

# Punishment in Humans: From Intuitions to Institutions

Fiery Cushman\*  
Harvard University

---

## Abstract

Humans have a strong sense of who should be punished, when, and how. Many features of these intuitions are consistent with a simple adaptive model: Punishment evolved as a mechanism to teach social partners how to behave in future interactions. Yet, it is clear that punishment as practiced in modern contexts transcends any biologically evolved mechanism; it also depends on cultural institutions including the criminal justice system and many smaller analogs in churches, corporations, clubs, classrooms, and so on. These institutions can be thought of as a kind of ‘exaptation’: a culturally evolved set of norms that exploits biologically evolved intuitions about when punishment is deserved in order to achieve cooperative benefits for social groups.

---

Building a psychological model of human punishment is something like building a psychological model of Valentine’s Day: perhaps as alluring, and certainly as fraught. What should a psychological model of Valentine’s Day consist of? Human behavior on February 14 is deeply grounded in a long adaptive history of sexual reproduction, in the biology of pair bonding, in the psychology of courtship, and so on. It would be a tremendous folly to regard Valentine’s Day as an arbitrary bundle of cultural norms, as if we might have been equally likely to give Hershey’s kicks as kisses. Yet, it would be equally foolish to approach Valentine’s Day as a pure psychological object. Valentine’s Day is a holiday, after all, and not just a concept or a behavior. It cannot be understood in isolation of the influence of Hallmark, or the symbolic value of roses and chocolate, or even the unlikely historical significance of a fifth century Christian martyr (one for whom the theoretical analogy between love and punishment became all too real). Valentine’s Day lives somewhere at the intersection of these two forces: it is a cultural institution that has hijacked a set of evolved psychological mechanisms.

Punishment is much the same. No doubt, there is some sense in treating punishment as a psychological process. There exist a set of stimuli that the brain perceives; it processes these stimuli according to a set of computations; this ultimately results in a behavior such as punishment, or the judgment that it is deserved. This psychological level of analysis is a genuinely helpful approach as we try to understand the ways in which punishment in the world reflects something about processes in our heads. But ending the story of human punishment there might be nearly as unsatisfying as ending the story of Valentine’s Day with the evolutionary dynamics of sexual reproduction and neurological effects of oxytocin. In any event, it would certainly be as incomplete. Much punishment as practiced by humans today is regulated and institutionalized. An obvious example is the criminal justice system, but similar mechanisms exist in clubs, schools, families, churches, and corporations. Thus, a model of ‘human punishment’ cannot merely specify a process of perception, computation, and behavior. We must also explain the social function of this process, and its interaction with cultural institutions.

This essay summarizes current research on the psychological mechanisms that guide our judgments about punishment, as well as their adaptive history. It also attempts to explain how they do, and don’t, contribute to modern institutions that carry out punishment. Finally, it argues that modern institutions have hijacked an evolved psychology of punishment, in much the

way that Hallmark and Godiva have hijacked love. In other words, institutionalized punishment is an ‘exaptation’: a repurposed adaptation. A long evolutionary history gave rise to psychological mechanisms that supply finely tuned intuitions about when punishment is deserved. Yet, our ability to act upon this motive for revenge is often suppressed in the face of disadvantageous circumstances. (The words ‘he’ll get what he deserves’ more often express a hope than a promise). A key feature of modern institutionalized punishment is to allow these latent motives to attain relatively unrestricted behavioral expression. And these institutions appear to play a key role in the emergence of large-scale cooperative societies.

## 1. *Psychological Mechanisms*

### 1.1. PUNISHMENT IS MOTIVATED BY RETRIBUTION

Philosophical theories identify several candidate motives for punishment. Punishment can deter wrongdoing, both on the part of the person punished (specific deterrence) and on the part of others who observe the punishment (general deterrence). Punishment can also incapacitate the wrongdoer: Few people harm society from jail, and fewer still from the looped end of a noose. Rehabilitation might also follow from punishment, converting a wrongdoer into a do-gooder. Each of these three motives appeals to the beneficial consequences of punishment, and so they are often referred to as consequentialist motives. There is, however, another type of motive for punishment: retribution. According to this rationale, punishment is simply what wrongdoers deserve, whether or not it ultimately brings about beneficial consequences.

When people are simply asked which motivations underlie their behavior, a large majority say that they are influenced by all of them and, when forced to choose, split roughly equally between deterrence and retribution (Carlsmith 2008). But a long history of research indicates that people’s self reports of their own motivations and attitudes are unreliable (e.g., Nisbett and Wilson 1977). Consistent with this finding, a strikingly different picture emerges when people are presented with vignettes that manipulate features that ought to matter according to consequentialist versus retributive theories of punishment. According to a consequentialist approach, people ought to punish more when the punishment is highly public (increasing general deterrence), when the crime is committed frequently and is likely to be committed again (making deterrence important), and when the crime is hard to detect or punish (demanding greater severity of punishment to achieve deterrence). Yet, these factors have virtually no influence upon people’s allocations of punishment in hypothetical vignettes (Carlsmith 2008). Rather, research suggests that punishment depends upon retributive motivations stoked by a sense of moral outrage (Carlsmith et al. 2002; Weiner 1995).

We know a lot about the specific features of an action that provokes such outrage, and many of these features have an apparent consequentialist rationale. This is no accident: There is ample evidence that our retributive motivations have evolved in large part for the purpose of deterrence. But there is an important distinction to be drawn between the ultimate adaptive function of a behavior and its proximate psychological cause (Tinbergen 1952). Consider sex: The ultimate adaptive function of sex is to make babies, but only a small minority of sex acts are undertaken with that proximate psychological motive. Rather, evolutionary pressures have produced a set of specialized motives, triggered in appropriate social circumstances, that prompt a fitness-maximizing behavioral response.

There is a very puzzling feature of this argument. If retributive motivations are adapted to the purpose of deterrence, then why are those retributive motivations not sensitive to precisely the features that Carlsmith and colleagues identified as central to deterrence: the likelihood that a crime will be detected, the likelihood of a repeat offense, the degree to which punishment is

public, and so forth? One answer, which is partially true but also greatly unsatisfying, is that evolution is imperfect. Pigs might be better off with wings, and surely people would be better off if they naturally synthesized penicillin, but the fact that natural selection can explain some adaptive traits does not commit us to predict the existence of every conceivable adaptive trait. A more satisfying answer would explain why some features correlated with deterrent value would be included among the triggering conditions of retribution, while others are not. A first, necessary step toward this more satisfying resolution is to consider these triggering features in detail.

## 1.2. PUNITIVE JUDGMENTS DEPEND ON HARMFUL INTENT AND CAUSAL RESPONSIBILITY

Retributive motivations are triggered most strongly in the presence of three features: (1) an *action* was performed, (2) that action *caused harm*, and (3) the action was performed with a *culpable mental state* such as malicious desire to harm, a callous indifference to the knowledge that harm will occur, or a reckless disregard for its possibility. These same factors are fundamental to the assignment of liability in the Anglo-American legal tradition, presumably because our intuitions about appropriate punishment play an important role in constraining the law (Mikhail 2011; Robinson and Darley 1995).

In psychological research, this three-part landscape is often simplified by contrasting two dimensions: outcome and intent. In this scheme, ‘outcome’ stands in for both the severity of the harm that occurs and the actor’s causal responsibility for it, while ‘intent’ stands in for both the mental states preceding an action (in law a guilty mind, or *mens rea*) and the act itself (a guilty act, or *actus reus*). These factors can be arranged in a  $2 \times 2$  design that yields four cells. Outcome and intent match in cases of intentional harm (Annie aims a gun at Frank, shoots, and kills him) and benign action (Annie aims at a stump, shoots, and hits it). They are mismatched in cases of accidental harm (Annie shoots at the stump but misses, hitting Frank instead) and attempted harm (Annie shoots at Frank but misses, hitting a stump instead).

This basic experimental design has been used widely in moral judgment research, with several important results. Among the earliest and most consistent findings is that young children tend to place more emphasis on information about outcomes, while older children and adults tend to place more emphasis on information about intent (Armsby 1971; Costanzo et al. 1973; Cushman et al. 2013; Hebble 1971; Imamoglu 1975; Karniol 1978; Killen et al. 2011; Nobes et al. 2009; Shultz 1986; Yuill and Perner 1988; Zelazo et al. 1996; but see Nobes et al. 2009). More recently, researchers have used this design to interrogate the neural substrates of outcome-based and intent-based judgment processes (Buckholtz et al. 2008; Young et al. 2007) and have made particularly impressive progress in elucidating the contribution of a network of brain regions to the assessment of culpable mental states (Young and Saxe 2008; Young et al. 2010a, 2010b).

This experimental design has also been used to reveal a distinctive characteristic of punishment judgments, which set them apart from other categories of moral evaluation such as moral wrongness, moral permissibility, or moral character: Punishment depends especially strongly on accidental outcomes. Consider, for instance, two drunk drivers who each fall asleep at the wheel. One is lucky and runs into some bushes, while another is unlucky and hits a pedestrian, killing her. In this circumstance, they have identically culpable mental states but are causally responsible for outcomes that vary greatly in severity. Research indicates that people would tend to judge the two drivers to have acted roughly equally wrongly, and to be roughly equally bad people, yet assign much greater punishment to the second driver based on the outcome that he is responsible for (Cushman 2008). Philosophers refer to this phenomenon as ‘moral luck’ (Nagel 1979; Williams 1981) because it reveals a manner in which our moral evaluations depend on chance.

Why is moral luck a prominent feature of punishment judgments, while it is largely absent from other categories of moral evaluation? At a proximate, psychological level, it is argued that this dissociation reveals two distinct processes of moral judgment (Cushman 2008; Cushman et al. 2013). One is triggered by the occurrence of a harmful outcome and seeks out the causally responsible party (we'll call this the 'causal process'). The other begins with an action and queries the mental state that gave rise to that action (we'll call this the 'mental state process'). While the mental state process appears to be ubiquitous in moral judgment, the causal process appears to be a special feature of punitive judgments. Of course, this raises an obvious question at an ultimate, adaptive level: What is the function of the causal process in moral judgment, such that moral luck might be a desirable property? We will take up this question shortly.

### 1.3. EXCEPTIONS TO THE RULE

Although many instances of intentional harm are punished, many others are not – here, again, we find an elegant correspondence between the law and the ordinary moral intuitions of non-experts. For instance, harmful actions can be justified when the actor's life, or the lives of others, is threatened by extraordinary circumstances (*force majeure*), if she is attacked or provoked (self-defense), or if she ultimately acts as a caretaker of the best interest of the 'victim' (e.g., as a doctor or a parent). These doctrines of justification have received relatively less attention in the literature on punishment (but see Malle et al. 2014; Robinson and Darley 1995; Weiner 1995). However, a common perspective is that justifying circumstances are processed subsequent to an initial assignment of blame. That is, it appears that first people assign blame based on causal responsibility for harm and then engage the processes of assessing whether the harm was justified.

Another set of exceptions concerns mental capacity and controllability: We assign less punishment to individuals who could not control their behavior or who otherwise acted with diminished capacity (Darley et al. 2000; Fincham and Roberts 1985; Robinson and Darley 1995). Children and individuals with mental disorders are often considered to act with diminished capacity. A lack of control over behavior might be attributed to a temporary disruption of mental functioning (e.g. hitting a person during a seizure or acting aggressively under the influence of an untreated brain tumor) or to circumstances that inspire passions beyond control (such as when a person discovers their spouse with another lover *in flagrante delicto*).

Finally, there are exceptions of the opposite variety: circumstances in which we tend to punish despite a lack of harm. For instance, as a legal matter, we punish drug use, prostitution, and polygamy even when there is no evidence that a person's conduct has led to any harmful outcome. Although one could argue that each of these behaviors leads to harm as a general category, if not in every specific instance (Gray et al. ), the weight of the evidence suggests that in fact people do condemn some actions for reasons other than their connection to a harmful outcome (Cushman 2013; Greene 2013; Haidt 2012; Young and Saxe 2011). Notably, however, there is much greater variation between individuals in the degree to which they think that such non-harmful acts should be punished, compared with harmful acts (Graham et al. 2009; Robinson and Kurzban 2007). These findings might be taken to suggest two broad categories of behaviors that trigger punishment: those that cause harm, which are relatively invariant across cultures, and those that violate culturally contingent norms. This distinction stands out as an important topic for further research. Notably, it appears to be reflected in a developmentally early-emerging ordinary folk distinction between moral and conventional violations (Turiel 1983).

## 2. Adaptive Rationale

### 2.1. PUNISHMENT AS PEDAGOGY

With some effort, it is possible to see punishment as an evolutionary mystery. In evolutionary game theory, punishment is typically modeled as a behavior that carries some cost for the punisher and also imposes some cost on the recipient. So the mystery is: Why would it maximize my fitness in order to hurt myself while hurting somebody else, too? But this question has an easy answer, which is that punishment can be used as a way to modify the future behavior of a social partner, preventing them from causing you harm in the future. Let us call this the pedagogical perspective. It proposes that you pay a short-term cost of teaching in order to maximize long-term gains when a social partner learns.

Evolutionary modeling bears out this logic. When punishment modifies others' future behaviors, converting them from harm-doers to do-gooders, it is adaptively favored (Boyd and Richerson 1992; Clutton-Brock and Parker 1995; Fehr and Gächter 2002; Henrich and Boyd 2001). Moreover, it appears that we have either evolved or learned to anticipate the pedagogical consequences of punishment (Krasnow et al. 2013). In this experiment, participants played a trust game with a partner who either cooperated with or defected against them. Participants were then allowed to either punish the defector or desist from punishment. Finally, the participants were given the opportunity to play again with the same partner, either trusting them or not. Participants were, of course, very trusting of partners who had previously cooperated. Remarkably, however, they were equally as trusting of partners who had defected but whom they then punished, while they showed little trust of partners who had defected but whom they had not punished. In other words, they acted on the expectation that punishment will fully reform an erstwhile sinner.

The pedagogical perspective can help to explain the characteristic triggers of our retributive motivations (Cushman 2013; McCullough et al. 2013). First, and most obviously, we direct punishment at individuals who have performed actions that cause harm so that they will learn not to perform those actions in the future. We are especially harsh on actions that are performed intentionally, presumably because individuals who intend harm are especially in need of an adjustment to their social motivations. However, we also punish people who acted accidentally, so long as their behavior was controllable – the phenomenon moral of luck. This may be because accidents are 'teachable moments'. Consider a puppy who pees on the carpet or a young child who spills her milk. Both of these behaviors are accidents – they are certainly not malicious! – but they are also moments that a teacher can exploit through punishment. Because the puppy and the child are both ultimately capable of controlling their behavior in relevantly similar circumstances in the future, punishing them may ultimately pay off. Recent research demonstrates that in the absence of controllability, however, the punishment of accidents is diminished (Martin and Cushman in preparation).

Although the pedagogical perspective explains many cases of punishment, it is certainly not the only explanation available (reviewed in Nakao and Machery 2012). For instance, adaptive models show that it can be valuable to impose fitness costs on social partners to lower the frequency of unshared trait (Rand and Nowak 2011). There are also some circumstances in which a social interaction is disadvantageous to one partner but advantageous to another (e.g. persistent food theft during joint foraging). If the disadvantaged partner has the ability to unilaterally disengage from the social partnership (e.g. forage on her own), this imposes a cost on the advantaged partner that might be conceived of as a form of punishment (Hirshleifer and Rasmusen 1989). This mechanism of partner choice imposes a selective pressure within social groups to engage in mutually advantageous interactions. It is argued that many cases – perhaps

most – of punishment among non-human animals are explained by these alternatives to the pedagogical perspective (Baumard et al. 2013). One likely explanation is that humans have a uniquely powerful capacity to learn from punishment, both by inferring punitive intent and because of natural language (Raihani et al. 2012).

This evidence raises an important question about the appropriate technical definition of punishment. At one extreme, Nakao and Machery (2012) define punishment as ‘an action that harms another organism’ (p. 834) and explicitly include ‘a failure to cooperate’ (p. 835) within the scope of harmful behavior. It is worth reflecting on just what a broad definition this is. For instance, if a male pidgin courts a female but fails to impress her and then she flies away, her choice not to mate would qualify as a ‘punishment’. A definition this broad is well suited to the project of identifying and contrasting all the various functional explanations that might exist for an act of harm-doing, and this was precisely the purpose of Nakao and Machery’s treatment. However, it is poorly suited to identifying a class of behavior that will have much internal coherence, precisely because so many biological functions are compatible with the class. Here, I pursue an approach at more or less the opposite extreme, focusing exclusively on actions that harm another organism for the purposes of modifying their behavior. In other words, I take one particular function of punishment to define the relevant object of study. This choice does not deny other functions of other-directed harm, or to downplay their importance, but rather targets a category likely to exhibit some internal coherence: a functionally defined natural kind of social behavior.

Even as an explanation of (some) punishment in humans, however, the pedagogical perspective raises important questions. First, if the ultimate adaptive function of punishment aligns with consequentialism, then why are people retributive rather than consequentialist in their proximate psychological motivations? Second, for whose benefit are the ‘lessons’ of punishment ultimately taught: the benefit of the punisher or the benefit of the broader society that she lives in?

## 2.2. WHY ARE WE PSYCHOLOGICAL RETRIBUTIVISTS IF WE ARE ADAPTIVE CONSEQUENTIALISTS?

Given that punishment has a consequentialist adaptive rationale and that people are perfectly capable of reasoning about the punishment from a consequentialist perspective, it is remarkable that they find consequentialism so utterly unmotivating. Why does the psychology of punishment reduce to a blinkered obsession with retribution? It is almost as if evolutionary forces had deliberately blinded us.

This possibility is not as strange or self-defeating as it sounds. In fact, it is argued that emotions often play precisely this role: blindly committing us to a course of action that reason would undermine (Elster 1979; Frank 1988). It turns out that such self-blinding can pay handsome dividends in the context of strategic multiparty interactions. A familiar example is the doctrine of mutual assured destruction that dominated strategic thinking during the cold war. According to the doctrine, a nation should pre-commit itself to the complete nuclear annihilation of any rival that launches a strike against them. This pre-commitment establishes the maximum possible deterrent to rivals. But it also demands blind commitment to a strategy that becomes irrational once executed. After all, if your nation already faces certain and complete doom from a pre-emptive nuclear strike, there is little benefit in spitefully launching a counterstrike merely to guarantee the annihilation of the enemy. You won’t be around to gloat.

One way that evolution might accomplish such an irrational pre-commitment is via a dedicated emotional response. Although we did not evolve to avert nuclear war, similar logic can be applied in less apocalyptic circumstances. For instance, the emotion of love can be viewed as a blind emotional commitment to fidelity, enabling mutual trust in a romantic partnership. Critically, in order for such an emotional mechanism to work, it must circumvent reasoning

in order to perpetrate a local irrationality that achieves a larger strategic end. That is, a smitten spouse must 'irrationally' forgo safe opportunities for infidelity in order for love to act as an honest signal that enables trust with his partner.

Is blind revenge similarly far-sighted at a strategic level? Suppose that people adopted a strictly rational and consequentialist approach to punishment. This would leave them susceptible to exploitation by a social partner who strategically refuses to learn from punishment. The unfortunate consequences are clear enough in an everyday setting. Imagine a parent who is trying to teach a child her manners. The child protests, cries, and whines. A rational conclusion on the part of the parent would be, 'My child is untrainable, and I should give up'. This is a strategic victory for the child that exploits her parents' rationality, and an utter disaster for the parent. Alternatively, the parent may blindly commit herself to a strategy of punishment, forcing the child's hand. In a sense, the logic is similar to mutual assured destruction: If any harm to my welfare is met with righteous anger and harsh punishment, then you are forced to respect my welfare. Thus, the most effective means of accomplishing the ultimate consequentialist aims of punishment may be to eschew consequentialist analysis in favor of retributive emotions at the proximate level.

This framework offers some hope for explaining why retributive motivations are sensitive to some, but not all, of the features that predict the retributive value of punishment. Broadly speaking, retributive punishment is highly sensitive to retrospective features of a situation (Who caused harm? Did they mean to? Was it controllable?), but it is not sensitive to prospective features (Will punishment teach them a lesson? Will it teach others a lesson? Are they likely to reoffend?). Of course, a would-be harm-doer cannot avoid punishment by manipulating retrospective features of their action while still reaping the benefits of transgression. If you don't want to be punished for having intentionally transgressed, your only course of action is not to intentionally transgress. On the other hand, prospective features are ripe for manipulation; a person can first transgress and then ensure that punishment will not deter them from future transgression. Thus, our retributive motivations may depend exclusively upon retrospective features because prospective features are exploitable.

### 2.3. SECOND-PARTY AND THIRD-PARTY PUNISHMENT

Much human punishment occurs in dyadic contexts. For instance, Jake steals Matt's food, so Matt punches Jake. This is sometimes called 'second-party punishment', and its adaptive rationale is straightforward. Although Matt pays a short-term cost (punching), he will presumably derive a long-term benefit, as Jake becomes less likely to steal food from him in the future. Both the existence of second-party punishment and its adaptive rationale are largely uncontroversial.

But, there is some evidence that humans at least occasionally punish in third-party contexts. For instance, Jake steals Matt's food, so Sally punches Jake. This form of punishment requires a more sophisticated adaptive analysis to explain because, according to a straightforward accounting of the costs and benefits, Sally has paid a personal cost in order to benefit Matt. In this sense, her act of punishment resembles what biologists often call altruism. Her behavior is challenging to explain in the same way that it is challenging to explain why Sally would do anything nice to Matt.

Several potential answers to this challenge readily come to mind, each mirroring a typical adaptive explanation for other forms of altruism. Invoking kin selection, we could explain Sally's behavior if she is genetically related to Matt. Invoking reciprocal altruism, we could explain her behavior if Matt is likely to repay her somehow in the future. In addition, it might be the case that Sally derives a direct fitness benefit from enhancing her own reputation as a force not to be reckoned with, thus preventing others from transgressing against her directly (Brandt et al. 2003).

These explanations would not suffice, however, if it could be shown that Sally would punish Matt in the context of an anonymous and one-shot interaction. Fehr and Fischbacher (2004) tested participants in this context by allowing one player (A) to divide resources between himself and another player (B) and then informing a third player (C) of the resource allocation and giving C the opportunity to punish unfair allocations by A. They observed that over 60% of participants in the role of C punished unfair allocations (but see Pedersen et al. 2013 for a methodological critique and evidence for much lower rates of third-party punishment). This and other related empirical findings prompted a set of theoretical models that purported to explain norms of third-party punishment in terms of group selection, often involving cultural rather than genetic adaptation (Boyd and Richerson 1992; Boyd et al. 2003; Gintis 2000; Gintis et al. 2003). In other words, it is proposed that people have a norm of third-party punishment because individuals in cultural groups that possess such a norm outcompete individuals in cultural groups that lack one. This is because punishment can stabilize cooperative behavior (Fehr and Gächter 2002), and cooperative behavior promotes group welfare.

In order to make these evolutionary dynamics successful, it is sometimes assumed that there is an additional norm of second-order punishment: Individuals who fail to engage in third-party punishment are, in turn, punished. Empirical support for such second-order punishment is mixed, at best. For instance, Kiyonari and Barclay (2008) showed that people are no more likely to punish those who *don't* punish than to punish those who *do*. This poses a challenge for many theories that rely on group selection to explain third-party punishment. One response has been to investigate mechanisms other than second-order punishment that could stabilize third-party punishment under group selection (Henrich and Boyd 2001), while another has been to investigate evolutionary dynamics that stabilize third-party punishment without group selection and with minimal requirements of second-order punishment (Fowler 2005).

Below, we will consider the merits of several of these models in greater detail. But it is worth pausing for a moment to remark on the scope of the phenomenon that they aim to explain. Far from being limited to the laboratory behavior of anonymous undergraduate trading partners in Cambridge or Zurich, these theories seek to explain nothing less than the emergence of large-scale human societies (Henrich et al. 2006, 2010). The claim is that a central challenge of life in such societies is the maintenance of cooperation when many social interactions are anonymous. And the still bolder claim is that cultural norms of altruistic punishment, fostered through group selection, are a key part of the answer to this challenge. In order to give these claims the attention they deserve, we must ask how and when punishment actually occurs in practice.

### 3. Punishment in Practice

#### 3.1. DO PEOPLE ACTUALLY PUNISH EACH OTHER?

Many discussions of the psychology of punishment conspicuously dodge a pivotal question: Outside of the laboratory or the hypothetical setting of a vignette, how often – and in what circumstances – do people actually punish each other? There are, of course, childhood food fights, grown-up bar fights, acts of vandalism, road rage fisticuffs, Internet tirades, whistleblowers, lynch mobs and vigilantes, strikes, boycotts, and other such legitimate forms of punishment. But if you carefully separate the fiction of TV dramas and sensationalization of the evening news from the more prosaic experiences of everyday life, it is remarkable just how little punishment actually occurs. When we study punishment, what exactly are we studying?

There is a category of gossip, slander, and social backstabbing that does impose more substantial costs on the victim, which is referred to in the literature as indirect, relational, or social aggression (Archer and Coyne 2005). These are non-physical acts that occur throughout the life

span but reach their apogee in the preteen and early teenage years. However, careful analysis of these behaviors reveals that they do not serve the pedagogical function of punishment but rather function as a mechanism to undermine and exclude undesired social partners. Consequently, it is argued that indirect social aggression is a categorically distinct kind of behavior from punishment (Archer and Coyne 2005).

One obvious category of 'real-world' punishment is institutional punishment, that is, punishment by agents of the state, of corporations, or of other institutions (unions, clubs, schools, sports teams, etc.). Another, distinct category of punishment occurs within families: parent to child, between spouses, and so forth (often in the form of domestic violence). These contexts are alike in the respect that they typically involve stark asymmetries of power, with the punisher holding an advantage over the punished. Beyond these two categories, the landscape becomes much more complicated, but two features stand out prominently.

The first feature is threat: In many circumstances, we signal anger without much imposing actual cost. For instance, we might make a testy remark to a co-worker who leaves their dirty dishes in the communal sink. If the remark itself does not impose any cost on the co-worker, then it does not meet the technical definition of a punishment. Rather, such actions are presumably favored because they act as a signal or a threat of potential future punishment. Even in circumstances where some minimal cost is imposed, its magnitude is slight compared with the implied threat of future action. Presumably, it is more the threat implied than the harm imposed that achieves a deterrent effect.

The second feature is withdrawal: In many instances, we withhold (or threaten to withhold) the benefits of cooperation, rather than imposing (or threatening to impose) a spiteful cost. Consider again the testy remark directed at our discourteous co-worker. What kinds of threats are realistically implied? Many of these are threats of discontinued cooperation: the loss of friendship, social support, favors, or even promotion. Relatively fewer are likely to be threats of spiteful harms imposed: destruction of personal property, physical assault, the active thwarting of goals, or demotion.

According to some recent treatments, the withdrawal of future cooperative benefits might be viewed not as a form of punishment but in fact as a rival explanation for the emergence and maintenance of prosocial behavior (e.g., Baumard et al. 2013). The crux of the issue is a distinction between partner control (using punishment to modify the behavior social partners) and partner choice (biasing social interaction selectively toward prosocial partners). Theoretical models show that when individuals have the ability to choose their social partners and bias their choice toward partners who provide the greatest benefits to them, this promotes mutualistic social interactions (e.g., cooperation in a prisoner's dilemma) and discourages antisocial behavior (Bull and Rice 1991; Noë et al. 1991; Roberts 1998). There is widespread evidence for the influence of partner choice among humans and non-human animals (reviewed in Baumard et al. 2013).

Partner choice is sometimes claimed to have a key advantage over partner control: Ceasing to interact with an antisocial social partner is typically assumed to be costless (or even profitable), while punishing an antisocial partner is typically assumed to be costly. Instantiated as such, both formal models and empirical research suggest an advantage to partner choice (e.g., Dreber et al. 2008). Yet, to cast partner choice as an alternative to partner control overlooks the possibility that the threat of withdrawing social interaction can accomplish the goal of modifying a social partner's future behavior. In other words, the threat of partner choice can operate as a mechanism for partner control. For instance, imagine that an unscrupulous car repairman overcharges you for a part. If you simply cease to interact with the repairman, this corresponds to partner choice. On the other hand, if you threaten the repairman by saying, 'if you do that again, I'll never come back', this corresponds to partner control: The threat is adaptive not because it shields you from interacting with a deterministically antisocial partner, but rather because it is likely to modify your social partner's future behavior.

It is not surprising from a strategic or adaptive perspective that humans often rely on the threat of social withdrawal rather than the imposition of actual costs in order to punish, especially outside of strict dominance hierarchies. Imposing direct costs on social partners carries the formidable risk of retaliation. Still worse, retaliation may prompt counter-retaliation, and so on, leading into a downwards spiral of fitness costs. ('Blood will have blood', in Shakespeare's succinct phrasing.) For this reason, punishment is typically not strategically favored unless the punisher holds a strong position of power over the target of punishment (Clutton-Brock and Parker 1995). This may explain the relative prevalence of punishment in institutionalized settings and in adult-child relations, compared with settings among more coequal actors. It may also explain why people who do not enjoy the benefits of a strong dominance relation often favor a signal of their displeasure over an actual punitive act. For instance, in a review of social sanctioning mechanisms practiced in small-scale societies, Boehm (1999, p. 70–89) suggests a widespread pattern of behavior in which ridicule, criticism, and ostracism are used to signal discontent and bring an deviant actor back into line with community norms. Only when such attempts fail will this trigger extreme group action (e.g., complete ostracism or even assassination of the transgressor, consistent with partner choice).

Experimental evidence reveals the costs of retributive cycles of violence in practical terms (Dreber et al. 2008; see also Nikiforakis and Mitchell 2013). In this experiment, participants were paired with several anonymous partners, interacting with each several times before moving on to the next. The pairwise interactions were structured around a typical prisoner's dilemma, but with a twist: In addition to the ordinary options of cooperation and defection, participants could also choose to punish each other. Thus, if Mary defected against Bob on round 1, Bob could retaliate either by defecting against Mary on round 2 (i.e., withholding the benefits of cooperation from her) or by actually imposing a cost on her at a cost to himself. Some participants tended to opt for defection, while others chose punishment. Dreber and colleagues tracked the earnings of each group over time. Defectors strongly outperformed punishers, and the culprit in punishers' poor performance was largely the risk of retributive cycles of violence (but see Gächter et al. 2008). This may explain why it is much easier to think of instances in which we respond to others' antisocial actions by passive disengagement than by active harm. Moreover, punishment is often found to *decrease*, rather than increase, group-level payoffs, challenging a central pillar of cultural group selection models (for a summary of these results, see Dreber et al. 2008; for an alternative perspective, see Gächter et al. 2008).

It is helpful to anchor the abstract features of an economic game in the real world. In a recent study of third-party punishment in naturalistic settings (Balafoutas and Nikiforakis 2012), an experimenter traveled into the subway stations of Athens, waited until they were standing by a single naïve passenger, and then threw litter on the ground in full view of that passenger. Although 91% of passengers surveyed said that they would be bothered by this kind of behavior, only 4% actually intervened – even at the minimal level of saying something to the litterer. By far the most common explanation that people offered for non-intervention was the fear of retaliation. It is instructive to compare these results to a standard third-party punishment economic game played in a laboratory-setting Athens, in which 89% of participants engaged in costly third-party punishment.

Collectively, these studies reveal a shortcoming in much of the existing research on punishment. We have reviewed many studies that ask people to indicate how much punishment is 'deserved' for various transgressions. And we have seen that there is also a large body of research in behavioral economics that shows that people are willing to pay a personal cost in order to carry out punishment. Yet, both methods inadvertently avoid accounting for the costs of retribution. To say that a person 'deserves' punishment in a survey is quite different from saying that you would punish them yourself. And the context in which punishment is typically carried

out in economic games is in one-shot anonymous interactions, where fear of retaliation is moot. Relatively fewer studies have asked whether participants are willing to engage in punishment in non-anonymous situations where they have a legitimate fear of retaliation.

### 3.2. CULTURES OF HONOR

One social context in which humans are famously prone to bear extraordinary personal costs in the name of punishment is the ‘culture of honor’, and so it is useful to attempt to understand why. Cultures of honor are a well-described and widespread phenomenon; some prominent case studies include the early settlers of Iceland (Miller 1990) and Montenegrins (Boehm 1984) and herders of the American south and their decedents (Nisbett and Cohen 1996). Other examples include Bedouin and the mafia. Cultures of honor typically emerge in areas with scarce natural resources and no centralized state-like authority. Individuals place a high value on their personal honor, which is typically defined in terms of violent self-defense for men and sexual fidelity (or chastity) for women. Perceived slights or actual violations of honor are met with extreme violence.

Additionally, cultures of honor typically involve tight family-based social groups (‘clans’) and a norm of vicarious punishment. That is, if a man from one clan murders a man from another clan, the victims’ brother would feel obligated to retaliate. In the event that he could not directly retaliate against the murderer, it would be considered acceptable to retaliate by murdering any man in the rival clan of roughly equivalent age and social rank to the victim – for instance, the murderer’s brother. For obvious reasons, this can lead to a protracted cycle of retaliation, or blood feud. The decades-long blood feud between the Hatfields and the McCoys is a famous historical example.

Belying the cultural stereotype of hotheaded cowboy hell-bent on revenge, ethnographic evidence shows that the participants in blood feuds are often ambivalent or outright reluctant retaliators (Boehm 1984; Miller 1990). Their behavior is deeply shaped by their cultural and social setting. This explains why phenomena like vicarious punishment are relatively restricted to specific cultures, and not a human universal. Within cultures of honor, women play a particularly important role in goading fathers, husbands, and sons into retaliatory violence. For instance, Boehm (1984) describes ‘a mother [who] repeatedly showed a container of her dead husband’s blood to her young sons to remind them, as they grew up, that since there was no one else to do the job, they must avenge him’.

But while punishment in cultures of honor is remarkable for its effect in prompting punitive behavior, it appears to build upon many of the ordinary psychological foundations that we considered at the beginning of this essay in terms of punitive judgments. Punishment is triggered by causal responsibility, and although it need not be targeted at the responsible *individual*, at least it must be targeted at the responsible individual’s *clan*. There are clear norms of proportionality between the severity of a transgression and the harshness of the response. Although accidents are sometimes punished, on the whole, retaliation for accidental transgressions is less severe than for intentional transgressions (Boehm 1984; Miller 1990). There are typically mechanisms by which apologies, accompanied by compensation, can erase or at least substantially diminish the likelihood of revenge. It is likely that these features are no accident and that vengeance in cultures of honor is parasitic on widely shared psychological mechanisms for computing moral responsibility and punishment. Indeed, in those instances where the punitive norms of cultures of honor deviate from the psychological template described in Section 1 – for instance, in the practice of vicarious punishment, in which the male relative of a transgressor is a legitimate target of revenge – there is some evidence that punishment is actually not accompanied by an intuitive sense that its target is morally responsible (Cushman et al. 2012).

## 3.3. INSTITUTIONALIZED PUNISHMENT

Modern western norms and institutions surrounding punishment are every bit as unique as those of the culture of honor and also equally grounded in widely shared psychological mechanisms. At the level of states, corporations, schools, and clubs, mechanisms of punishment are highly formalized in modern western culture. Rules are commonly stated explicitly and communicated *ex ante*. Determinations of guilt and punishment are made by designated individuals, and their social roles are often professionalized. It only takes a moment of reflection to realize that these highly institutionalized mechanisms of punishment are both unusual and specialized from a broad historical and cultural point of view (Fukuyama 2011).

Institutionalized punishment has two key advantages at a group level. First, it solves the problem of sanctioning public goods violations. If an individual engages in a behavior that impacts all members of the group equally (e.g., does not pay taxes), punishing that individual becomes a public good: The punisher must pay individual costs to provide a collective benefit. Punitive institutions like a state solve this problem by professionalizing the role of punishment, thus compensating the punisher for her effort. Second, it solves the problem of retributive cycles of violence. When the state has a professionalized and highly armed enforcement arm, for instance, there is little utility in mounting a campaign of retaliatory violence against it.

There is a very abstract sense in which institutionalized punishment can be thought of as a variety of ‘third-party punishment’, but this analogy does not hold at the level of psychological mechanism. Imagine, for a moment, an experimental economist bringing three undergraduates into the laboratory. She shows that one undergraduate is willing to pay to punish another who stiffed a third. Does this behavior rely upon the same psychological mechanisms that govern a judge’s behavior when, as a salaried professional, she assigns a sentence to a criminal in accordance with legal standards? Does it rely on the same psychological mechanisms as a police officer who responds to a call and makes an arrest? More pointedly, is the behavior of the judge or the police officer best explained by appealing to a cultural norm of third-party punishment – one that causes people to spontaneously pay personal costs in order to punish those who transgress against third parties, anonymously, as in the laboratory? Surely many public servants are motivated by noble aims but also – and often overwhelmingly – by institutional systems that align personal self-interest (salary, promotion, and the law) with a public good.

These considerations invite a new perspective on the role of third-party punishment in the emergence and maintenance of large-scale society. From this new perspective, the central figure is not a norm of third-party punishment in the experimental economists’ sense (‘When a person sees one stranger harm another stranger, that person will spontaneously pay a personal cost in order to punish the harm-doer’). Rather, the central figure is an institution: the state, the corporation, the school, or the club. Institutions work by creating incentives for individuals to police third-party norm violations, but, critically, those incentives mean that the individuals doing the policing *need not pay personal costs*. By engaging in ‘third-party punishment’, these individuals advance the interests of the group, but they also advance their own personal interests in a very direct and straightforward manner. Some attempts have been made to model this process experimentally, and with promising results (Hilbe et al. 2014).

Thus, while modern institutions of punishment can be squeezed into the corset of ‘cultural group selection’, the resulting fit is unflattering. Much depends on what is meant by ‘culture’ and what is meant by ‘group selection’. Formal institutions undoubtedly depend upon culture, but they have distinctive properties that demand a level of explanation beyond mere ‘shared norms’. (By analogy, human culture depends upon a human genetic endowment, but it is not useful to therefore describe human culture as simply another biologically evolved trait.) A similar critique applies to the claim that third-party punishment is explained by group-level selection. Institutional punishment undoubtedly provides group-level benefits, and these benefits

undoubtedly explain its prevalence. However, to describe a behavior as a product of ‘group selection’ is often taken to imply that the behavior carries a fitness cost for the actor – i.e., that the individual pays a cost that is recouped by the collective (West et al. 2007). Yet, the chief innovation of state-like institutions is to align the self-interest of the individual with the interests of the collective. In other words, one does not need to look beyond the individual level of selection to understand why a policeman arrests a criminal when his salary depends upon it.

In both of these respects, cultures of honor provide a useful foil to modern institutionalized punishment. The norms governing behavior in a culture of honor were often explicit and sometimes even codified, but mechanisms of retributive violence were never institutionalized in the sense of being professionalized, bureaucratized, salaried, and so forth. In this sense, the culture of honor is, in fact, best described as a ‘culture’ of shared norms, rather than as an institution. And, for this very reason, it was often the case that individual actors in a culture of honor were called upon to pay extreme personal costs (e.g., death) for the benefit of the collective. This contrast between cultures of honor and modern state institutions illustrates the usefulness of the concept of cultural group selection – including for the purpose of understanding human punishment, for instance in cultures of honor – but also its limitation as an explanation for the emergence of large-scale societies.

In one pivotal respect, however, modern institutionalized punishment is very similar to cultures of honor: It relies heavily on humans’ intuitive sense of justice. This point is well established in existing research. For instance, Robinson and Darley (1995) compared criminal law and sentencing guidelines against ordinary peoples’ judgments of liability and punishment. Although their work reveals several instances of divergence, the similarities were more striking. The concepts of causal responsibility, proportionality, gradations of culpability, doctrines of justification and excuse, and the degree of action required for criminal conduct all exhibit broadly similar contours across both legal standards and the untutored judgments of ordinary people. Mikhail (2009, 2011) finds a similar correspondence. According to the theory of universal moral grammar, a wide range of intuitions about the permissibility of harmful actions can be captured through a formalized set of cognitive operations that closely parallel key elements of both criminal and tort law. Mikhail suggests that our legal institutions evolve over time to fit the underlying psychological mechanisms that give rise to moral intuitions.

#### 4. *Synthesis: Institutional Exaptation*

Humans have a taste for retribution and an exquisite sense for when it is deserved. Because personally engaging in punishment carries the potential for a downward spiral of retaliation, however, we tend not to observe much punishment except in restricted cases with a large power asymmetry (e.g., parent to child). Institutional mechanisms, such as a state-sponsored criminal justice system, provide a handy workaround by establishing an authority that acts upon our sense of desert without fear of retaliation. Thus, institutions can release – and, in fact, exploit – a psychological urge that is otherwise often behaviorally repressed. This is a familiar evolutionary process called exaptation: A trait originally evolved for one purpose is later exploited for another purpose. In the case of punishment, the proposed exaptation bridges between a biologically evolved urge toward second-party retaliation and a culturally adaptive institution of third-party punishment.

This model of ‘institutional exaptation’ stands in contrast to recent models of cultural group selection, yet it also shares a number of important features with them. According to the cultural group selection hypothesis, individuals’ punitive motivations are largely the product of culturally inherited norms. Moreover, the structure of third-party punishment is best understood by appealing to the group-level adaptive value of third-party punishment. Finally, the utility of norms such as third-party punishment explains the emergence of large-scale society.

According to the institutional release hypothesis, individuals' punitive motivations are largely innate (although unquestionably, they are strongly influenced by cultural forces as well). Moreover, the adaptive function responsible for shaping the structure of punitive motivations was mostly the direct pedagogical benefit of second-party punishment. Due to the risk of retaliation, however, it was often maladaptive to actually engage in punishment. Thus, punitive motivations were often behaviorally expressed in terms of criticism, ridicule, threats, and anger (Boehm 1999). Institutions such as states allow for the 'release' of punitive motivations because they can carry out punishment unilaterally, and without fear of retaliation.

Like models of cultural group selection, the institutional exaptation hypothesis credits robust third-party punishment with a key role in the emergence and success of large-scale societies. But it emphasizes the role of institutions in making third-party punishment compatible with personal self-interest via professionalized roles (like police), rather than supposing that individual punitive actors constitute the backbone of third-party punishment or that punitive action carries individual costs (like the anonymous third-party punishers of laboratory experiments).

Many of the features that make punishment a daunting object of psychological study also make it an exemplar of the moral domain. If the psychology of punishment can only be understood as an interaction between biology, culture, and institutions, then surely the same is true of the psychology of cooperation, forgiveness, generosity, fairness, character, trust, and so forth. In other words, if punishment is even a little bit like Valentine's Day, then all of morality is very much like punishment.

### *Short Biography*

Fiery Cushman is an Assistant Professor of Psychology at Harvard University, where he directs the Moral Psychology Research Laboratory. His research investigates the cognitive mechanisms responsible for human moral judgment, along with their development, evolutionary history, and neural basis. His work often draws from classic philosophical dilemmas and has focused in particular on the psychology of punishment and the aversion to harmful action. He received his BA and PhD from Harvard University, where he also completed a post-doctoral fellowship. He was an Assistant Professor of Cognitive, Linguistic, and Psychological Sciences at Brown University from 2011 to 2014.

### *Note*

\* Correspondence: Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138, USA. Email: [cushman@wjh.harvard.edu](mailto:cushman@wjh.harvard.edu)

### *Works Cited*

- Archer, J., and S. M. Coyne. 'An Integrated Review of Indirect, Relational, and Social Aggression.' *Personality and Social Psychology Review* 9.3 (2005): 212–30.
- Armsby, R. 'A Reexamination of the Development of Moral Judgments in Children.' *Child Development* 42, (1971): 1241–1248.
- Balafoutas, L., and N. Nikiforakis. 'European Economic Review.' *European Economic Review* 56.8 (2012): 1773–85. DOI: 10.1016/j.euroecorev.2012.09.008.
- Baumard, N., J. B. André, and D. Sperber. 'A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice.' *Behavioral and Brain Sciences* 36.1 (2013): 59–78.
- Boehm, C. *Blood Revenge: The Anthropology of Feuding in Montenegro and Other Tribal Societies*. Lawrence, KS: University Press Kansas, 1984.

- Boehm, C., & Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge MA: Harvard University Press.
- Boyd, R., et al. 'The Evolution of Altruistic Punishment.' *Proceedings of the National Academy of Sciences, USA* 100 (2003): 3531–5.
- Boyd, R., and P. Richerson. 'Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups.' *Ethology and Sociobiology* 13.3 (1992): 171–95.
- Brandt, H., et al. Punishment and Reputation in Spatial Public Goods Games. *Proceedings: Biological Sciences*, 2003: 1099–104.
- Buckholtz, J. W., C. L. Asplund, P. E. Dux, D. H. Zald, J. C. Gore, O. D. Jones, and R. Marois. 'The Neural Correlates of Third-Party Punishment.' *Neuron* 60.5 (2008): 930–5.
- Bull, J. J., and W. R. Rice. 'Distinguishing Mechanisms for the Evolution of Co-operation.' *Journal of Theoretical Biology* 149.1 (1991): 63–74.
- Carlsmith, K. M. 'On Justifying Punishment: The Discrepancy between Words and Actions.' *Social Justice Research* 21.2 (2008): 119–37. DOI: 10.1007/s11211-008-0068-x.
- , et al. 'Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment.' *Journal of Personality and Social Psychology* 83.2 (2002): 284–99.
- Clutton-Brock, T. and G. Parker. Punishment in Animal Societies. *Nature.com*, 1995.
- Costanzo, P., et al. 'A Reexamination of the Effects of Intent and Consequence on Children's Moral Judgments.' *Child Development* 44.1 (1973): 154–61.
- Cushman, F. A. 'Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment.' *Cognition* 108.2 (2008): 353–80. DOI: 10.1016/J.Cognition.2008.03.006.
- . 'Action, Outcome, and Value: A Dual-System Framework for Morality.' *Personality and Social Psychology Review* 17.3 (2013): 273–92. DOI: 10.1177/1088868313495594.
- Cushman, F. A., A. J. Durwin, and C. Lively. 'Revenge without Responsibility? Judgments about Collective Punishment in Baseball.' *Journal of Experimental Social Psychology* 48.5 (2012): 1106–10. DOI: 10.1016/j.jesp.2012.03.011.
- Cushman, F. A., R. Sheketoff, S. Wharton, and S. Carey. 'The Development of Intent-Based Moral Judgment.' *Cognition* 127.1 (2013): 6–21. DOI: 10.1016/J.Cognition.2012.11.008.
- Darley, J. M., et al. 'Incapacitation and Just Deserts as Motives for Punishment.' *Law and Human Behavior* 24.6 (2000): 659–83.
- Dreber, A., et al. 'Winners Don't Punish.' *Nature Materials* 452.7185 (2008): 348–51. DOI: 10.1038/nature06723.
- Elster, J. *Ulysses and the Sirens*. Cambridge: Cambridge University Press, 1979.
- Fehr, E. and U. Fischbacher. 'Third-Party Punishment and Social Norms.' *Evolution and Human Behavior* 25.2 (2004): 63–87. DOI: 10.1016/S1090-5138(04)00005-4.
- Fehr, E. and S. Gächter. 'Altruistic Punishment in Humans.' *Nature Materials* 415 (2002): 137–40.
- Fincham, F. and C. Roberts. 'Intervening Causation and the Mitigation of Responsibility for Harm Doing. II: The Role of Limited ...' *Journal of Experimental Social Psychology* 21 (1985): 178–94.
- Fowler, J. H. 'Altruistic Punishment and the Origin of Cooperation.' *Proceedings of the National Academy of Sciences of the United States of America* 102.19 (2005): 7047–9. DOI: 10.1073/pnas.0500938102.
- Frank, R. H. *Passion Within Reason: The Strategic Role of the Emotions*. New York: Norton, 1988.
- Fukuyama, F. *The Origins of Political Order: From Prehuman Times to the French Revolution*. Profile Books: Columbia University Press, 2011.
- Gächter, S., E. Renner, and M. Sefton. 'The Long-Run Benefits of Punishment.' *Science* 322.5907 (2008): 1510. DOI: 10.1126/science.1164744.
- Gintis, H. 'Strong Reciprocity and Human Sociality.' *Journal of Theoretical Biology* 206.2 (2000): 169–79. DOI: 10.1006/jtbi.2000.2111.
- Gintis, H., et al. 'Explaining Altruistic Behavior in Humans.' *Evolution and Human Behavior* 24.3 (2003): 153–72.
- Graham, J., J. Haidt, and B. A. Nosek. 'Liberals and Conservatives Rely on Different Sets of Moral Foundations.' *Journal of Personality and Social Psychology* 96.5 (2009): 1029–46. DOI: 10.1037/a0015141.
- Gray, K., C. Schein, and A. F. Ward. *The Myth of Harmless Wrongs in Moral Cognition: Automatic Dyadic Completion from Sin to Suffering*. nd.
- Greene, J. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. New York, 2013.
- Haidt, J. *The Righteous Mind*. Pantheon: New York, 2012.
- Hebble, P. W. 'Development of Elementary School Children's Judgment of Intent.' *Child Development* 42.4 (1971): 583–8.
- Henrich, J. and R. Boyd. 'Why People Punish Defectors Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas.' *Journal of Theoretical Biology* 208.1 (2001): 79–89. DOI: 10.1006/jtbi.2000.2202.
- Henrich, J., et al. 'Markets, Religion, Community Size, and the Evolution of Fairness and Punishment.' *Science* 327.5972 (2010): 1480. DOI: 10.1126/science.1182238.
- . 'Costly Punishment across Human Societies.' *Science* 312.5781 (2006): 1767–70.

- Hilbe, C., et al. 'Democratic Decisions Establish Stable Authorities that Overcome the Paradox of Second-Order Punishment.' *Proceedings of the National Academy of Sciences* 111.2 (2014): 752–6.
- Hirshleifer, D., and E. Rasmusen. 'Cooperation in a Repeated Prisoners' Dilemma with Ostracism.' *Journal of Economic Behavior & Organization* 12.1 (1989): 87–106.
- Imamoglu, O. 'Children's Awareness and Usage of Intention Cues.' *Child Development* 46 (1975): 39–45.
- Karniol, R. 'Children's Use of Intention Cues in Evaluating Behavior.' *Psychological Bulletin* 85.1 (1978): 76–85.
- Killen, M., et al. 'ScienceDirect – Cognition, Volume 119, Issue 2, Pages 149–312 (May 2011).' *Cognition* (2011). <http://www.sciencedirect.com.ezp-prod1.hul.harvard.edu/science?\_ob=PublicationURL&\_tockey=%23TOC%234908%232011%23998809997%233038760%23FLA%23&\_cdi=4908&\_pubType=J&\_auth=y&\_acct=C000014438&\_version=1&\_urlVersion=0&\_userid=209690&md5=f1628babb1b6c3645fb6ec3b2512d752>
- Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *Journal of personality and social psychology*, 95(4), 826.
- Krasnow, M. M., et al. 'Meeting Now Suggests We Will Meet Again: Implications for Debates on the Evolution of Cooperation.' *Scientific Reports* 3 (2013).
- Malle, B. F., S. Guglielmo, and A. E. Monroe. *A theory of blame. Psychological Inquiry*, 25(2) (2014): 147–86.
- Martin, J., and F. Cushman. 'Why we forgive what can't be controlled.' Unpublished manuscript.
- McCullough, M. E., R. Kurzban, and B. A. Tabak. 'Cognitive Systems for Revenge and Forgiveness.' *Behavioral and Brain Sciences* 36.1 (2013): 1–15.
- Mikhail, J. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press, 2011.
- . Moral grammar and intuitive jurisprudence: a formal model of unconscious moral and legal knowledge. *Psychology of learning and motivation*, 50 (2009): 27–100.
- Miller, W. I. *Bloodtaking and Peacemaking: Feud, Law, and Society in Saga Iceland*. University of Chicago Press: Chicago, 1990.
- Nagel, T. *Mortal Questions*. Cambridge: Cambridge University Press, 1979.
- Nakao, H., and E. Machery. 'The Evolution of Punishment.' *Biology & Philosophy* 27.6 (2012): 833–50.
- Nikiforakis, N., and H. Mitchell. 'Mixing the Carrots with the Sticks: Third Party Punishment and Reward.' *Experimental Economics* 17.1 (2013): 1–23. DOI: 10.1007/s10683-013-9354-z.
- Nisbett, R. E., and D. Cohen. *Culture of Honor: The Psychology of Violence in the South*. Boulder: Westview Press Inc., 1996.
- Nisbett, R., and T. Wilson. 'Telling More than We Can Know.' *Psychological Review* 24.3 (1977): 231.
- Nobes, G., G. Panagiotaki, and C. Pawson. 'The Influence of Negligence, Intention and Outcome on Children's Moral Judgments.' *Journal of Experimental Child Psychology* 104 (2009): 382–97.
- Noë, R., C. P. van Schaik, and J. A. van Hooff. 'The Market Effect: An Explanation for Pay-off Asymmetries among Collaborating Animals.' *Ethology* 87.1–2 (1991): 97–118.
- Pedersen, E. J., R. Kurzban, and M. E. McCullough. 'Do Humans Really Punish Altruistically? A Closer Look.' *Proceedings of the Royal Society B: Biological Sciences* 280.1758 (2013): 20122723.
- Raihani, N. J., A. Thornton, and R. Bshary. 'Punishment and Cooperation in Nature.' *Trends in Ecology & Evolution* 27.5 (2012): 288–95.
- Rand, D. G., and M. A. Nowak. 'The Evolution of Antisocial Punishment in Optional Public Goods Games.' *Nature Communications* 2 (2011): 434.
- Roberts, G. 'Competitive Altruism: From Reciprocity to the Handicap Principle.' *Proceedings of the Royal Society of London, Series B: Biological Sciences* 265.1394 (1998): 427–31.
- Robinson, P. H., and J. M. Darley. *Justice, liability, and blame: Community views and the criminal law*. Boulder CO: Westview Press, 1995.
- Robinson, P. H., and R. Kurzban. 'Concordance & Conflict in Intuitions of Justice.' *Minnesota Law Review* 91.6 (2007): 1633.
- Shultz, T. 'Assignment of Moral Responsibility and Punishment.' *Child Development* (1986): 177–184.
- Tinbergen, N. 'Derived Activities: Their Causation, Biological Significance, Origin and Emancipation during Evolution.' *Quarterly Review of Biology* 27 (1952): 1–32.
- Turiel, E. *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press, 1983.
- Weiner, B. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. New York: Guilford Press, 1995. xvi, 301.
- West, S. A., A. S. Griffin, and A. Gardner. 'Social Semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection.' *Journal of Evolutionary Biology* 20.2 (2007): 415–32. DOI: 10.1111/j.1420-9101.2006.01258.x.
- Williams, B. *Moral Luck*. Cambridge: Cambridge University Press, 1981.
- Young, L., et al. 'Damage to ventromedial prefrontal cortex impairs judgment of harmful intent.' *Neuron* 65 (2010a): 1–7.
- Young, L., et al. 'Disruption of the Right Temporoparietal Junction with Transcranial Magnetic Stimulation Reduces the Role of Beliefs in Moral Judgments.' *Proceedings of the National Academy of Sciences* 107.15 (2010b): 6753.

- Young, L., et al. 'The Neural Basis of the Interaction between Theory of Mind and Moral Judgment.' *Proceedings of the National Academy of Sciences* 104.20 (2007): 8235–40. DOI: 10.1073/Pnas.0701408104.
- Young, L., and R. Saxe. 'The neural basis of belief encoding and integration in moral judgment.' *NeuroImage* 40 (2008): 1912–20.
- . 'When Ignorance Is No Excuse: Different Roles for Intent across Moral Domains.' *Cognition* 120.2 (2011): 202–14.
- Yuill, N. and J. Perner. 'Intentionality and Knowledge in Children's Judgments of Actor's Responsibility and Recipient's Emotional Reaction.' *Developmental Psychology* 24.3 (1988): 8358–65.
- Zelazo, P., C. Helwig, and A. Lau. 'Intention, Act, and Outcome in Behavioral Prediction and Moral Judgment.' *Child Development* 11(1) (1996): 37–63.