COMMENTARY

The roots of modern justice: cognitive and neural foundations of social norms and their enforcement

Joshua W Buckholtz & René Marois

Among animals, *Homo sapiens* is unique in its capacity for widespread cooperation and prosocial behavior among large and genetically heterogeneous groups of individuals. This ultra-sociality figures largely in our success as a species. It is also an enduring evolutionary mystery. There is considerable support for the hypothesis that this facility is a function of our ability to establish, and enforce through sanctions, social norms. Third-party punishment of norm violations ("I punish you because you harmed him") seems especially crucial for the evolutionary stability of cooperation and is the cornerstone of modern systems of criminal justice. In this commentary, we outline some potential cognitive and neural processes that may underlie the ability to learn norms, to follow norms and to enforce norms through third-party punishment. We propose that such processes depend on several domain-general cognitive functions that have been repurposed, through evolution's thrift, to perform these roles.

Human beings are the most richly social creatures that our planet has ever known. Although other species are well known for their high level of social organization, none share our capacity for stable large-scale cooperation between genetically unrelated individuals. This unique feature of human culture is made possible by cognitive capacities that permit us to establish, transmit and enforce social norms[1,2]. Social norms—widely shared sentiments about what constitutes appropriate behavior—comprise a basic "grammar of social interaction"[3]: sets of prescribed and proscribed rules that serve to foster social peace, stabilize cooperation and enhance prosperity. These norms take a variety of forms, ranging from highly culturally specific standards of behavior that lack a strong moral valence ("Thou shalt not wear white after Labor Day"), to more universal norms possessed of a moral valence that varies in magnitude by culture ("Thou shalt not commit adultery"), to norms that possess such a universally strong moral valence and such widespread agreement about the necessity of their compliance that they are formalized and codified into laws ("Thou shalt not kill"). Specific prosocial norms that operate to promote cooperation include prohibitions against physical harm, taking of property through theft or intimidation, and cheating in economic exchanges, as well as resource distribution norms pertaining to equity and fairness.

Although the idea that social norms promote cooperation is not new, the specific mechanisms through which social norms act to induce prosociality have only recently been a target of scientific inquiry. In this commentary, we synthesize behavioral, cognitive and neuroscientific data on norm enforcement to propose that the ability of human culture to create modern institutions of justice—a crucial force for social stability that is chiefly characterized by fair and impartial enforcement of widely endorsed norms of moral conduct—is enabled by the evolutionary elaboration of domain general cognitive processes for value learning, decision making and perspective taking.

Joshua W. Buckholtz is in the Department of Psychology and Center for Brain Science, Harvard University, Cambridge, and the Department of Psychiatry, Psychiatric Neuroimaging Division, Massachusetts General Hospital, Boston, Massachusetts, USA, and René Marois is in the Department of Psychology and Center for Integrated and Cognitive Neuroscience, Vanderbilt University, Nashville, Tennessee, USA.
e-mail: rene.marois@vanderbilt.edu or joshuabuckholtz@fas.harvard.edu

The mystery of human cooperation

That widespread cooperation exists even in the face of significant incentives to behave selfishly is an enduring evolutionary mystery. Several models have been proposed to explain this distinguishing feature of human behavior.

Kin selection theory posits that non–self-interested (altruistic) behavior among related individuals occurs because individuals will accept costs when the net result is an increased likelihood of transmitting shared genes to future generations[4]. However, this theory has been challenged on the grounds that it can't account for cooperation among individuals who do not share genes, as is common in human culture. Direct reciprocity (reciprocal altruism) suggests that cooperation in bilateral interactions, even when initially costly, is incentivized owing to the selfish benefits that may be accrued in the long-term ("I scratch your back now, you scratch my back later"). But cooperation under direct reciprocity models is only evolutionarily stable in small groups (<10); empirical data suggests that natural selection wouldn't favor cooperation by reciprocal altruism among unrelated individuals on the

scale of human culture[5]. Theories of indirect reciprocity focus instead on the self-interest that is served by accruing a good reputation through altruistic behavior. Though reputational enhancement clearly has a role in human cooperation, some have suggested that indirect reciprocity alone is unlikely to account for the widespread nature of human cooperation, where one-shot (unrepeated) interactions are common and attendant reputational benefits likely to be small[6].

Strong reciprocity: norm enforcement through punishment

While reciprocity and kin-based models appear inadequate to account for the natural selection of cooperation in humans, a compelling alternative has been proposed. According to the theory of strong reciprocity, long-term widespread cooperation is made possible by the presence of "strong reciprocators": individuals who reward norm-followers (for example, cooperators) and punish norm-violators (for example, defectors) even when such actions are costly, and in the absence of any material future gain for the strong reciprocator. Evolutionary models support the idea that strong reciprocity can maintain cooperation among genetically unrelated individuals[6], and studies in primates show that chimpanzees will sanction (punish) conspecifics who have directly harmed them, suggesting phylogenetic origins for norm-enforcement behavior[7]. Data from behavioral economic studies provide empirical support for the critical role of punishment in enforcing norms of distributional fairness and cooperation. For example, in the ultimatum game, a 'proposer' is endowed with a certain amount of money to split with a 'responder' however she chooses. If the responder accepts the proposer's offer, then both individuals keep the money according to the proposed split; if the responder rejects the offer, neither party receives any money. Although it would always be in the responder's rational self-interest to accept any offer greater than zero, low ('unfair') offers are frequently rejected by responders. In effect, the responder is punishing the proposer for violating fairness norms, even though it is costly for them to do so. This form of second-party norm enforcement is evident in societies all over the world. Across societies, variation in the willingness to engage in costly punishment predicts inter-society differences in altruistic behavior[8]. Both findings are consistent with recent evolutionary models of norm-enforcement[1,8].

Experimental work using public goods games demonstrates the extent to which 'altruistic punishment' of norm-violators is crucial in stabilizing cooperation. In each round of these games, participants have the opportunity to behave selfishly (keep money for themselves) or prosocially (contribute money to a common account). Moderate levels of cooperation (that is, transfers to the common account) are present initially and then decay rapidly as individuals begin to defect, and most participants behave selfishly by the last round. However, when participants are allowed to sanction other participants for defecting, they do so avidly, even when the opportunity to punish comes at a cost (altruistic punishment). When punishment is possible, contributions to the common account increase markedly, leading to almost complete cooperation by the last round of the experiment[9]. Thus, in the absence of punishment, cooperation is impossible to maintain, and even individuals who are initially predisposed to cooperate will begin to behave antisocially. By contrast, selfish individuals can be incentivized to cooperate when there is the threat of punishment by strong reciprocators.

Neurocognitive foundations of norm-based cooperation

As detailed above, widespread norm-compliance in humans is contingent on the willingness of individuals to sanction behavior that deviates from widely agreed-upon norms. This ability to punish in this manner implies the presence of a set of mental faculties that, together, may form a cognitive foundation for norm-based cooperation. Strong reciprocity requires that individuals have the capacity to learn norms; integrate predictions about norm-related action outcomes into decision making to guide their own behavior; assess other individuals' beliefs, desires and behavior in the context of these norms; and use subjective responses to norm violations to appropriately sanction defection. However, it seems unlikely that *H. sapiens* evolved specific neural pathways or cognitive modules devoted exclusively to norm compliance and enforcement[10]. Rather, we suggest that this unique, and uniquely successful, aspect of human culture is enabled by the elaboration of brain systems that support basic or domain-general cognitive processes, which have been used over evolution to 'bootstrap' more specific cognitive mechanisms for adaptive norm-related behaviors. In the sections that follow, we propose a potential neurobiological architecture that may underpin norm learning, norm compliance and norm enforcement.

Norm learning and compliance

The widespread propagation and consistent intergenerational transmission of beliefs about appropriate behavior highlight the importance of social learning processes[11] in promoting norm-compliant behavior. Social learning involves integrating information about others' beliefs, goals, actions and outcomes into self-relevant reinforcement learning algorithms. Recent evidence from human neuroimaging demonstrates the involvement of mesolimbic circuitry, particularly the ventromedial prefrontal cortex (vmPFC) and ventral striatum, in the generation of these observational learning signals[12]. This circuitry is crucial for basic forms of reward learning (for example, representing reward value, learning stimulus–value associations and acquiring predictive value representations) in humans, nonhuman primates and rodents[13], raising the possibility that the capacity for social norm transmission in human culture developed by the elaboration of basic reinforcement learning processes mediated by pre-existing neural circuitry.

Norm compliance generally, and cooperative behavior specifically, is diminished in individuals with vmPFC damage, as is the level of guilt experienced for violating cooperation norms[14]. This is particularly interesting given the suggestion that guilt is a "moral emotion"; specifically, one that is elicited by one's own violation of learned moral norms, and that serves to facilitate the production of prosocial behavior[15]. These findings are consistent with the idea that brain systems that originally evolved to handle a basic and survival-critical cognitive process, value learning, can be adapted to facilitate the learning of higher-order action values (moral norms), an essential component of norm-based cooperation in *H. sapiens*[2,16].

Above, we reviewed evidence that the threat of punishment induces prosocial behavior, incentivizing cooperation in people who are otherwise inclined to defect. In contrast to the ventral frontostriatal network that we believe may be involved in norm learning, neuroimaging data suggest that a dorsal frontostriatal circuitry is essential for integrating information about sanction threats into decision making to incentivize norm-compliant behavior. Spitzer and colleagues[17] scanned participants (Player A) while they made decisions about how much of a monetary endowment to split with another, anonymous participant (Player B). On some trials, Player B was permitted to punish Player A (at a cost to themselves) if they deemed the amount of the split to be unfair. During trials in which Player A faced the threat of punishment, they transferred substantially more money to Player B, indicating enhanced compliance to fairness norms. Notably, the change in transfer amount between punishment and

no punishment conditions was positively correlated with condition differences in functional magnetic resonance imaging (fMRI) signal in a frontostriatal circuit comprising, in part, the dorsal striatum and right dorsolateral prefrontal cortex (DLPFC). Individuals exhibiting greater sensitivity in this circuit to the threat of punishment showed the largest increases in transfer amount, indicating greater norm compliance[17].

Dorsal striatum receives inputs from DLPFC and the dopaminergic midbrain, and transmits outputs to primary motor cortex. It is thus in a key position to guide action selection based on predicted action values. It has been suggested that dorsal striatum receives an array of response options and goal representations from DLPFC and uses learned action–outcome associations to bias response selection in favor of the action that has the best probability of maximizing reward[18]. Further, interactions between dorsal caudate and DLPFC are thought to be particularly salient for the guidance of behavior according to long-term rewards[19], consistent with the ability of DLPFC to maintain stable goal representations over time[20]. It is possible that the threat of punishment may increase cooperation by biasing reward-related action selection mechanisms mediated by dorsal frontostriatal circuitry. In this context, punishment threat may change the reinforcement contingencies associated with potential responses, and increased corticostriatal fMRI signal in the punishment condition may reflect this updating process.

Interestingly, disrupting DLPFC function with transcranial magnetic stimulation (TMS) decreases norm compliance during a cooperative task (the trust game), impeding participants' ability to form positive reputations for cooperation[21]. This result has been interpreted as reflecting the diminished capacity of DLPFC to override a prepotent response to behave selfishly following TMS. Alternatively, TMS may alter cooperative behavior by narrowing the response repertoire maintained in DLPFC, by interfering with the transmission to dorsal striatum of DLPFC goal representations, or by disrupting the integration of DLPFC response options and goals with dorsal striatal action value predictions. All of these may disrupt dorsal frontostriatal control over action preparation for a long-term social reward (reputation) derived from norm compliance.

The research reviewed above is consistent with the hypothesis that the evolution of norm-based decision making may have been facilitated by the presence of neural circuitry that supports domain-general cognitive processes for value-based action selection. Such processes may have been co-opted and expanded to operate in the social domain, where they promote action selection according to higher-order action values linked to social rewards, such as reputation and trust, thereby facilitating cooperation.

Neural basis of second-party norm enforcement

The costly punishment of anonymous defectors in one-shot interactions is no less curious for it being so common. What motivates people to accept a significant personal cost for the opportunity to retaliate against those who have treated them unfairly? It has been suggested that negative emotional responses toward norm violators drive costly punishment in two-party interactions, and neuroimaging investigations have shed light on the relevant neurobiology. In an influential study, rejection of unfair offers in the ultimatum game was found to be associated with enhanced activity in right DLPFC and in insular cortex, where the magnitude of activation was associated with the probability of rejection[22]. The insula is thought to contribute to the visceral experience of negative emotions by representing aversive interoceptive states, and may therefore represent a negatively valenced affective signal that biases response options maintained in DLPFC[22]. Further supporting a causal role for negative affective biasing signals in second-party norm-enforcement is the recent finding that selective pharmacological attenuation of amygdala activation during the ultimatum game reduces the rejection of unfair offers[23].

Together, these data suggest that negative emotional responses lead individuals to punish those who have treated them unfairly, even though this punishment is costly. However, an alternative view proposes that negative emotion follows as a consequence of moral judgment rather than serving as its source[24]. By extension, this model would predict that the experience of negative emotion in the ultimatum game is a consequence of a moral judgment that drives both the rejection and the emotional response.

Third-party norm enforcement and evolution of stable societies

It is noteworthy that all of the studies highlighted above, and most published brain imaging studies of norm enforcement, deal with second-party punishment ("you harm me, I punish you"). However, sanctioning of norm violations in modern societies is primarily achieved through the punishment of norm-violators by impartial, state-empowered enforcers (third-party punishment). Behavioral economics work has shown that individuals will accept costs to sanction individuals who have violated fairness and distribution norms even when they were not directly affected by the norm violation[1,11]. This form of punishment is particularly important given evidence that the development of stable social norms in human societies specifically required the evolution of third-party sanction systems[25]. Norm compliance in modern societies cannot alone be maintained by the self-interest that is often served by cooperation, nor can it thrive with a two-party retaliatory system beyond the small-scale level[26]. Given this, the development of stable large societies hinged on the ability to involve impartial third parties to evaluate and sanction harms[25–27]. Laboratory experiments have shown that third parties will bear costs to punish defectors even though the defection did not materially harm them[28], and field studies support these findings[8].

The stability of modern, large-scale societies therefore depends on the ability and willingness of disinterested third-parties—impartial decision makers who were not directly affected by the norm violation and who derive no direct benefits from its sanction—to enforce moral norms through punishment. The codification of these norms into laws, and the attendant establishment of state-administered systems of criminal justice that are charged with norm compliance, is arguably one of the most important developments in human culture. In many modern criminal justice systems, the deciding parties consist of jurors and judges, who evaluate evidence, determine guilt or innocence and arrive at a punishment that, ideally, accords with citizens' intuitions about the severity of the norm violation[29]. In courts of law, impartiality—expressed in the maxim *Judex non potest injuriam sibi datum punier* ("A judge cannot punish a wrong done to himself")—is a foundational principle that guides the proceedings. Only by being uninvolved third parties to the presumed criminal act can judges and juries ensure that anyone accused of a crime receives a fair trial.

Cognitive and neural mechanisms of third-party norm enforcement

In modern systems of criminal justice, an individual will typically be convicted of a crime if the state proves beyond a reasonable doubt that (i) he committed a prohibited act (*actus reus*) and (ii) the act was accompanied by a bad or guilty intent (*mens rea*)[30]. The amount of punishment imposed will be affected by both the intent of the accused and the severity of the harm that he caused (or intended to cause)[29]. This implies that third-party decision makers

must be equipped with cognitive mechanisms that permit (i) evaluation of the criminal act and the mental state of the criminal actor, (ii) evaluation of the harm caused by that actor, (iii) integration of these evaluations with representations of relevant legal codes (for example, sentencing guidelines) and internal motivations for punishment, and (iv) action selection from among an array of punishment response options. In this section, we will outline potential cognitive and neurobiological underpinnings for these mechanisms (**Fig. 1**).

Evaluation of mental states

Whereas determination of *actus reus* is usually fact-based, guilty intent, and therefore blameworthiness, is more difficult to establish because it refers to a subjective state of mind. Further, the law distinguishes between different levels of intent, varying from acting "purposely" to acting "knowingly" to acting "recklessly" to acting "negligently." The blameworthiness of an accused, and hence his assigned punishment, is modified according to these differing standards of intent (though the extent to which lay people—and, by extension, jurors—can meaningfully distinguish between these legally distinct mental states has been questioned[31]). The determination of blameworthiness is also affected by the presence of mitigating circumstances (for example, duress or psychosis) that bear directly on a defendant's state of mind. Such factors must therefore be taken into full account when determining whether the accused possessed a blameworthy state of mind at the time of the criminal act.

On the basis of the proposed sequence of cognitive processes outlined above, it follows that third-party punishment should begin with an assessment of the criminal act and an inference of the mental state of the defendant. Mitchell and colleagues have argued that we evaluate the state of mind of others through a process of "self-projection," whereby another's mental state is inferred ("simulated") by referencing it to a projection of our own state in that situation. Self-projective "mentalizing" engages a core network that includes the medial prefrontal cortex (mPFC), temporo-parietal junction (TPJ) and posterior cingulate[32]. Given the need for mentalizing in assessing the blameworthiness of accused offenders, this network should be strongly recruited during third-party punishment. The TPJ in particular may be key in this process, as it has been shown to be important for belief attribution in moral judgments[33].

Evaluation of harm

In legal contexts harm can take many forms, ranging from depriving someone of their property (theft) to depriving someone of their health (assault and battery) or their life (murder). Such harms, particularly physical harms, may be processed in a manner that is similar to processing of environmental threats. Consistent with this, brain regions that are important for threat detection and the generation of aversive emotional states (for example, amygdala) are active when participants process information about bodily harm during moral judgments tasks[34]. The engagement of neural systems for threat detection may engender negative affective arousal, which may be used to guide intuitions about how much punishment is deserved for a given harm. Negative emotional states have been shown to affect legal decision making, both directly and indirectly (that is, even when they are irrelevant to the context of the crime)[35]. One explanation for this phenomenon, the 'affect-as-information' model, suggests that individuals use their emotional state as a source of information when making decisions. Rather than making a calculated judgment based on factual information about the case, individuals may use subjective emotional tone as a heuristic device to guide punishment[36]. Neural circuitry involved in threat detection and in regulating affective tone, such as amygdala and vmPFC, may therefore play a role in translating harm severity into punishment severity by influencing an individual's level of affective arousal.

Integration of mental state and harm information for punishment selection

A mentalizing-based evaluation of blameworthiness may be conjoined with affective heuristics related to harm severity to create a rough 'intuition' of deserved punishment. To select a specific punishment response, this intuitive judgment must be further integrated with information about the specific set of available punishment options in a particular context. Given the high need for integration among several areas inherent in this process, heteromodal association 'hub' regions such as mPFC may be key. mPFC has strong anatomical and functional connectivity to other mentalizing network nodes, including the temporo-parietal junction[37], as well as to amygdala and DLPFC[38]. We speculate that DLPFC, possibly in concert with intraparietal sulcus (a region that is often coactivated with the lateral prefrontal cortex[38] and which is known to be involved in representing ordinal 'number lines'[39]) may translate rough intuitions about deserved punishment into a precise punishment response by anchoring it to a context-specific punishment scale. As such, DLPFC may represent a final output node that is crucial for selecting a context-appropriate punishment response from among competing response options. This would be consistent with data indicating that disrupting DLPFC function with TMS selectively impairs the ability to make punishment decisions in second-party contexts, but does not affect the (presumably antecedent) evaluation of fairness[40].

In sum, the capacity for third-party punishment may be grounded in several domain-general cognitive processes—self-projection (mentalizing), threat detection, scale representation, rule selection and action selection—that have been co-opted and elaborated to enable this important force for social stability. In the section that follows, we detail results from a brain imaging experiment in which we sought to test the neurocognitive hypothesis of third-party punishment outlined above.

Testing a neurocognitive hypothesis of third-party norm enforcement

We used fMRI to identify neural activity associated with determining blameworthiness and assigning punishment during third-party punishment decisions about criminal offenders (that is, legal decision making)[41]. Specifically, participants read scenarios that depicted criminal violations and indicated, on a scale of 0 (no punishment) to 9 (extreme punishment), how much punishment the scenario protagonist should receive for his actions against another party. The scenarios ranged in harm severity from petty theft to rape and murder. They also varied with respect to criminal responsibility: some scenarios (responsibility scenarios; R) depicted prototypical criminal behavior for which intent was clear (for example, purposefully stealing an item), whereas other scenarios (diminished-responsibility scenarios; DR) involved similar harms, but with added details that might mitigate culpability and reduce blameworthiness (for example, mental illness or duress). Pairs of R and DR scenarios were matched for harm severity and differed only on the presence or absence of exculpating or mitigating information. Consistent with expectation, participants assigned less punishment and blameworthiness to the protagonist, and also reported less subjective arousal, in DR scenarios.

Consistent with the involvement of mental state attribution in third-party punishment decision-making, bilateral TPJ was robustly activated across all conditions, though it was most strongly engaged by the DR scenarios. As the scenario protagonists' mental state is most ambiguous in the DR scenarios, participants should engage mentalizing processes more strongly in these scenarios to resolve this

ambiguity and arrive at an intuitive evaluation of blameworthiness. Examination of the time course of TPJ activity revealed that it began early and peaked before the participant's actual decision, as might be expected for a region that is involved in making an antecedent evaluation that is necessary for a determination of appropriate punishment.

Harm evaluation was assessed by using participants' punishment magnitude as a proxy for harm severity, as punishment ratings increased linearly according to ranked category of harm (that is, taking of property by theft < taking of property by force < physical assault < grievous bodily harm < homicide). We found that activity in several regions, particularly amygdala, posterior cingulate and mPFC, correlated with the magnitude of punishment assigned to the protagonist in R scenarios, as well as with self-reported negative arousal. This self-reported arousal mediated the relationship between brain activity and punishment scores, confirming that participants used their negative affect as a heuristic to gauge appropriate punishment (J.W.B. and R.M., unpublished data). Notably, engagement of these regions did not merely reflect the fact that high harm (for example, rape or murder) scenarios were inherently more arousing than low harm (for example, theft) scenarios. For the regions above, responsibility condition differences (R versus DR) in fMRI signal for each harm severity–matched scenario pair predicted the difference in assigned punishment between those matched scenarios. Thus, these regions demonstrate joint sensitivity to both harm severity and blameworthiness. Though the experimental design did not permit a decisive test, this result is consistent with the idea that mPFC and posterior cingulate hub regions integrate intuitions about mental states (derived from self-projective mentalizing) and harm severity (gleaned from affective arousal) to arrive at a rough sense of deserved punishment magnitude.

The final tenet of our hypothesis was that there should be brain regions that translate information about blame and harm into a precise punishment decision on the basis of a context-specific scale of punishment. We posited that such regions would be most engaged during scenarios—either R or DR—in which participants actually made a decision to punish. The right DLPFC and bilateral intraparietal sulcus were not only found to be more engaged in R than DR scenarios, but activity in these regions was also greater in DR scenarios for which participants chose to assign punishment than in DR scenarios in which participants assigned no punishment. The observed pattern of activity resembles that predicted for regions that are involved in selecting a response from among an array of possible punishment options. This final decision about deserved punishment magnitude may be based on biasing signals conveyed from the regions highlighted above (for example, mPFC).

On the whole, these findings are consistent with a model of norm enforcement behavior wherein outputs from domain-general cognitive processes in TPJ and amygdala produce 'blame' and 'harm' signals, which are integrated by posterior cingulate and mPFC hubs and used to bias action selection mechanisms in DLPFC (**Fig. 1**). However, although these results suggest an ordered sequence of brain activity during norm enforcement, these cognitive processes may not necessarily proceed in a topically and temporally segregated manner. Indeed, there is evidence that in addition to affecting punishment decisions directly[42], emotional reactions to harm may influence punishment reactions by affecting how people attribute blame[42] and moral responsibility[43]. Thus, it may not always be the case that judgments of causality, responsibility, harm assessment and punishment proceed in a sequential and independent fashion.

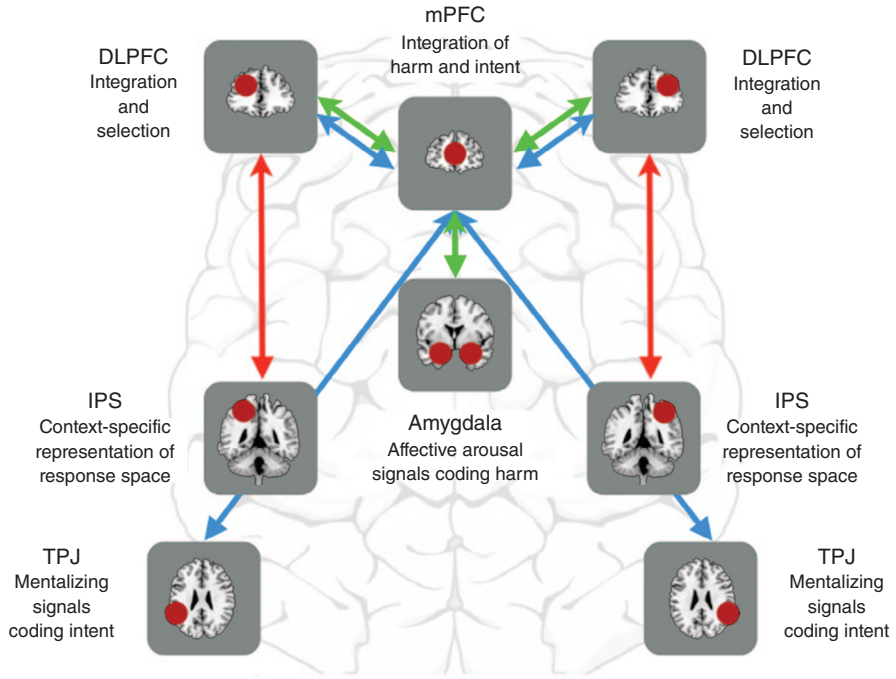

Figure 1 A neurocognitive hypothesis for third-party punishment behavior. According to this hypothesis, the TPJ encodes information about criminal intent and blameworthiness, extracted from mentalizing computations that assess an accused's state of mind. The amygdala may generate an affective arousal signal based on the magnitude of the accused's harm, which may be used as a heuristic to guide punishment severity. mPFC may integrate intent representations from TPJ and harm magnitude signals from the amygdala, and this information is conveyed to DLPFC. We hypothesize that DLPFC receives a multiplexed intent and harm signal from mPFC that is integrated with activity in the intraparietal sulcus (IPS), which may maintain a context-specific representation of the response space that is used to construct a scale of punishment. Together, these inputs could bias selection among an array of competing punishment response options by DLPFC. Bidirectional arrows emphasize the interactive computations likely to take place along the core components of this network. This hypothesis presumes a 'just desserts' (retributivist) motivation for punishment, which is supported by data[29]. The model would require modification to account for punishment enacted from utilitarian or consequentialist concerns. Long red arrows, scale representation information from IPS; blue arrows, mentalizing output from TPJ; short green arrows, affective arousal signals from the amygdala.

The universality of punishment and its neurobiological origins

As discussed earlier, the urge to punish is common to both third-party and second-party interactions, although it may be stronger in the latter case[1]. There is consistent evidence that intuitions of justice are shared across societies and cultures[44] (though intercultural differences do, of course, exist[45]). Cross-cultural data indicate that costly second- and third-party punishment is practiced across the globe, from Tanzanian villages and Amazonian rainforests to American college campuses[8]. Worldwide, the degree of altruism present in a society is

directly relatable to its punishment practices: when norms are enforced, cooperation follows[8]. Taken together, this evidence suggests that the mechanisms of third-party punishment may be deeply ingrained in *H. sapiens*. We have suggested that the evolutionary elaboration of domain-general cognitive processes is a driving force in the development of increasingly complex forms of norm-enforcement behavior. In this light, the consistency of brain activation patterns across moral judgment, second-party punishment and third-party punishment is striking. In particular, the prefrontal cortex area activated in our third-party legal decision-making task corresponds very well to the prefrontal region engaged during norm enforcement behavior in two-party economic exchanges[22,40] and norm compliance[17], suggesting that this region of the prefrontal cortex may serve a common core function that enables the practice of norm-related behaviors. This is not to suggest that this area is specifically and exclusively devoted to such behaviors; indeed, the same brain region has been observed in a wide variety of higher-level functions, including working memory, inhibitory control and rule-guided response selection[20]. Thus, a more likely and parsimonious explanation is that the basic cognitive functions subserved by this region have been adapted to serve this role in norm-relevant contexts. A natural question is, of course, what are those functions?

The role of DLPFC in norm enforcement

One proposed role of the prefrontal cortex in norm enforcement is that its involvement reflects cognitive control: namely, the inhibition of a prepotent response to behave selfishly[40]. This would explain why the DLPFC is activated when people refuse to accept an unfair economic deal in the ultimatum game for the sake of punishing the proposer for his unfairness (even though that punishment is costly), and why disruption of DLPFC activity with TMS reduces second-party altruistic punishment behavior[40]. Similarly, the engagement of this region during norm compliance has been proposed to signal the need to inhibit prepotent selfish responses after an individual has received a warning that defection will be sanctioned[17]. However, the cognitive control hypothesis is less consistent with the finding that this brain region is more activated when participants decide to punish protagonists in third-party interactions[41] than when they withhold doing so because of mitigating circumstances. Moreover, a recent TMS study suggests that disrupting this same region of right DLPFC during third-party punishment decreases punishment ratings when the offender is fully blameworthy, but not under conditions of diminished responsibility (J.W.B. and R.M., unpublished data). If the role of DLPFC in norm enforcement consisted of inhibiting prepotent responses, this would predict that disrupting DLPFC would increase punishment during diminished responsibility trials, when the prepotent response to punish would need to be constrained by mitigating information. Moreover, individuals with impaired cognitive control show more, not less, punishment in second-party contexts (even when it is costly to do so)[46]. Similarly, pharmacological diminution of cognitive control leads to higher, not lower, levels of altruistic punishment[47]. Such findings are difficult to reconcile with a pure cognitive control model of costly norm enforcement.

An alternative hypothesis for the role of DLPFC in punishment may be that it selects a specific response from among possible response options by integrating information about harm and blame with context-specific rules about how to apply this information. Accordingly, disruption of DLPFC may interfere with this process by impairing response selection mechanisms of the DLPFC, by interfering with biasing inputs from mPFC, or by altering the integration of context-specific response space representations maintained in intraparietal sulcus. This proposed 'integration-and-selection' function of right DLPFC seems to apply equally well to situations where the motive for punishment is the violation of a fairness norm in dyadic economic exchange or when responding to the violation of a codified moral norm (law) in a disinterested third-party context.

The integration-and-selection hypothesis of prefrontal cortex function in norm enforcement behavior does not imply that cognitive control over prepotent behavior isn't a crucial component of norm compliance and enforcement. However, we suggest that it may not be the sole underlying process that accounts for the engagement of DLPFC across norm compliance, reputation building, second-party norm-enforcement and third-party norm-enforcement tasks. It is clear that discerning the specific roles of cognitive control and impulsivity in typical and aberrant norm-related behavior is a crucial topic for further study.

Conclusions

In this commentary, we have reviewed evidence that large-scale cooperation among unrelated individuals—a defining signature of *H. sapiens* culture—is predicated on our unique ability to establish norms, to transmit these norms from generation to generation and to enforce these norms through punishment. We argue that this capacity is enabled not by specialized cognitive modules that have evolved for this purpose, but rather through evolution's thrifty repurposing of basic cognitive mechanisms, such as value learning, threat detection, self projection and response selection, that were already in place. The success of our species, made possible by strong reciprocity and its mandate of bidirectional norm enforcement, required the further elaboration of these mechanisms if we were to survive in large and genetically heterogeneous groups. Today, our welfare depends on the social order that is made possible by modern third-party systems of justice. These systems have their roots in cognitive mechanisms that originally supported fairness-related behaviors in dyadic interactions, which in turn may have developed from the basic, domain-general cognitive processes described above.

Cross-species and developmental experiments offer particularly useful insights into the phylogeny and ontology of norm enforcement behavior. Although chimpanzees do engage in costly second-party punishment, there is evidence that they will not punish third-party conspecifics for norm violations (for example, food theft). By contrast, 3-year-old human children respond strongly to third-party norm violations[48]. Together, these data raise the intriguing possibility that third-party norm-enforcement is a uniquely human behavior that appears early in human development.

Finally, although this commentary has highlighted the function of punishment in promoting large-scale cooperation in our species, it is clearly not the only factor. Positive reinforcement or reward is also an important behavior-shaping tool that incentivizes cooperation at both short- and long-term timescales. Indeed, Rand and colleagues have argued that, in certain contexts, reward actually outperforms punishment in maintaining cooperation[49]. These findings, along with the known involvement of mesocorticolimbic 'reward circuitry' (for example, striatum and vmPFC) in norm compliance and enforcement, highlight the importance of understanding the common and distinct contributions of reward and punishment in promoting cooperation and social welfare. We believe that future behavioral and neurobiological studies of bidirectional norm enforcement (rewarding norm compliance and punishing norm violation) will reveal much about the cognitive and neural architectures that enable the development of an increasingly just and peaceful society[50].

ACKNOWLEDGMENTS

We thank K. Jan for literature research and O. Jones for comments on an earlier draft of this manuscript, and the MacArthur Foundation Research Network on Law and Neuroscience for support (KK9127 and KK1031).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Fehr, E. & Fischbacher, U. *Trends Cogn. Sci.* **8**, 185–190 (2004).
2. Kitcher, P. *The Ethical Project* 422 (Harvard Univ. Press, 2011).
3. Bicchieri, C. The Grammar of Society: The Nature and Dynamics of Social Norms (Cambridge Univ. Press, 2006).
4. Hamilton, W.D. *J. Theor. Biol.* **7**, 1–16 (1964).
5. Boyd, R. & Richerson, P.J. *J. Theor. Biol.* **132**, 337–356 (1988).
6. Bowles, S. & Gintis, H. *Theor. Popul. Biol.* **65**, 17–28 (2004).
7. Jensen, K., Call, J. & Tomasello, M. *Proc. Natl. Acad. Sci. USA* **104**, 13046–13050 (2007).
8. Henrich, J. *et al. Science* **312**, 1767–1770 (2006).
9. Fehr, E. & Gächter, S. *Nature* **415**, 137–140 (2002).
10. Churchland, P.S. *Braintrust* 273 (Princeton Univ. Press, 2011).
11. Boyd, R., Richerson, P.J. & Henrich, J. *Proc. Natl. Acad. Sci. USA* **108**, 10918–10925 (2011).
12. Burke, C.J., Tobler, P.N., Baddeley, M. & Schultz, W. *Proc. Natl. Acad. Sci. USA* **107**, 14431–14436 (2010).
13. O'Doherty, J.P. *Curr. Opin. Neurobiol.* **14**, 769–776 (2004).
14. Krajbich, I., Adolphs, R., Tranel, D., Denburg, N.L. & Camerer, C.F. *J. Neurosci.* **29**, 2188–2192 (2009).
15. Haidt, J. in *Handbook of Affective Sciences* (eds. R.J. Davidson, K.R. Scherer & H.H. Goldsmith) 852–870 (Oxford University Press, 2003).
16. Seymour, B., Singer, T. & Dolan, R. *Nat. Rev. Neurosci.* **8**, 300–311 (2007).
17. Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. & Fehr, E. *Neuron* **56**, 185–196 (2007).
18. Balleine, B.W., Delgado, M.R. & Hikosaka, O. *J. Neurosci.* **27**, 8161–8165 (2007).
19. Tanaka, S.C. *et al. Nat. Neurosci.* **7**, 887–893 (2004).
20. Miller, E.K. & Cohen, J.D. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
21. Knoch, D., Schneider, F., Schunk, D., Hohmann, M. & Fehr, E. *Proc. Natl. Acad. Sci. USA* **106**, 20895–20899 (2009).
22. Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. & Cohen, J.D. *Science* **300**, 1755–1758 (2003).
23. Gospic, K. *et al. PLoS Biol.* **9**, e1001054 (2011).
24. Mikhail, J. *Trends Cogn. Sci.* **11**, 143–152 (2007).
25. Bendor, J. & Swistak, P. *Am. J. Sociol.* **106**, 1493–1545 (2001).
26. Marlowe, F.W. *et al. Proc. R. Soc. Lond. B* **278**, 2159–2164 (2011).
27. Boyd, R., Gintis, H. & Bowles, S. *Science* **328**, 617–620 (2010).
28. Fehr, E. & Fischbacher, U. *Evol. Hum. Behav.* **25**, 63–87 (2004).
29. Darley, J. Morality in the law: the psychological foundations of citizens' desires to punish transgressions. *Annu. Rev. Law Soc. Sci.* **5**, 1–23 (2009).
30. LaFave, W. *Criminal Law* (West Group, St. Paul, Minnesota, USA, 2003).
31. Shen, F.X., Hoffman, M.B., Jones, O.D., Greene, J.D. & Marois, R. *New York Univ. Law Rev.* **86**, 1307–1360 (2011).
32. Waytz, A. & Mitchell, J.P. *Curr. Dir. Psychol. Sci.* **20**, 197–200 (2011).
33. Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A. & Saxe, R. *Proc. Natl. Acad. Sci. USA* **107**, 6753–6758 (2010).
34. Heekeren, H.R. *et al. Neuroimage* **24**, 887–897 (2005).
35. Feigenson, N. & Park, J. *Law Hum. Behav.* **30**, 143–161 (2006).
36. Bright, D.A. & Goodman-Delahunty, J. *Law Hum. Behav.* **30**, 183–202 (2006).
37. Buckner, R.L., Andrews-Hanna, J.R. & Schacter, D.L. *Ann. NY Acad. Sci.* **1124**, 1–38 (2008).
38. Duncan, J. *Trends Cogn. Sci.* **14**, 172–179 (2010).
39. Dehaene, S., Spelke, E., Pinel, P., Stanescu, R. & Tsivkin, S. *Science* **284**, 970–974 (1999).
40. Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. *Science* **314**, 829–832 (2006).
41. Buckholtz, J.W. *et al. Neuron* **60**, 930–940 (2008).
42. Oswald, M.E., Orth, U., Aeberhard, M. & Schneider, E. *J. Appl. Soc. Psychol.* **35**, 718–731 (2005).
43. Cushman, F. & Greene, J.D. *Soc. Neurosci.* doi:10.1080/17470919.2011.614000 (23 September 2011).
44. Robinson, P.H., Kurzban, R. & Jones, O.D. *Vanderbilt Law Rev.* **60**, 1634–1649 (2007).
45. Henrich, J. *et al. Science* **327**, 1480–1484 (2010).
46. Koenigs, M., Kruepke, M. & Newman, J.P. *Neuropsychologia* **48**, 2198–2204 (2010).
47. Crockett, M.J., Clark, L., Lieberman, M.D., Tabibnia, G. & Robbins, T.W. *Emotion* **10**, 855–862 (2010).
48. Reidl, K., Jensen, K., Call, J. & Tomasello, M. *Int. J. Primatol. Abstr.* **314** (2010).
49. Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M.A. *Science* **325**, 1272–1275 (2009).
50. Pinker, S. The Better Angels of Our Nature: Why Violence Has Declined (Viking Penguin, New York, 2011).