 2014;45:9–18.

171. Petrides M. Lateral prefrontal cortex: architectonic and functional organization. Philos Trans R Soc B Biol Sci. 2005;360:781–95.

172. Nee DE, D'Esposito M. The hierarchical organization of the lateral prefrontal cortex. Elife. 2016;5:e12112.

173. Badre D, Nee DE. Frontal cortex and the hierarchical control of behavior. Trends Cogn Sci. 2018;22:170–88.

174. Hare TA, Camerer CF, Rangel A. Self-control in decision-making involves modulation of the vmPFC valuation system. Science. 2009;324:646–8.

175. Hare TA, Malmaud J, Rangel A. Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. J Neurosci. 2011;31:11077–87.

176. Hare TA, Hakimi S, Rangel A. Activity in dlPFC and its effective connectivity to vmPFC are associated with temporal discounting. Front Neurosci-Switz. 2014;8:50.

177. Nihonsugi T, Ihara A, Haruno M. Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. J Neurosci. 2015;35:3412–9.

178. Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E. The neural signature of social norm compliance. Neuron 2007;56:185–96.

179. Knoch D, Schneider F, Schunk D, Hohmann M, Fehr E. Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. Proc Natl Acad Sci. 2009;106:20895–9.

180. Ruff CC, Ugazio G, Fehr E. Changing social norm compliance with noninvasive brain stimulation. Science 2013;342:482–4.

181. Soutschek A, Sauter M, Schubert T. The importance of the lateral prefrontal cortex for strategic decision making in the Prisoner's dilemma. Cogn Affect Behav Neurosci. 2015;15:854–60.

182. Strang S, Gross J, Schuhmann T, Riedl A, Weber B, Sack AT. Be nice if you have to — the neurobiological roots of strategic fairness. Soc Cogn Affect Neur. 2015;10:790–6.

183. FeldmanHall O, Dalgleish T, Thompson R, Evans D, Schweizer S, Mobbs D. Differential neural circuitry and self-interest in real vs hypothetical moral decisions. Soc Cogn Affect Neur. 2012;7:743–51.

184. Yamagishi T, Takagishi H, Fermin A, de SR, Kanai R, Li Y, et al. Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. Proc Natl Acad Sci. 2016;113:5582–7.

185. Yin Y, Yu H, Su Z, Zhang Y, Zhou X. Lateral prefrontal/orbitofrontal cortex has different roles in norm compliance in gain and loss domains: a transcranial direct current stimulation study. Eur J Neurosci. 2017;46:2088–95.

186. Yin L, Reuter M, Weber B. Let the man choose what to do: neural correlates of spontaneous lying and truth-telling. Brain Cognition. 2016;102:13–25.

187. Abe N, Greene JD. Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. J Neurosci. 2014;34:10564–72.

188. Hu J, Hu Y, Li Y, Zhou X. Computational and neurobiological substrates of cost-benefit integration in altruistic helping decision. J Neurosci. 2021;41:3545–61

189. Maréchal MA, Cohn A, Ugazio G, Ruff CC. Increasing honesty in humans with noninvasive brain stimulation. Proc Natl Acad Sci. 2017;114:4360–4.

190. Zhu L, Jenkins AC, Set E, Scabini D, Knight RT, Chiu PH, et al. Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. Nat Neurosci. 2014;17:1319–21.

191. Chang LJ, Smith A, Dufwenberg M, Sanfey AG. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron. 2011;70:560–72.

192. Suzuki S, Niki K, Fujisaki S, Akiyama E. Neural basis of conditional cooperation. Soc Cogn Affect Neur. 2011;6:338–47.

193. Cutler J, Campbell-Meiklejohn D. A comparative fMRI meta-analysis of altruistic and strategic decisions to give. Neuroimage. 2019;184:227–41.

194. Baumgartner T, Dahinden FM, Gianotti LRR, Knoch D. Neural traits characterize unconditional cooperators, conditional cooperators, and noncooperators in group-based cooperation. Hum Brain Mapp. 2019;40:4508–17.

195. Wills J, FeldmanHall O, Collaboration NP, Meager MR, Bavel JJV, Blackmon K, et al. Dissociable contributions of the prefrontal cortex in group-based cooperation. Soc Cogn Affect Neur. 2018;13:nsy023-.

196. Knoch D, Fehr E. Resisting the power of temptations. Ann Ny Acad Sci. 2007;1104:123–34.

197. Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 2001;24:167–202.

198. Morris A, Cushman F. A common framework for theories of norm compliance. Soc Philos Policy. 2018;35:101–27.

199. Buckholtz JW, Martin JW, Treadway MT, Jan K, Zald DH, Jones O, et al. From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron. 2015;87:1369–80.

200. Duverne S, Koechlin E. Rewards and cognitive control in the human prefrontal cortex. Cereb Cortex. 2017;27:5024–39.

201. Pornpattananangkul N, Zhen S, Yu R. Common and distinct neural correlates of self-serving and prosocial dishonesty. Hum Brain Mapp. 2018;39:3086–103.

202. Dogan A, Morishima Y, Heise F, Tanner C, Gibson R, Wagner AF, et al. Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. Sci Rep. 2016;6:33263.

203. Hu J, Li Y, Yin Y, Blue PR, Yu H, Zhou X. How do self-interest and other-need interact in the brain to determine altruistic behavior? Neuroimage. 2017;157:598–611.
204. Hackel LM, Wills JA, Bavel JJV. Shifting prosocial intuitions: neurocognitive evidence for a value-based account of group-based cooperation. Soc Cogn Affect Neur. 2020;15:371–81.
205. Vaidya AR, Fellows LK. Under construction: ventral and lateral frontal lobe contributions to value-based decision-making and learning. F1000research. 2020;9:F1000 Faculty Rev–158.
206. Qu C, Météreau E, Butera L, Villeval MC, Dreher J-C. Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. Plos Biol. 2019;17:e3000283.
207. Shuster A, Levy DJ. Contribution of self- and other-regarding motives to (dis) honesty. Sci Rep. 2020;10:15844.
208. Tusche A, Hutcherson CA. Cognitive regulation alters social and dietary choice by changing attribute representations in domain-general and domain-specific brain circuits. Elife. 2018;7:e31185.
209. Baar JM, van, Chang LJ, Sanfey AG. The computational and neural substrates of moral strategies in social decision-making. Nat Commun. 2019;10:1483.
210. Baumeister RF, Stillwell AM, Heatherton TF. Guilt: an interpersonal approach. Psychol Bull. 1994;115:243–67.
211. Boyd R, Richerson PJ. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol. 1992;13:171–95.
212. DeScioli P, Kurzban R. Mysteries of morality. Cognition. 2009;112:281–99.
213. Smith A. The theory of moral sentiments. New York: Oxford University Press: 1976[1761]
214. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. The neural basis of economic decision-making in the ultimatum game. Science. 2003;300:1755–8.
215. Baumgartner T, Knoch D, Hotz P, Eisenegger C, Fehr E. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat Neurosci. 2011;14:1468–74.
216. Strobel A, Zimmermann J, Schmitz A, Reuter M, Lis S, Windmann S, et al. Beyond revenge: neural and genetic bases of altruistic punishment. Neuroimage 2011;54:671–80.
217. Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science. 2006;314:829–32.
218. Krueger F, Hoffman M. The emerging neuroscience of third-party punishment. Trends Neurosci. 2016;39:499–501.
219. Zhong S, Chark R, Hsu M, Chew SH. Computational substrates of social norm enforcement by unaffected third parties. Neuroimage. 2016;129:95–104.
220. Civai C, Huijsmans I, Sanfey AG. Neurocognitive mechanisms of reactions to second- and third-party justice violations. Sci Rep. 2019;9:9271.
221. Corradi-Dell'Acqua C, Civai C, Rumiati RI, Fink GR. Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc Cogn Affect Neur. 2013;8:424–31.
222. Stallen M, Rossi F, Heijne A, Smidts A, Dreu CKWD, Sanfey AG. Neurobiological mechanisms of responding to injustice. J Neurosci. 2018;38:2944–54.
223. Zinchenko O. Brain responses to social punishment: a meta-analysis. Sci Rep. 2019;9:12800.
224. Bellucci G, Camilleri JA, Iyengar V, Eickhoff SB, Krueger F. The emerging neuroscience of social punishment: meta-analytic evidence. Neurosci Biobehav Rev. 2020;113:426–39.
225. Glass L, Moody L, Grafman J, Krueger F. Neural signatures of third-party punishment: evidence from penetrating traumatic brain injury. Soc Cogn Affect Neur. 2016;11:253–62.
226. Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, et al. The neural correlates of third-party punishment. Neuron 2008;60:930–40.
227. Treadway MT, Buckholtz JW, Martin JW, Jan K, Asplund CL, Ginther MR, et al. Corticolimbic gating of emotion-driven punishment. Nat Neurosci. 2014;17:1270–5.
228. Ginther MR, Bonnie RJ, Hoffman MB, Shen FX, Simons KW, Jones OD, et al. Parsing the behavioral and brain mechanisms of third-party punishment. J Neurosci. 2016;36:9420–34.
229. Patil I, Calò M, Fornasier F, Cushman F, Silani G. The behavioral and neural basis of empathic blame. Sci Rep. 2017;7:5200.
230. Gert B. Common morality: deciding what to do. New York, NY: Oxford University Press; 2004.
231. Clark MS, Mils J. The difference between communal and exchange relationships: what it is and is not. Pers Soc Psychol B. 1993;19:684–91.
232. Preuss TM, Goldman-Rakic PS. Myelo- and cytoarchitecture of the granular frontal cortex and surrounding regions in the strepsirhine primate Galago and the anthropoid primate Macaca. J Comp Neurol. 1991;310:429–74.
233. Neubert F-X, Mars RB, Thomas AG, Sallet J, Rushworth MFS. Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. Neuron. 2014;81:700–13.
234. Koechlin E. Frontal pole function: what is specifically human? Trends Cogn Sci. 2011;15:241.
235. Vaccaro AG, Fleming SM. Thinking about thinking: a coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. Brain Neurosci Adv. 2018;2:2398212818810591.
236. Rouault M, McWilliams A, Allen MG, Fleming SM. Human metacognition across domains: insights from individual differences and neuroimaging. Personal Neurosci. 2018;1:e17.
237. Domenech P, Koechlin E. Executive control and decision-making in the prefrontal cortex. Curr Opin. Behav Sci. 2015;1:101–6.
238. Meyniel F, Sigman M, Mainen ZF. Confidence as Bayesian probability: from neural origins to behavior. Neuron. 2015;88:78–92.
239. Tsujimoto S, Genovesio A, Wise SP. Evaluating self-generated decisions in frontal pole cortex of monkeys. Nat Neurosci. 2010;13:120–6.
240. Tsujimoto S, Genovesio A, Wise SP. Frontal pole cortex: encoding ends at the end of the endbrain. Trends Cogn Sci. 2011;15:169–76.
241. Miyamoto K, Setsuie R, Osada T, Miyashita Y. Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. Neuron. 2018;97:980. e6
242. Semendeferi K, Armstrong E, Schleicher A, Zilles K, Hoesen GWV. Prefrontal cortex in humans and apes: a comparative study of area 10. Am J Phys Anthropol. 2001;114:224–41.
243. Mansouri FA, Koechlin E, Rosa MGP, Buckley MJ. Managing competing goals — a key role for the frontopolar cortex. Nat Rev Neurosci. 2017;18:645–57.
244. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010;329:1541–3.
245. Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain. 2014;137:2811–22.
246. Martino BD, Fleming SM, Garrett N, Dolan RJ. Confidence in value-based choice. Nat Neurosci. 2013;16:105–10.
247. Boorman ED, Behrens TEJ, Woolrich MW, Rushworth MFS. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron. 2009;62:733–43.
248. Boorman ED, Behrens TE, Rushworth MF. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. Plos Biol. 2011;9:e1001093.
249. Fleming SM, Putten EJ, van der, Daw ND. Neural mediators of changes of mind about perceptual decisions. Nat Neurosci. 2018;21:617–24.
250. Miyamoto K, Trudel N, Kamermans K, Lim MC, Lazari A, Verhagen L, et al. Identification and disruption of a neural mechanism for accumulating prospective metacognitive information prior to decision-making. Neuron. 2021;109:1396–408.
251. Soutschek A, Ugazio G, Crockett MJ, Ruff CC, Kalenscher T, Tobler PN. Binding oneself to the mast: stimulating frontopolar cortex enhances precommitment. Soc Cogn Affect Neur. 2016;12:635–42.
252. Bang D, Fleming SM. Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci. 2018;115:6082–7.
253. Wittmann MK, Kolling N, Faber NS, Scholl J, Nelissen N, Rushworth MFS. Self-other mergence in the frontal cortex during cooperation and competition. Neuron. 2016;91:482–93.
254. Drummond N, Niv Y. Model-based decision making and model-free learning. Curr Biol. 2020;30:R860–65.
255. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron 2013;80:312–25.
256. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8:1704–11.
257. O'Doherty JP, Lee S, Tadayonnejad R, Cockburn J, Iigaya K, Charpentier CJ. Why and how the brain weights contributions from a mixture of experts. https://psyarxiv.com/ns6kq/. 2020.
258. Fetsch CR, DeAngelis GC, Angelaki DE. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci. 2013;14:429–42.
259. Seilheimer RL, Rosenberg A, Angelaki DE. Models and processes of multisensory cue combination. Curr Opin Neurobiol. 2014;25:38–46.
260. Ma WJ, Pouget A. Linking neurons to behavior in multisensory perception: a computational review. Brain Res. 2008;1242:4–12.
261. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron 2014;81:687–99.

262. Weissengruber S, Lee SW, O'Doherty JP, Ruff CC. Neurostimulation reveals context-dependent arbitration between model-based and model-free reinforcement learning. Cereb Cortex. 2019;29:4850–62.
263. Charpentier CJ, Iigaya K, O'Doherty JP. A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. Neuron. 2020;106:687. e7
264. Diaconescu AO, Stecy M, Kasper L, Burke CJ, Nagy Z, Mathys C, et al. Neural arbitration between social and individual learning systems. Elife. 2020;9:e54051.
265. Park SA, Sestito M, Boorman ED, Dreher J-C. Neural computations underlying strategic social decision-making in groups. Nat Commun. 2019;10:5287.
266. Bang D, Ershadmanesh S, Nili H, Fleming SM. Private–public mappings in human prefrontal cortex. Elife. 2020;9:e56477.
267. Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW. Mapping cortical change across the human life span. Nat Neurosci. 2003;6:309–15.
268. Shaw P, Kabani NJ, Lerch JP, Eckstrand K, Lenroot R, Gogtay N, et al. Neurodevelopmental trajectories of the human cerebral cortex. J Neurosci. 2008;28:3586–94.
269. Travis K, Ford K, Jacobs B. Regional dendritic variation in neonatal human cortex: a quantitative golgi study. Dev Neurosci-Basel. 2005;27:277–87.
270. Marshall J, Mermin-Bunnell K, Bloom P. Developing judgments about peers' obligation to intervene. Cognition. 2020;201:104215.
271. Marshall J, Wynn K, Bloom P. Do children and adults take social relationship into account when evaluating people's actions? Child Dev. 2020;91:e1082–e1100.
272. Trommershäuser J, Kording K, Landy MS. Sensory cue integration. New York, NY: Oxford University Press; 2011.
273. Crockett MJ. Models of morality. Trends Cogn Sci. 2013;17:363–66.
274. Cushman F. Action, outcome, and value. Pers Soc Psychol Rev. 2013;17:273–92.
275. Nelson TO. Metamemory: a theoretical framework and new findings. Psychol Learn Motiv. 1990;26:125–73.
276. Frith CD, Frith U. Interacting minds–a biological basis. Science 1999;286:1692–5.
277. Tomasello M, Carpenter M. Shared intentionality. Dev Sci. 2007;10:121–5.

AUTHOR CONTRIBUTIONS

YZ, SWCC, and MJC conceived and planned the paper. YZ and MJC wrote the paper with edits from SWCC.

FUNDING AND DISCLOSURE

This work was supported by the National Institute of Mental Health (MJC: 1R21MH124093-01; SWCC: R01MH110750 and R01MH120081). YZ was supported by a Doctoral Study Abroad Fellowship from the Korea Foundation for Advanced Studies. The authors declare no competing interests.

GLOSSARY

Cooperation
any behavior that is potentially costly to an individual but benefits at least one other individual [8].

Self-regulation
adjusting one's inner states or behaviors according to personal goals, expectations or standards [34].

Metacognition
ability to monitor, assess and orchestrate one's own cognitive processes and their quality for the guidance of behavior[48, 50, 53, 275].

Precommitment
voluntary restriction of access to temptations on the basis of a metacognitive insight that one's own self-regulation is likely to fail [40–46].

Mentalizing
capacity to infer and represent mental states (i.e., desires, intentions and beliefs) of oneself and others and thereby better predict and regulate future behaviors [276].

Shared intentionality
ability to build a common understanding of joint commitment to a collective goal with others to engage in cooperative acts and to regulate individual desires when they conflict with the collective interest [277].

Social norm
commonly known rules or standards of behavior that are based on widely shared views about how individual group members ought to behave in a given situation [84].

Cooperative norm
a type of social norm that facilitates cooperation [84, 86, 87].

Norm compliance
adoption of behaviors constrained by normative considerations: pursuing prescribed behaviors and avoiding proscribed behaviors according to social norms [198].

Norm enforcement
inducing others to obey a social norm through sanction in the event of norm violations [86].

Norm arbitration
flexible selection of norms to guide one's behavior in a context-appropriate way, especially when the situation is ambiguous and no predominant norm is inferred.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to M.J.C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.