In press, R. Joyce, K. Sterelny, B. Calcott & B. Fraser (Eds.) *Signaling, Commitment and Emotion, Vol. 2: Psychological and Environmental Foundations of Cooperation* MIT Press

The role of learning in punishment, prosociality, and human uniqueness

Fiery Cushman

1. Introduction

At your local natural history museum rows of tiny dripping noses press on display cases, peering at the impalas, the grizzlies, and the komodo dragon. Like the glass that separates stuffed noses from stuffed animals, something separates humans from other animals—something substantial, but hard to see. An alchemic combination of accumulated change must explain language, science, culture, art and civilization; in short, why humans build museums and the other animals inhabit them.

This essay focuses on just one ingredient of that alloy: the uniquely rich, complex and successful range of human prosocial behaviors. Even more narrowly, it focuses on the role of punishment in enforcing prosociality. In its approach, however, it aims for a broader insight: to illustrate the important relationship between abstract evolutionary models of behavior and the specific psychological mechanisms that actually produce behavior. This natural union improves evolutionary models, clarifies the structure of psychological mechanisms, and helps to reveal the foundations of human uniqueness.

Evolutionary theorists posit a simple relationship between punishment and prosocial[1] behavior (e.g. Boyd & Richerson, 1992; Clutton-Brock & Parker, 1995). In a population where some individuals punish antisocial behavior, it pays to be prosocial selectively with the punishers. And, in a population where some individuals behave prosocially only when threatened with punishment, it pays to punish antisociality. Put in concrete terms: I ought to stop stealing from you if you hit me in retaliation, and therefore the immediate costs of retaliation might be worth the long-term benefit of securing your property. This co-dependent relationship between punishment and prosocial behavior is well understood.

Far less understood are the psychological mechanisms that actually produce prosocial and punitive behaviors. At first glance, one might assume that psychology does not matter to the larger evolutionary question. Can't we understand and model the abstract relationships between evolutionary strategies without troubling ourselves with their psychological implementation? By analogy, formal models of economics do not concern themselves with the molecular structure of coins or bills.

This is a seductive perspective. It certainly simplifies the task of modeling the evolution of social behavior to ignore the underlying mechanisms. Ultimately, however, it is deeply flawed. The functional design of punishment and prosociality depend critically on psychological details, in much the same way that formal economic models cannot ignore the peculiar irrationalities of human actors (Tversky & Kahneman, 1981). Psychological details explain when, how and who organisms decide to punish. They explain why punitive strategies (and especially reactive aggression) are more often observed than rewarding strategies among non-human animals. And, they explain the pervasive, complex and flexible nature of human social behavior.

Of course, the benefits of integration run the opposite way as well. Our understanding of the psychological mechanisms supporting punitive and prosocial behavior are enriched by considering their functional design.

This essay takes up both challenges: first, to demonstrate how the functional design of punishment and prosociality mirror psychological constraints, and second to demostrate how the psychological mechanisms underlying punishment and prosociality are illuminated by considering their functional design. I argue that punishment is a specialized behavioral adaptation that exploits the ability of social partners to learn. Implicit in this argument is a distinction between specialized behavioral adaptations that function in a fixed manner in limited contexts, and general mechanisms of learning and behavioral choice that function flexibly across diverse contexts (Fodor, 1983;

[1] In this essay, I define prosocial behavior as a behavior that has worse fitness consequences than some alternative (the antisocial choice) for the agent, while having better fitness consequences than the antisocial alternative for some social partner. Grooming, sharing food, and alarm calling could all be prosocial behaviors, on this definition. Choosing not to steal food from a social partner, or not to encroach on his territory, could also be prosocial behaviors. In accounting for the costs and benefits of a prosocial behavior, I am specifically excluding the contingent response of a social partner (e.g. punishment or reciprocation). Thus, grooming with the expectation of reciprocation, or respecting another's property under the threat of punishment, both count as prosocial behaviors.

Hirschfeld & Gelman, 1994; Shiffrin & Schneider, 1977; Spelke, 2000). Many organisms can flexibly learn to avoid behaviors that have negative consequences: Eating toxic foods, approaching open flames, jumping in cold water, etc. This general capacity for learning presents an opportunity for social exploitation. Specifically, organisms can use punishment to teach social partners to act prosocially by exploiting their general learning capacity. As we will see, the evolutionary dynamics of this relationship between punishment and prosociality make it likely that punishment will operate via a specialized behavioral adaptation that is relatively fixed and limited, while prosociality will be supported in part by general, flexible mechanisms of learning and behavioral choice.

The general ability to learn associations between behavior and consequence is highly constrained in most organisms, however. For instance, the consequence must be rapidly and salient in order for learning to occur (Mackintosh, 1975; Renner, 1964; Rescorla & Wagner, 1965; Schwartz & Reisberg, 1991). Thus, psychological constraints on learning will influence the functional design of punishment. In the abstract, any sort of punishment could motivate any sort of prosocial behavior. In reality, punishment must conform itself to the circumscribed ability of organisms to learn.

There is one species with a substantially expanded learning abilities: humans. It is no accident, therefore, that we also exhibit a uniquely flexible and productive prosocial behavior. To the extent that punishment (and also reciprocity and reward) depend on general learning mechanisms to motivate prosocial action, humans' uniquely powerful capacities in learning, reasoning and behavioral choice stand to vastly expand their range of prosocial behavior beyond non-human animals'.

I conclude by considering an irony of the human situation: Our punitive instincts (and possibility our instincts for reciprocity and reward) may not have "caught up" with our new capacities for learning, reasoning and deciding. In some respects, the functional design of human punishment may still be adapted to the substantially more limited minds of our non-human ancestors.

2. Specialized versus general mechanisms in psychology

Psychological detail is starkly absent in most evolutionary models of punishment and prosociality. In these models a population of agents interacts and reproduces, leading to evolutionary change over successive generations. The modeler specifies a set of behavioral strategies that the agents can employ. For instance, possible strategies include "always behave prosocially", "punish antisocial behavior with 80% probability", and "cease antisocial behavior if it has been punished twice". These strategies are typically specified in very abstract terms, which is appropriate to the task of creating formal models that can be generalized across diverse cases. But there are many different ways that these abstract strategies could be implemented at a psychological level. Below, I describe a very coarse distinction between two classes of psychological mechanism: general mechanisms of associative learning and behavioral choice versus specialized behavioral adaptations.

Consider a rat that pushes a lever in order to avoid an electric shock. Decades of psychological research suggests that the rat's learned behavior is guided by something like a simplified calculation of expected value (reviewed in Daw & Doya, 2006). That is, the rat learns to associate certain behavioral choices with their likely future consequences in terms of subjective value, conditioned on some set of perceptual inputs. It then selects behaviors as a function of the value associated with each. The rat continuously updates these associations as it experiences punishments and rewards following its behavior. Critically, it has broad flexibility[2] in the kinds of associations that it can form. This allows it to adaptively adjust its behavior, guiding it towards optimal patterns of choice.

At the other extreme, a behavioral strategy can depend on an innate, rigid psychological response; what is called a "fixed action pattern" in behavioral biology and ethology. A classic example of a fixed action pattern is the motor routine by which a goose retrieves an egg that rolls out of its nest (Lorenz & Tinbergen, 1938). This behavior does not seem to be learned and regulated by general, flexible cognitive processes employing associative learning or value maximization. Rather, geese appear to have an innate mechanism that recognizes the perceptual input of an egg rolling out of the nest and automatically triggers a highly specific motor routine for retrieval.

Interpreted literally, most formal models of the evolution of social behavior use behavioral strategies like the fixed action pattern of the goose. They are innate, fixed over the course of the lifetime, and do not involve associative learning or the computation of expected value (Boyd & Richerson, 1992; Clutton-Brock & Parker, 1995; Maynard Smith, 1982; Nowak, 2006). A handful of studies, however, model social behavior using general, flexible learning mechanisms something like the rat (Gutnisky & Zanutto, 2004; Macy & Flache, 2002; Nowak & Sigmund, 1993). These models demonstrate that, in principle, it is possible to achieve prosociality without the biological evolution of a domain-specific prosocial strategy.

So, which is the more accurate model of social behavior: the goose, or the rat? That question

[2] Broad, but certainly not unlimited (Garcia & Koelling, 1996).

motivates much of the remainder of this essay. Before charging into the fray, it will help to arm ourselves with two general observations.

First, general learning mechanisms are "free" from an adaptive perspective. Basic mechanisms supporting associative learning and reward-maximizing choice—the essence of operant conditioning—exist in fruit flies, zebra fish, pigeons, rats, sophomores, and virtually animal in between. Thus, if we can explain some behavior in terms of general learning, then it will usually not be necessary to postulate any further adaptation. For instance, imagine that we observe a dog retrieve a newspaper. This behavior might the product of general learning, or it might be the product of a specialized adaptation. All else being equal, if the behavior can be explained in terms of general learning processes, this hypothesis is more likely than the alternative hypothesis that the dog has a specialized adaptation for newspaper retrieval. The general learning hypothesis comes for free, whereas the specialized adaptation hypothesis requires some additional evolutionary event. This is an argument from parsimony.

Second, general learning mechanisms show characteristic constraints. First, and most obviously, general learning mechanisms require experience. Rats aren't born knowing when to push a lever; this behavior is only acquired given sufficient experience. Moreover, successful association between stimuli, behavioral choice, and punishment or reward requires special conditions. For instance, learned associations typically require (1) salient events that occur within (2) a relatively short period of time (Mackintosh, 1975; Renner, 1964; Rescorla & Wagner, 1965; Schwartz & Reisberg, 1991). These constraints may explain why geese do not rely on general learning processes to acquire egg-retrieval behavior. The relevant experiences are probably infrequent, and each negative experience is very costly to reproductive success. The feedback (one less chick born than egg laid) is probably not very salient to goose, and it comes only after a long temporal delay.

In summary, general learning processes are adaptively "free", but mechanistically constrained. By contrast, specific behavioral adaptations require new adaptive events, but they can move beyond the constraints of general learning processes. With these considerations in mind we can assess, first, whether punishment and prosociality are more likely to be supported by specific behavioral adaptations versus general learning processes and, second, how these psychological details affect the functional design of punishment.

3. The evolutionary dynamics of punishment and prosociality

The next two sections argue from complimentary perspectives that we should expect organisms to punish like geese (using specialized adaptations) and behave prosocially like rats (using general mechanisms of learning and choice). Some technical detail is required, but I will begin by sketching the argument from evolutionary dynamics in broad strokes before painting in the freckles and flies.

First, consider punishment. If punishment depends on a specialized mechanism, it can be inflexible: We might punish certain situations no matter how we estimate the costs or benefits. By contrast, if punishment depends on general learning mechanisms, it will be flexible: We will only punish situations when the anticipated benefits outweigh the anticipated costs. It turns out that there is a great benefit to inflexible punishment. Think of an unruly toddler who throws a tantrum whenever his parents attempt discipline: If parents flexibly decide whether to punish (like the rat), then they may conclude that the costs are greater than the benefits and give up. In this case, the toddler can strategically persist in misbehavior until his parents learn that punishment is hopeless—so much the worse for the parents. But if the parents are inflexibly committed to punishment, the toddler cannot profit persistent tantrums. Rather, his best strategy is to behave. Hence, it pays to punish inflexibly, like the goose.

Behavioral flexibility is favored for prosociality, however. Imagine that the toddler's parents wisely adopt a strategy of inflexible punishment for misbehavior, but that his grandmother is a doting pushover. If the toddler is an inflexible devil, he benefits with grandma but pays the costs of punishment with his parents. If the toddler is an inflexible angel, he benefits with his parents but misses the valuable opportunity to exploit grandma's dotage. The optimal path for this toddler is to flexibly adopt prosociality depending on the costs and benefits: behave around the parents, misbehave with grandma. In short, it pays to adopt prosociality flexibly, like a rat .Now, it is possible to imagine a specialized adaptation that facilitates this contingent choice, finely tuned to distinguishing the enforcers from the pushovers and selecting innately specified appropriate behavioral responses to each. General learning mechanisms provide this behavior strategy "for free", however, using past experience to learn when to be naughty and when to be nice.

With this rough argument on the table, I'll now turn to a more detailed consideration of the evolutionary dynamic between punishment and prosociality and the psychological mechanisms that we might expect to support each. It will help to give labels

to the strategies we have considered. Imagine an interaction in which an agent (A) can act either prosocially or antisocially, and in response a partner (P) can either punish or not punish. Here are the six strategies that we must consider:

Fixed Prosociality: A behaves prosocially towards P

Fixed Antisociality: A behaves antisocially towards P

Contingent Prosociality: A behaves prosocially towards P only if P punishes antisocial behavior

Fixed Punishment: P always responds to A's antisocial behavior by punishing A

Fixed Non-Punishment: P never responds to A's antisocial behavior by punishing A

Contingent Punishment: P responds to A's antisocial behavior by punishing A only if this tends to decrease A's antisocial behavior

I will assume that punishment is more costly than non-punishment and prosociality is more costly than antisociality (setting aside the future benefits of social partners' behavior).

A "rat-like" learner that maximizes expected utility based on past experiences will adopt contingent prosociality (only pay the costs of prosocaility if antisociality carries the greater cost of punishment) and contingent punishment (only pay the costs of punishment if it succeeds in inducing prosocial behavior). Meanwhile, it is possible to imagine "goose-like" fixed action patterns that are either fixed or contingent in their operation. As noted in the last section, explaining contingent behavior by "adaptively free" general learning mechanisms is more parsimonious than invoking a specialized behavioral adaptation, all else being equal. Thus, I will be treating contingent strategies as products of general learning and choice mechanisms, and fixed strategies as products of specialized adaptations.

The following sections ask whether the co-evolutionary dynamic between prosociality and punishment can emerge from various combinations of the strategies listed above. In particular, they ask whether a population of *antisocial non-punishers* can be successfully invaded to yield a population of *prosocial punishers*. That is, can prosociality and punishment jointly arise where neither existed before?

3.1 Fixed prosociality and fixed punishment

A population of antisocial non-punishers is unlikely to be invaded by either fixed prosociality or fixed punishment. The combination of these strategies does not provide a reliable path towards prosociality enforced by punishment.

To begin with, fixed prosociality is a clear loser in a population that never punishes antisocial behavior—this strategy pays the costs of generosity with no contingent benefit. On the other hand, fixed prosociality would be favored over fixed antisociality in a population of fixed punishers. In this population, fixed prosociality avoids the costs of punishment for antisocial behavior[3]. So, could a population of fixed punishers emerge and maintain stability? No: fixed punishment is disfavored whenever agents adopt prosociality or antisociality as fixed strategies. The fixed punishment strategy pays an extra cost (of punishment) whenever it encounters an antisocial agent, but this extra cost does not yield any contingent future benefit. The fixed antisocial agent is just as likely to adopt antisociality in future interactions, whether or not P adopts a punitive strategy. Likewise, fixed prosocial agents are just as likely to adopt prosocial behavior in future interactions, again whether or not P adopts a punitive strategy. Thus, the costs associated with punishment yield no selective benefit to punishers when either fixed prosociality or fixed antisociality dominates.

One helpful lens to apply to this interaction is the tragedy of the commons. Over time a sufficient number of punishers can make fixed prosocialty stable, but they pay a cost to do so. Unfortunately, their costly efforts are exploitable by non-punishers, who avoid the costs of punishment but equally reap the benefits of prosociality.

3.2 Contingent prosociality and fixed punishment

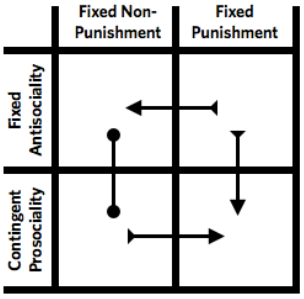
One solution to this tragedy of the commons is clear: Punishment will be favored if it yields *exclusive* benefits to the punisher. This condition is met when social partners engage in contingent prosociality. That is, if agents apply the rule, "Only act prosocially when it is enforced by punishment", and if they are sufficiently adept at discriminating between social partners, then they will end up adopting prosociality only in interactions with individuals who punish. So, is it possible for contingent prosociality and fixed punishment to invade a population of antisocial non-punishers?

Figure 1 charts the evolutionarily dynamics between these two strategies. First, consider a social environment in which neither strategy is employed

[3] Note, however, that contingent prosociality performs better still. A fixed prosocial strategy avoids punishment by punishing social partners, but misses the opportunity to exploit non-punitive partners by behaving antisocially selectively with them. By contrast, a contingently prosocial strategy avoids punishment by punishers, while selectively exploiting non-punishers.

(upper left corner). Clearly, punishment will not be favored in this environment; it pays the costs of punishing antisocial acts without any contingent benefits. However, the strategy of contingent prosociality is neutral in this environment. An individual that adopts contingent prosociality will never experience punishment for its antisocial actions, and therefore will consistently behave in an antisocial manner (assumed to be fitness-maximizing). If the strategy of contingent prosociality attained sufficient frequency in a population, the strategy of contingent punishment can then also be favored (lower left corner). This presents a plausible path from a state in which there is neither punishment nor contingent prosociality to a state in which there is both punishment and contingent prosociality (from the upper left to lower right of Figure 1), which is evolutionarily stable. The key first step is the emergence of a strategy that responds to the punishment of antisocial behavior by switching to prosocial behavior.

Figure 1.



As described above, it is possible that contingent prosociality could be supported either by general learning mechanisms or by a specialized adaptation. Which is more likely? Again, we return to argument by parsimony: General learning processes are adaptively free (in the sense that they don't require any new adaptive event) — they are sitting, waiting to be exploited. By contrast, a specialized adaptation for contingent prosociality is relatively unlikely, especially because there is no selective pressure favoring such an adaptation prior to the emergence of punishment.

This point is best appreciated in concrete terms. Imagine a population of rats that have no specialized adaptations for punishment or contingent prosociality, but do have standard, general learning mechanisms. A punitive strategy that induces prosociality by exploiting rats' general learning mechanisms can immediately invade this population. By contrast, a punitive strategy that induces prosociality by exploiting a specialized adaptation for contingent prosociality must wait until such a specialized adaptation emerges. As described above, the emergence of such a specialized adaptation is not disfavored in fitness terms; it is a neutral change. But, the emergence of such a specialized innate behavior is unlikely given the absence of any selective pressure that favors it.

Now, once punishment has invaded a population, it becomes relatively easier for specialized adaptations for contingent prosociality to emerge. Consider again a hypothetical population of rats. Suppose that a punitive strategy has invaded this population by exploiting the general learning processes available in the population. But recall that general learning processes are characterized by certain constraints: the rat must experience sufficient punishment, this punishment must be timely and salient, etc. Within this social environment of punishment, there is a selective pressure for the emergence of specialized mechanisms that detect punishment and respond contingently with prosocial behavior more quickly and reliably than generalized learning processes (Cushman & Macendoe, 2009). In this case, general learning mechanisms establish an initial behavioral repertoire that facilitates the subsequent emergence of specialized adaptations. The ability of general learning mechanisms to pave the way towards specialized adaptations is a well-studied evolutionary phenomenon known as the "Baldwin Effect". Thus, while I have argued in this section that punitive strategies are most likely to have emerged by exploiting contingent prosociality as a property of general learning processes, the present dynamic of punishment and prosociality may be supported by more specialized psychological mechanisms of contingent prosociality, or by some mix of specialized and general mechanisms. I return to this point in Section 6.

3.3 Contingent punishment

So far I have argued that contingent prosociality is necessary for *fixed* punitive strategies to be successful and, moreover, that contingent prosociality is relatively more likely to be supported by general learning mechanisms. I now turn to the same question regarding punishment. Can *contingent* punishment coevolve with prosociality?

On its face, contingent punishment looks superior to fixed punishment. Whereas fixed punishment pays the cost of punishing individuals who never respond with contingent prosociality (fixed antisocial actors), contingent punishment avoids these costs. Simply put, contingent punishers learn not to bother punishing where it can't help, and focus the costs of punishment solely where the maximize benefits: changing the behavior of contingent prosocialists.

Despite these apparent advantages, however, contingent punishment does not provide a reliable path toward prosociality. The difficulty is that, faced with contingent punishment, both fixed and contingent prosociality are disfavored strategies. Rather,

individuals do best by adopting fixed antisocial behavior. After all, a purely antisocial actor can largely avoid punishment in an environment dominated by contingent punishment by teaching social partners, "don't bother punishing me". It thereby reaps the rewards of antisociality, while avoiding the costs of punishment. Thus, the strategy of contingent punishment will tend to suppress prosociality, rather than to promote it.

More formally, this argument has two facets. First, in a population dominated by fixed antisocial behavior, the emergence of contingent punishment will not promote prosocial behavior. This follows directly from the logic of the previous paragraph. Neither fixed nor contingent prosociality outperforms fixed antisociality when played against contingent punishment.

Second, in a population that has achieved the fixed punishment / contingent prosociality dynamic described in the previous section, contingent punishment may be unable to invade. Specifically, contingent punishment is at best neutral, and possibly inferior, compared with fixed punishment. If contingent punishers must learn to adopt punishment, then they suffer the costs of this learning process (lost opportunities for prosociality obtained via punishment) compared with fixed punishers, who adopt the optimal punitive strategy immediately. At best, if contingent punishers have a strong initial bias towards punishment and a capacity to *un*learn punishment if it is unsuccessful, then they fare no worse (but no better) than fixed punishers when playing against contingent prosocialists. Critically, the only circumstance in which contingent punishment outperforms fixed punishment is when playing against a *fixed antisocial* partner. However, a population dominated by fixed punishers presents an extremely unforgiving environment for fixed antisocial players. Thus, fixed punishment sustains a social environment unfavorable to invasion by contingent punishment.

In summary, a stable dynamic between prosocial and punitive behavior requires inflexible punishment. This requirement of inflexibility among certain behavioral strategies is well-recognized (Frank, 1988). In principle, a sufficiently sophisticated cognitive mechanism capable of general strategic reasoning could recognize this inflexibility requirement and adopt fixed prosociality. Humans sometimes do this: the doctrine of "mutual assured destruction" is an example. But the kind of general learning mechanisms possessed by most animals are unlikely to support this kind of abstract strategic reasoning. Rather, these learning mechanisms will tend to support punitive behavior only when it demonstrably promotes prosocial behavior: i.e., flexibly. Consequently, the punitive behaviors of non-humans are more likely to be the product of a specialized adaptation resembling "fixed punishment" than the product of general learning processes.

As we have seen, the evolutionary dynamics of co-dependent punishment and prosociality suggest that punishment more likely emerged as a specialized adaptation, whereas contingent prosociality was more likely initially supported by general learning processes. To put it another way, punishment is a mechanism that exploits general learning processes; it gets social partners to adopt prosocial behavior roughly by operant conditioning. This sets up a clear prediction about the functional design of punishment. Punishment should be designed to match the constraints of general learning processes, obtaining the maximum response from social partners at the minimum cost. Thus, to understand the functional design of punishment we will need to understand the psychology of learning. Section 5 reviews experimental evidence that supports this functional match, and Section 7 considers its relevance to human behavior in particular.

First, however, Section 4 illustrates how certain structural aspects of many social interactions also favor specialized adaptation or general learning processes for punishment and prosociality. This discussion depends not on considerations of evolutionary dynamics and arguments from parsimony, but rather on the other conceptual tool we established in Section 2: the psychological constraints characteristic of general learning processes.

4. The structure of social interactions and the constraints of learning

As noted above, general learning processes typically require that the reinforcement of actions occurs in a relatively quick and salient manner (Mackintosh, 1975; Renner, 1964; Rescorla & Wagner, 1965; Schwartz & Reisberg, 1991). For instance, if you want to train your dog to pick up the newspaper, it makes more sense to reward her with a biscuit each time she fetches than with a new collar for Christmas. General learning mechanisms are simply not sufficient to associated a paper fetched in May with an new neck accessory in December.

Turning to the relationship between punishment and prosociality, certain kinds of social interactions provide the opportunity for quick, salient reinforcement, while others do not. This provides an additional basis on which to predict whether, and when, punishment and prosociality are likely to be supported by general learning processes versus specialized adaptations. By analogy, if dogs readily adopt some behavior incentivized by immediate biscuits, it is plausible that their behavior depends on general learning mechanisms. But, if dogs adopts some behavior incentivized by collars at Christmas, it is unlikely that that behavior was learned by general

mechanisms—rather, it indicates a specialized adaptation.

Consider again the social interaction in which an agent (A) harms a social partner (P). Possible harmful actions by A include aggression, resource theft, territorial violation or sexual contact, for instance. Then, P's punishes A by physical aggression. As a consequence, A does not perform this harmful action in future encounters. As above, we are faced with two questions. First, did A adopt prosociality because of a general learning mechanism or a specialized adaptation? Second, did P punish because of a general learning mechanism or a specialized adaptation? I will approach these questions in the following way: Is it plausible that each of these behaviors could be product of general learning mechanisms, given the constraint that those mechanisms require quick and salient feedback?

In order for A to adopt prosociality via general learning processes, P's punishment should follow A's harmful action very quickly. If we assume that P is present when the harmful action occurs, there is no obstacle to rapid punishment: P can initiate physical aggression towards A immediately. On the other hand, relatively more delayed forms of punishment will be disfavored. For instance, if P punishes A by destroying a food resource of A's several days hence, A is relatively less likely to associate this punishment with her prior harmful act.

Similarly, in order for P to adopt punishment via general learning processes, A's desisting from future harm must follow P's punishment very quickly. In some cases this criterion will be easy to meet. For instance, if A is encroaching on P's territory and P's punishment drives A away without resistance[4], this positive reinforcement of P's punishment occurs immediately. However, in other cases this criterion will be hard to meet. For instance, if A consumes some resource of P's and P punishes A, the positive reinforcement of P's punishment only occurs at some point in the future when A next has the opportunity to steal resources from P, but instead desists. In this case, the basic temporal structure of the social interaction — the fact that opportunities for A to steal from P arise only occasionally — makes it relatively easier for A to adopt prosociality via general learning processes (because P's punishment can be immediate and takes the form of a relatively salent aggressive action) but relatively harder for P to adopt punishment via general learning processes (because A's prosociality must be delayed, and takes the form of a relatively non-salient "omission" of harm).

To summarize, antisocial acts can be punished immediately, facilitating learning. But the value of punishment is harder to learn, because the behavioral changes it promotes follow at a longer temporal delay. Thus, general mechanisms of learning and behavioral choice may be sufficient to support prosocial action, whereas specialized psychological mechanisms may be required to support punishment.

Similar considerations may help to resolve a puzzle concerning punishment. A provocative series of studies shows that it is preferable to promote prosocial behavior in social partners by withholding aid from antisocial actors, rather than actively punishing antisocial actors (Dreber, Rand, Fudenberg, & Nowak, 2008; Rand, Dreber, Ellingsen, Fudenberg, & Nowak, 2009; Rand, Ohtsuki, & Nowak, 2009). These studies model dyadic social interactions in which each actor has three choices of behavior. She can cooperate with her partner (e.g. share food), which is costly to herself and beneficial to their partner. She can defect against her partner (e.g. withhold food sharing), which is costless to herself and yields no benefit to her partner[5]. Or, she can punish her partner (e.g. by physical attack), which is costly to herself and even more costly to her partner. In brief, it turns out that both individual- and group-level fitness is maximized when players respond to defection with reciprocal defection. Both individual- and group-level fitness is lower when players respond to defection with punishment. This finding is puzzling because it predicts that we should not observe costly punishment as a response to antisocial behavior — rather, we should observe reciprocal antisociality. Yet, costly punishment is a common response to antisocial behaviors both in experimental and natural settings (Clutton-Brock & Parker, 1995; Fehr & Gachter, 2002; Henrich, et al., 2005). Why?

This puzzle may be partially explained by the considerations introduced in this section: the structure of social interactions and the constraints of general learning mechanisms. When A defects against P, P may have an immediate opportunity to respond with costly punishment of A — for instance, by physical aggression. By contrast, the opportunity for P to respond with defection against A is necessarily delayed until a circumstance arises in which P has an opportunity to cooperate with A. The temporal delay imposed by the structure of the social interaction may prevent reciprocal defection from effectively exploiting general learning processes to promote future prosociality.

[4] Of course, if A retaliates against P's punishment, then the most likely learned association for P in the short-term is: punishing A is costly and harmful.

[5] Of course, to respond to defection with defection might be regarded as punishment of a sort—call it "passive" punishment (failing to provision a benefit), and contrast it with "active" punishment (imposing a cost). When using the term "punishment" in this chapter, it is the active sort that I have in mind. Part of the argument of this section is that active punishment can more salient and rapid, and therefore more learnable, than passive punishment (i.e. defection).

Of course, the same point applies if we re-frame "reciprocal defection" as "reciprocal altruism"—the identical strategy framed in terms of prosociality-for-prosociality, rather than antisociality-for-antisociality. Acts of reciprocal prosociality must away *opportunities* for reciprocal prosociality arise. If you share food with me, for instance, I may not be able to reward this prosocial action until a situation arises where I am the one with surplus food. Thus, while mathematics favor reciprocal altruism, there are difficult psychological obstacles to implementing it via general mechanisms of learning and choice. This may explain the key role for trust in human cooperation (Knack & Keefer, 1997; McNamara, Stephens, Dall, & Houston, 2008; Mui, Mohtashemi, & Halberstadt, 2002; Silk, 2003; Zak & Knack, 2001). Possibly, trust functions as a specific behavioral adaptation that facilitates reciprocal exchanges of goods without requiring a general psychological mechanism to successfully associate prosocial acts with subsequent rewards.

The magnitude of the cognitive constraints on learned social behavior are difficult to overstate. For instance, experimental work in pigeons shows that even a short delay between choice and reinforcement can have an severe consequences for cooperation in a prisoner's dilemma (Baker & Rachlin, 2002). In this study, pigeons played an iterated prisoner's dilemma game against a computer that adopted the strategy tit-for-tat, cooperating one trial after the pigeon cooperated, and defecting one trial after the pigeon defected. The delay between trials was varied. When there was no enforced delay between trials — that is, when pigeons experienced defection from the computer immediately following their own act of defection — cooperation rates averaged 64%. But when reciprocated defection was delayed by just 18 seconds, cooperates rates dropped by a quarter, to 48%. This finding illustrates just how severe the constraint of rapid and salient response to antisocial behavior will be for punishment (or reward, defection, etc.) to promote social behavior by exploiting general processes of associative learning and behavioral choice in nonhuman animals (see also Stephens, McLinn, & Stevens, 2002).

In summary, the standard temporal structure of social interactions often allows punishment to follow rapidly after antisocial acts, but prevents contingent prosociality from following rapidly after punishment. This property makes possible a learned association between performing antisocial behaviors and receiving a punitive response, but makes more difficult a learned association between responding punitively and obtaining future benefits via prosociality. This affords an additional basis on which to predict that prosociality depends on rat-like learning mechanisms, while punishment depends on a goose-like specialized adaptation. Moreover, the same temporal constraints will often make punishment a more effective "teaching strategy" than reward. Physical aggression can often be employed for swift and salient reinforcement, whereas reward must often be delayed until an appropriate opportunity or resource is available.

5. The functional design of punishment

Human punishment furnishes more than its share of puzzles. Why do we execute prisoners who are soon to die of natural causes anyway? Why do we punish a malicious shooter who hits his target more than one who misses? Why do we excuse people for past crimes after a period of several years? Why is it illegal to push a child in a pond to drown, but perfectly legal not to throw a life preserver toward a drowning child?

Framed as moral, philosophical and legal puzzles, these questions have tickled and tortured scholarly minds for centuries. But they can also be framed as psychological puzzles, and in this capacity the arguments developed above offer insight. Two clear predictions follow from the claim that punishment is a specialized adaptation that exploits general learning processes in order to promote prosocial behavior among social partners. First, punishment should operate in a relatively inflexible manner; that is, more like the fixed action pattern of the goose than the learning behavior of the rat. Second, punitive behavior should be functionally designed in ways that reflect the particular constraints of general learning processes. These predictions allow us to understand puzzles of punitive behavior in terms of functional design.

5.1 Retribution

Philosophical and legal scholarship identifies several possible motives for costly punishment. These include deterrence (establishing a policy that discourages future harmful behavior), incapacitation (directly preventing future harmful behavior, e.g. by imprisonment or death), and retribution (harming morally responsible harmdoers for reasons of "just desert"). Notice that incapacitation and deterrence treat punishment as instrumentally valuable: it is a useful behavior because it maximizes the welfare of possible future victims. This kind motivational structure is compatible with punishment as product of general mechanisms of learning and behavioral choice, which also operate roughly by maximizing expected value. By contrast, retribution accords punishment itself primary value—retributive punishment occurs not because it is expected to bring secondary benefits, but rather because it is considered to be a necessary or deserved response. This motivational structure is more compatible with punishment as specialized behavioral response.

Several lines of psychological research suggest a basic process of assigning blame and punishment (Cushman, 2008; Darley & Shultz, 1990; Heider, 1958; Shaver, 1985; Shultz, Schleifer, & Altman, 1981; Weiner, 1995), and in its details it is fundamentally retributive. When a harm occurs, we begin by seeking out individuals who are causally responsible. We then assess the harm-doers' mental states at the time of their actions, determining whether they had a culpable mental state such as intent to harm or foresight. Finally, we assign punishment to the causally responsible parties in proportion both to the degree of the harm and the degree of their culpable mental state. On its face, this basic model of punishment described fits best with a retributive motive for punishment, as opposed to deterrence or incapacitation. It does not contain any explicit calculation of the probability of future transgression, as would be predicted if deterrence were the primary psychological motivation underlying punishment. Rather, it treats punishment itself as an object of primary value, as would be predicted if retribution were the primary psychological motivation underlying punishment.

Psychological studies have directly contrasted the predictions of incapacitation or deterrence as motivations for punishment against the predictions of a retributive motivation for punishment, consistently favoring the latter[6] (Carlsmith, 2006; Carlsmith, Darley, & Robinson, 2002; Darley, Carlsmith, & Robinson, 2000). Contrary to the predictions of an incapacitation or deterrence motivation, judgments of deserved punishment are not strongly modified by the probability that a perpetrator will re-offend or the probability that future offenses will be go undetected, two factors that should increase the amount of punishment assigned.

Additionally, several studies of actual punitive behavior in structured economic exchanges show that people punish harmful acts even in one-shot, anonymous interactions (Fehr & Gachter, 2002; Henrich, et al., 2005). This is a situation in which the punisher clearly has no personal stake in deterrence of future harms. It has been argued that the adaptive function of punishment in one-shot interactions is to deter future harms perpetrated against third parties (Fehr & Gachter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003). However, there is some evidence that people are more likely to punish in a one shot interaction if they have been harmed themselves than if a third party was the victim (Carpenter & Matthews, 2004). This casts doubt on the view that punishment is primarily motivated by a concern with future harms against third parties. Rather, the structure of punishment in one-shot interactions appears to be an inflexible, retributive response: You harmed me, so I harm you. Subjects' self-reported motivations match this conclusion: in one study of third party punishment, 14% of subjects said that they punished third parties in order to reduce the incidence of future harms (deterrence), 56% said they punished third parties in order to "get back" at those who acted antisocially (retribution), and 30% said they were motivated by both factors (Carpenter & Matthews, 2004).

Of course, retributive motivations might reliably produce deterrent or incapacitative effects. In fact, I have taken pains to argue that the best way to understand the functional value of punishment is precisely in terms of deterrence—i.e., eliciting contingent prosociality in future interactions. But the likely effects of punishment, and its adaptive function, need not constitute the psychological motivations that underlie it. Punishment may be adaptive for deterrent reasons at an "ultimate" adaptive level, and yet be instantiated by retributive mechanisms at a "proximate" psychological level. By analogy, the consumption of sugars and fats has future energetic benefits, and presumably the adaptive function of that consumption is to obtain the nutritive effects. The principle psychological motivation underlying the consumption of sugars and fats appears to be their taste, however, and not a learned association between consumption and future energetic states. Having an innate taste for sugar or fat circumvents the problem of learning by brute association which properties of potential foodstuffs are correlated with which future energetic states. Similarly, having an innate taste for punishment would circumvent the problem of learning associatively how to elicit future prosociality from social partners. Also, it would meet the inflexibility requirement introduced in Section 3: the requirement that punishers cannot be "taught out of punishment" by intransigent antisocial actors.

In summary, the standard psychological processes underlying individual punishment of harmful actions are best characterized by a retributive motivation, and not by reasoning about the long-term benefits of punishment. Retributive motives are typically triggered when a person performs an action that causes harm, and subsequent punishment depends both on the severity of the harmful outcome, and also the degree to which that harmful outcome was intended. Retributive behaviors are surprisingly inflexible, operating even in contexts where interactions are one-shot and anonymous. This psychological model of retribution matches the inflexibility requirement discussed above: Punitive behavior will tend to be maintained even against social partners that fail to adopt prosociality. As

[6] These studies target the punitive judgments of ordinary, non-expert respondents to psychological surveys. A separate but potentially related issue is the structure of the actual legal system, which in some instances is better described by retributive motives and at other times by deterrence or incapacitation motives.

we have seen, the apparent irrationality of this inflexible strategy actually has important consequences for the maintenance of co-dependent punishment and prosociality.

5.2 Punishing accidents

People often judge that some punishment is deserved for unintentionally harmful behaviors—that is, for accidents. Our sensitivity to accidental outcomes appears to be substantially greater for punitive judgment than for judgments of moral permissibility or wrongness (Cushman, 2008). This is surprising: Why is it that we tend to punish accidents to a greater degree that we actually consider them wrongful? Could the punishment of accidental outcomes reflects the function of exploiting general learning mechanisms to promote prosocial behavior?

Outcome-sensitive punishment has been observed in several vignette studies (Cushman, 2008; Darley, et al., 2000), and it is also widespread in the laws concerning negligent behavior[7]. More recently, it has been demonstrated in actual behavior using a probabilistic economic game (Cushman, Dreber, Wang, & Costa, 2009). In this "trembling hand" game, one player allocated money between herself and her partner either selfishly (everything for herself), fairly (an even split), or generously (everything for her partner). But, her allocation was subject to a probabilistic roll of a die — for instance, attempting to be selfish had a 4/6 chance of a selfish allocation, a 1/6 chance of a fair allocation, and a 1/6 chance of a generous allocation. Thus, the allocator could have selfish intent matched with a generous outcome (or vice versa). Finally, her partner was given the chance to respond by decreasing or increasing the allocator's payoff (i.e., punishing or rewarding the allocator). In doing so, the partner could respond to the allocator's intent, or the actual outcome of the allocation (even if unintended), or both. On average, responders punished both stingy intent and accidentally-stingy outcomes. If anything, they weighted accidental outcomes slightly more than intent.

It is surprising that people's judgments of deserved punishment are strongly influenced by accidental outcomes because their judgments of "moral wrongness" are not (Cushman, 2008). For instance, consider two drunk drivers, one who runs into a tree and another who runs into a person. People will tend to say that both behaved equally wrongly, but that the murder deserves substantially more punishment. This finding contradicts the commonsense assumption that we punish actions simply in proportion to their wrongfulness. To be sure, the moral status of an intention and an act play an important role in judgments of deserved punishment—but accidental outcomes count for a lot, too. The fact judgments of moral wrongness and punishment differ in this way may explain why the punishment of accidents has been a point of particular concern in law and philosophy (Hall, 1947; Hart & Honore, 1959; McLaughlin, 1925; Nagel, 1979; Williams, 1981).

So, why do we punish accidents? Possibly, because people learn from the punishment of accidents. To borrow a helpful term from education theory, accidents are "teachable moments". A may invade P's territory, eat P's food, consort with P's mate, etc., with no knowledge of P's claims. These transgressions are, in some sense, unintentional—the acts of walking, eating and mating are intentional, but their transgressive nature is unforeseen. Still, by punishing, P has the opportunity to teach A the boundaries of his territory, his property and his relationships. Consider an even more unintentional harm: A loses his grip of a log and drops it on P's foot. The harm is wholly unintentional, but punishment may teach A to exert greater care in future circumstances. It also teaches A what matters to ("my foot") and how much it matters ("as much as the whap you're about to get"). In these cases, A's unintentional harm provides an opportunity for P to teach a valuable lesson.

Of course, P's punishment will accomplish little if the type of behavior produced by A could not be successfully controlled even in future interactions. Indeed, experimental evidence suggests that the punishment of accidents is restricted to behaviors that could, in principle, be controlled (Alicke, 2000; Cushman, et al., 2009). This criterion is paralleled in Anglo-American law, as well (Kadish, Schulhofer, & Steiker, 2007). For instance, suppose a driver's brakes fail and he hits a pedestrian. Clearly the harm to the pedestrian is not intentional. However, the driver will tend to be exposed to greater liability if he failed to have his brakes maintained properly, and lesser liability if the flaw in the breaks was inherent to their original manufacture. In the former case, the brake failure was controllable by the driver; in the latter case, it was not. The factor of controllability also plays a key role in punitive behavior in the trembling hand game (Cushman, et al., 2009). When the allocator has some probabilistic control over the allocation by choosing one of three die, she is punished for accidental outcomes, as described above. But, when the allocator has no probabilistic control over the allocation—when the allocator is forced to roll a single die where stingy, fair or generous outcome are equally likely—she is

[7] In the Anglo-American tradition, when a person performs a negligent act, she assumes liability for the consequences of that negligent behavior. But, if no harm occurs, there is no liability. As the American judge Benjamin Cardozo wrote, "proof of negligence in the air, so to speak, will not do" ("Palsgraf v. Long Island Railroad Co.," 1928). If a harm is caused, then typically the extent of liability is increased in proportion to the degree of harm caused.

punished less, if at all, for accidental outcomes[8]. Again, a focus on controllability makes sense from the functional perspective of modifying social partners' future behavior. When A's behavior is controllable, P's punishment can effectively modify A's future behavior ("I know you didn't mean to be stingy this time, but I'm going to show you what will happen if you don't watch out"). When A's behavior is not controllable, P's punishment cannot modify A's future behavior, and so there is no value to teaching a lesson.

So, accidents may be punished because they are "teachable moments". But to work, they must teach a lesson that the transgressor is able learn. If I punishing you for swinging a hammer at a nail and hitting my thumb, will you learn not to hit my thumb (the lesson I am hoping for)? Or will you learn instead not to aim for nails (a lesson I have no reason to teach)? This depends on the structure learning mechanisms themselves. The key factor is whether the learning mechanism associates reward or punishment with the intended action, or instead with the outcome actually produced.

This is a fundamental distinction in reinforcement learning (Daw & Shohamy, 2008; Sutton & Barto, 1999). *Model-free* mechanisms of learning associate experienced reinforcement with the action selected (i.e., the agents intent). Thus, if a ball is thrown towards the plate and hits the batter, a model free mechanism reduces the value associated with "throwing the ball at the plate". By contrast, *model-based* mechanisms of learning associate experienced reinforcement with the outcome produced. These mechanisms will instead reduce the value of "hitting a batter", selecting future actions based on some model that relates possible actions (like throwing a ball) to outcomes (like hitting a batter). As the name suggests, model-based learning mechanisms require a working causal model of the world, one that relates actions to outcomes. For this reason they are relatively harder to implement than model-free learning mechanisms. But, they will tend to learn from the punishment of accidental outcomes more effectively.

This raises a key empirical issue: in social contexts, do people learn from punishment in a model-free way ("don't perform that action again!"), or in a model-based way ("don't cause that outcome again!")? This question has been assessed using a modified version of the trembling hand game implemented in a two-player game of darts (Cushman & Costa, in preparation). One player throws darts at a multi-colored board, and her shots will win or lose money for a second player. But the thrower doesn't know which colors will win money for her partner, and which colors will lose money. Her partner has the opportunity to teach the thrower, however, by rewarding or punishing the thrower after each throw. Critically, the thrower has to call her shots, and she receives a bonus from the experimenter every time she hits the color she calls. Thus, she has a clear incentive to be honest about what she is aiming for. The responder therefore knows both the thrower's intended target and, of course, what she actually hits.

Suppose that the thrower aims for a high-value target, but hits a low-value target. Should the responder reward the thrower (to encourage aiming at the high-value target) or punish the thrower (to discourage hitting the low-value target)? This question was assessed by using a confederate in the place of the responder who adopted either an intent-based policy of reward and punishment or an outcome-based policy of reward and punishment. The results showed that the thrower learned the relative value of the targets significantly better when the responder adopted an outcome-based punitive strategy, compared to an intent-based punitive strategy. Moreover, a model-based simulation of the thrower's learning process matched throwers' observed behavior substantially better than a model-free simulation.

Thus, evidence suggests that the actual structure of human learning processes in social situations makes the punishment of accidental outcomes advantageous. The experimental results presented here are only an initial foray into the complicated psychology underlying human learning in social contexts. Minimally, however, they suggest that when we exploit accidental outcomes as "teachable moments", our social partners are receptive pupils. Punishing accidents may contradict our moral attitude that "it's the thought that counts". But, as a matter of functional design, it is an effective way to leverage learning mechanisms in order to maximize future prosocial behavior.

5.3 Salience and the action/omission distinction

Some of the events, objects and properties we encounter tend to "pop out" and capture attention (Parkhurst, Law, & Niebur, 2002) and these salient stimuli are the easiest to learn about (Mackintosh, 1975; Rescorla & Wagner, 1965; Schwartz & Reisberg, 1991). Thus, in order for punishment to successfully discourage antisocial behavior via general learning processes, both the punishment and the antisocial act should be salient to the learner. The learner needs to notice that it is being punished and to infer what prompted that punishment. This constraint may explain the preference for punishing harmful actions versus harmful omissions.

[8] When the allocator has no intentional control over the allocation, the responder's behavior is still, on average, affected by the allocation amount. However, this pattern of behavior may be best understood as a product of inequity aversion (Fehr & Schmidt, 1999) rather than retributive punishment.

Consider a puppy with two antisocial traits: He wets the carpet (an action) and fails to pick up the newspaper (an omission). Imagine trying to induce prosocial behavior by punishing the puppy. Every time it wets the carpet you punish it, and every morning when it fails to bring in the newspaper you punish it. Intuitively, you might guess that the puppy will learn to stop wetting the carpet, but will never learn to fetch the newspaper. One way of putting this is that "wetting the carpet" is a more salient event to the dog than "not picking up the newspaper". Even if the puppy knew that it was being punished for not doing something, how would it know what that something is? After all, at any moment there is an infinity of actions we are *not* performing.

Both experimental evidence (Hineline & Rachlin, 1969; Hoffman & Fleshler, 1959) and formal modeling (Randløv & Alstrøm, 1998) bear out this intuition. Animals tend to respond to punishment by performing innate, stereotypic avoidance behaviors. It has been suggested that most "novel" behaviors successfully conditioned by contingent non-punishment are in fact close variants of innate avoidance behaviors, and that truly novel behaviors have only been obtained via punishment with extreme effort on the part of the experimenter (Bolles, 1970). On the other hand, schedules of contingent reward are more successful at conditioning novel behaviors. The way that they succeed, by employing processes of "shaping" and "chaining", is also revealing. In shaping, the experimenter begins by rewarding a behavior already in the animal's repertoire and then restricts reinforcement to variants of the behavior ever closer to the desired action. In chaining, the experimenter rewards the performance of several individual behaviors when performed sequentially. These techniques presumably work because the target the performance of salient actions for immediate reinforcement, narrowing the space of potential hypotheses that an organism must consider when associating its behavioral choices with pleasant or aversive outcomes.

Notably, the action/omission distinction is widely reflected in the law and intuitive moral judgment. Although a legal "duty of care" does mandate prosocial action (i.e., punishes an antisocial omission) in certain specialized cases such as a parent's obligation to provide active support for a child, criminal law is overwhelming focused on the punishment of harmful actions, and not harmful omissions (Kadish, et al., 2007). This distinction between actions and omissions is reflected in ordinary people's moral judgments as well (e.g. Cushman, Young, & Hauser, 2006; Spranca, Minsk, & Baron, 1991). Moreover, it appears that the action/omission distinction is particularly acute in judgments of deserved punishment, compared with judgments of moral wrongness (Cushman & Young, in press).

To restate the present proposal in very general terms: Actions constitute a more salient target for learning than do omissions. Thus, in order to discourage antisocial actions, punishment (of the performance of the antisocial action) is preferred over reward (of the absence of antisocial action). By contrast, in order to encourage prosocial actions, reward (of the performance of the prosocial action) is preferred over punishment (of the absence of the antisocial action). Section 3 argued that punishment has a relative temporal advantage over reward, however: Your fists are always available for immediate punishment, while opportunities for reward may be fewer and further between. Combining these proposals, basic learning constraints appear to make it easier for organisms to discourage each others' antisocial actions than to encourage each others' prosocial actions. Section 7.6 considers ways in which human cognition can move beyond these constraints.

5.4 Limited capacity

When individuals have severe impairments in their general mechanisms of learning or behavioral control, punishment cannot effectively leverage those general mechanisms to promote prosocial behavior. Consequently, one might expect retributive motivations to be lessened for perpetrators incapable of learning or behavioral control. There is some suggestive evidence from legal codes, which commonly differentiate between perpetrators with full versus diminished mental capacity (Kadish, et al., 2007). Psychological research also suggests that considerations of mental capacity affect people's judgments of deserved punishment (Darley, et al., 2000; Fincham & Roberts, 1985; Robinson & Darley, 1995). For instance, Darley, Carlsmith and Robinson (2000) found that subjects assigned substantially less punishment to a person who murdered as a consequence of hallucinations resulting from an inoperable brain tumor, compared to a person who murdered out of jealous rage. Subjects indicated that the tumor patient should be detained in a mental health facility, clearly evincing a non-retributive sensitivity to preventing future harm, but they did not report a motivation to see him punished with prison time.

In another study (Robinson & Darley, 1995), only 16% of subjects assigned punishment (versus civil commitment) in a case where a schizophrenic man killed an innocent bystander under the belief that the bystander was about to attack him. Darley and Robinson compare this case to another involving "mistaken identity" by a sane perpetrator. In this case, a shop-owner is robbed and chases the burglar. He misidentifies another man as the burglar during the chase, gets in a fight with that man, and kills him. In this case, 100% of subjects assigned punishment. Why

does the schizophrenic who mistakes an innocent as an assailant receive no punishment (but rather civil commitment), while the sane man who mistakes an innocent as an assailant receives substantial punishment? A simple analysis of the mental state of the perpetrator at the time of the crime is not sufficient to account for this discrepancy — both individuals murdered in a fight due to mistaken identity. A critical factor may be the greater capacity of sane versus schizophrenic individuals to learn from the experience of punishment and successfully regulate future behavior on the basis of that learning. Future research should explore, first, the extent to which an evaluation of diminished capacity is central to the human retributive instincts and, second, the specific categories of impairment that trigger such an assessment.

5.5 Immediacy

General learning processes form associations between behavior and reinforcement more efficiently when the delay between the two is short. Consequently, punishment should tend to follow as quickly as possible after the commission of a harmful act, and given a long enough delay the motivation for punishment should be extinguished. Notably, many legal systems impose a statute of limitations on the prosecution of a criminal act. For instance, in my home state of Massachusetts, standard criminal misdemeanors carry a statute of limitations of six years. However, this period is lengthened for some felonies and there is no statute of limitations on murder.

It is unclear whether the statute of limitations reflects an underlying feature of the psychology of punishment in ordinary people. Additionally, it should be noted that the statute of limitation on most criminal offenses is on the scale of years, whereas temporal constraints on general learning processes in non-human animals often apply on the scale of minutes (Renner, 1964). A key direction for future research is to test whether something analogous to a "statute of limitations" is a fundamental feature of human retributive psychology, and whether it plausibly reflects the temporal constraints imposed by domain general learning processes in non-human animals.

5.6 Conclusions

I have argued that several features of human punishment—retributive motives, the punishment of accidents, the preference for actions over omissions, the limited capacity excuse, and the statute of limitations—may reflect its functional design. Specifically, each of these features can be sensibly interpreted as elements of specialized mechanism that uses punishment to induce prosocial behavior among social partners by exploiting their general mechanisms of learning and behavioral control. Like any attempt to understand complex behavior in adaptive terms this proposal is speculative. Moreover, a general argument for functional design cannot, by itself, distinguish between the influences of biological adaptation, cultural adaptation, or human reasoning. Nevertheless, it shows how otherwise puzzling and disparate features of human social behavior can begin to cohere into a more sensible and unified schema by considering a simple question: What is this behavior designed to do?

6. Specialized mechanisms of prosocial behavior

If punishment matches the constraints of learned prosocial behavior, then it certainly must be the case that prosocial behavior is learned. But, is it? Or, alternatively, is prosocial behavior accomplished by specialized behavioral mechanisms more like the goose's egg retrieval than the rat's lever-press? Several lines of evidence are suggestive of innate mechanisms supporting prosociality in humans. However, there remains substantial scope for learned prosociality to have shaped the functional design of punishment.

The case for innate prosociality begins with adaptive considerations. The widespread existence of punishment and reward in human societies imposes a selective pressure to rapidly adopt prosocial behavior when it is enforced. As noted in Section 3, the Baldwin effect describes a tendency for general learning mechanisms to pave the way for specific adaptations. Along these lines, some form of domain-specific innate preparation to adopt prosocial behavior might be favored over pure reliance on general mechanisms of learning and behavioral choice. An initial attempt to test this claim in an agent-based simulation model shows that a moderate bias towards prosocial behavior is favored in an evolving population where the punishment of antisocial behavior dominates (Cushman & Macendoe, 2009).

Empirical evidence is more compelling than adaptive theory, and here too an innate preparation for prosocial behavior is suggested. Just as economists talk about a "taste for retribution", they have identified tastes for generosity, fairness or cooperation (Gintis, et al., 2005). Across diverse experimental paradigms, humans choose behaviors that provide benefits for others at a cost to themselves, without the motivation of reciprocal reward or punishment (Batson & Shaw, 1991; Henrich, et al., 2005). Moreover, prosocial behavior appears to be developmentally early-emerging. Human infants and some apes spontaneously engage in prosocial behavior, for instance picking up a pen that a stranger has dropped out of reach and returning it to him (Warneken & Tomasello, 2006).

At the same time, there is also compelling evidence that prosocial behavior has a substantial learned component. To begin with, there is substantial cross-

cultural variation in prosocial behavior as measured by standard behavioral-economic paradigms (Henrich, et al., 2005; Herrmann, Thöni & Gächter, 2008). Individuals determine their levels of prosocial behavior in part by assessing the behavior of peers (Frey & Meier, 2004). Prosocality is also acquired developmentally through experiences that direct attention to others' feelings and activate empathy (Hoffman, 2000). There is also evidence specifically for a role of punishment in learned prosociality. Research using economic games also shows that levels of prosocial behavior are sensitive to rates of punishment (Fehr & Gächter, 2002). Additional evidence comes from studies of psychopaths, who unquestionably lack the taste for generosity and fairness. Notably, psychopaths do not show an exclusive deficit in prosocial emotions, but rather a more general deficit in processing negative feedback (i.e. punishment) and integrating it into future behavioral plans (Blair, 2003).

Two broad conclusions are warranted. First, there are both innate and learned contributions to human prosociality. Second, human adults maintain flexibility in their prosocial behavior, adjusting levels of prosociality according to cultural norms and past experience of punishment and reward. So, where does this leave the argument that punishment should be adapted to match the constraints of general learning processes? One possible consideration is the origin of punishment: Although innate preparations for prosocial behavior presently exist, Section 3 argued that punitive strategies probably originated by exploiting general learning mechanisms. Perhaps the current structure of punitive instincts still reflect the original functional design. On the other hand, perhaps not. This argument has little appeal because it depends on unverifiable speculation about evolutionary stasis.

A second argument depends not on distant origins of punishment, but rather on its present scope. Let's begin with the strong hypothesis there is a fully-formed, innate "taste for generosity": People are born valuing prosocial behaviors and devaluing antisocial behaviors. For instance, imagine that sharing food feels intrinsically good, stealing food feels intrinsically bad. Insofar as these innate mechanisms cause people to share and not to steal, punishment will be unnecessary. But, in some circumstances, the importance of food may outweigh the intrinsic disutility of theft for an agent. In simple terms, hunger will sometimes hurt more than shame. When an agent engages in theft despite its intrinsic disutility, punishment then plays a critical role by assigning an additional source of disutility to theft: the disutility of the punishment. In the future, the agent must weigh its hunger not only against its own intrinsic guilt, but also the prospect of extrinsic retaliation. In order for punishment to be effective in this manner, of course, the agent must to associate the performance of antisocial actions with future punishment. The critical point is that punishment is required only when the intrinsic (possibly innate) value of prosociality is insufficiently motivating. Thus, however much prosocial behaviors are valued via innate mechanisms, punishment may still be required to exploit general processes of learning and behavioral control to obtain prosocial behavior. It is precisely when altruism fails that punishment must work[9].

Adopting somewhat weaker hypothesis, prosociality might depend on specialized learning mechanisms, rather than an innate valuation of prosociality. Thus, for instance, the human mind may be finely tuned to detect punishment that follows antisocial action, even when delayed or unobvious. Analogous mechanisms certainly exist outside the social domain; for instance, rats are innately prepared to reliably associate between novel tastes and subsequent illness (Garcia, Ervin, & Koelling, 1966; Garcia & Koelling, 1996). Within the social domain, animals use specialized behavioral routines to communicate violent threats without paying the costs of engaging in actual violent behavior (Maynard Smith & Price, 1973). In contexts where organisms deploy specialized learning mechanisms the functional design of punishment should reflect the specific constraints of those mechanisms, rather than the constraints of general mechanisms of learning and behavioral choice.

Yet, once again, where innate preparations end general learning mechanisms must suffice. Prosocial behaviors that fall outside of the scope of innate preparation must be supported by general mechanisms of learning and choice. And, when specialized learning mechanisms fail to sufficiently motivate prosocial behaviors, general mechanisms of learning can still be leveraged to provide additional motivation. These might be termed arguments from "scope": However large the scope of innate preparation for prosociality, it can be expanded via dependence on general mechanisms of learning and choice.

The argument from scope is particularly important when considering human social behavior, which is exhibits broad flexibility across diverse contexts that could not have been anticipated on the timescale of biological evolution. I have argued at length that "rat-like" general learning mechanisms are highly constrained, and that aspects of human retributive instincts match those constraints. In the following section, however, I will argue that human's general learning mechanisms are vastly less constrained, and this explains much about the unique complexity and successes of human social behavior.

[9] Similarly, there may be variation between individuals in levels of intrinsic prosociality, in which case punishment would play a key role in those individuals who are only minimally motivated by intrinsic prosocial concerns.

7. Cognition and social behavior in humans

Humans possess profoundly more sophisticated cognitive abilities than non-human animals, and prosociality is far more widespread and flexible in humans than in non-human animals. There is broad agreement that this no coincidence. On the one hand, it has often been argued that the existing demands of a complex social life may have provided a key selective pressure towards the development of more powerful general cognitive abilities (e.g. Byrne & Whiten, 1989; Trivers, 1971). On the other hand, it has been argued that prior development of powerful cognitive abilities allowed complex social systems to emerge (Stevens, Cushman, & Hauser, 2005), including especially the cognitive mechanisms that support cultural transmission (Richerson, Boyd, & Henrich, 2003). To claim a single direction of causation probably misstates a fundamentally coevolutionary relationship; in any event, complex social life demands powerful cognition. The relationship between punishment and prosociality that we have considered above—and, critically, the distinction between specialized cognitive mechanisms versus general learning and reasoning mechanisms—helps to illuminate why. As we will see, general processes of learning and reasoning are critically important to the richness of human social life, first, because they expand our capacity to learn from punishment and reward and, second, because they expand our capacity to identify acts warranting punishment and reward.

Consider an illustrative example. As I have argued above, the general learning ability of dogs dooms the strategy of training your dog to fetch the newspaper by withholding doggie treats each Christmas if it fails to do so. The temporal delay between the dog's behavior and the reinforcement, combined with the minimal salience to the dog of "not fetching the newspaper" and "not receiving a doggie treat on Christmas", make it highly unlikely that the dog will form the necessary learned association. However, a similar strategy might be far more effective in training your son to fetch the newspaper. You can explain what you want him to do, and the consequences of failure. He can rapidly comprehend this connection, has an available conceptual structure that relates prosocial action to reciprocal holiday rewards, and might be sufficiently motivated by that distant reward to modify his present behavior. Your son's general capacity to (1) acquire information via language, (2) rapidly integrate new information into rich conceptual models of the world, (3) and use that conceptual knowledge to guide behavioral planning and choice allows him to respond to social punishments and rewards far more flexibly than your dog.

7.1 Language

Language has a transformational impact on human learning. Without language, knowledge about the world will typically only be obtained via direct observation of or interaction with the with the relevant phenomenon; with language, the experience of one individual can ultimately support knowledge among others (Tomasello, 1999). Thus, for instance, I know a great deal about Rwanda, retirement and ribosomes despite very little direct interaction with each. More particularly, language plays a key role in acquiring conceptual abstractions. It allows us to generalize from the three stooges, the three kings and the three tenors to the conceptual abstraction "three", embedded within broader concepts of counting and numerosity (Carey, 2004). We can generalize from moving objects to "velocity", from unsavory characters to "psychopath", from missed chances to "opportunity cost".

Conceptual abstraction may be possible without language, but linguistic symbols learned from social partners create a cognitive placeholder in the mind. They focus attention on fruitful generalizations and allowing the learner to gradually fill the empty conceptual structure with rich, productive content. Like the grain of sand that starts a pearl, linguistic symbols provide a nucleus around which concepts can grow. Moreover, formal models show that the kinds of conceptual abstractions supported by language cascade downwards to support learning at lower, more concrete levels, as well (Goodman, Ullman, & Tenenbaum, in press). Thus, for instance, learning conceptual abstractions such as "belief" or "cause" can support the acquisition of knowledge about particular beliefs and particular causal relationships.

These consequences of language greatly enhance the potential of punishment (and reward) to elicit prosocial behavior via general learning processes. First, language allows threats of punishment and promises of reward to be communicated in advance of the relevant behavior. Absent language, threats and promises can only be inferred by the experience or observation of past instances of punishment and reward for sufficiently similar behaviors. Second, language allows a behavior to be readily associated with punishment or reward at a long temporal delay. Absent language, behavior will typically only be associated with rewards and punishments when they follow immediately. Third, language allows for the rapid communication of novel and complex behavioral expectations that do not already exist in the behavioral repertoire of social partners. That is, language provides a rapid solution to the problem of "shaping" new behaviors, described in Section 5.3. I have already used the example of "bring me the newspaper" as a demand that is easy to communicate by language and relatively harder to communicate without language. Demands like "bring a

ten-percent tithe", "bring the murderer dead or alive" and "bring back these tools by October" fall into the same category, and precisely these kinds of demands are central to the complexity of human social life.

7.2 Conceptual models

Of course, the power of language to communicate is constrained by the power of language-users to comprehend. Here, again, humans have fundamentally different mental resources available compared with non-human animals, certainly in magnitude and possibly in kind. Human thought is supported by rich mental models that make use of conceptual abstractions and can be productively combined (Carey, 2009; Fodor, 1975; Roberts & Mazmanian, 1988; Sloman, 1996; Thompson & Oden, 2000). Three particular conceptual competencies are likely to have a large impact on prosocial behavior. The first is our understanding of others' mental states—their perceptions, sensations, goals and beliefs—which allows us to rapidly and reliably infer what social partners want from us (Saxe, Carey, & Kanwisher, 2004; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Warneken & Tomasello, 2006). The second is our ability to construct complex causal theories relating spatially and temporally distant events (Carey, 2009; Murphy & Medin, 1985; Waldmann & Holyoak, 1992), which allows us to predict the likely consequences of our behavioral choices on social partners' welfare. The third is our ability to construct appropriate analogies between situations, and to infer abstract principles on the basis of those analogical constructions (Gentner, Holyoak, & Kokinov, 2001). Combining these three competencies, humans have the capacity to infer from specific experiences of punitive behavior ("When I steal apples from Billy, he punches me") to a general model of punitive behavior ("When my behavior interferes with others' goals, they punish me")[10]. Conversely, we have the ability to appreciate how a linguistically communicated rule stated in abstract terms ("Do onto others as you would have others do onto you") applies to particular circumstances ("Don't steal apples from Billy").

Each of these three aspects of humans' conceptual knowledge allows us to infer the appropriate course of action without direct experience of past punishment for a particular behavior. A conceptual abstraction like, "Help others achieve their goals – eventually they will do the same for you" depends on mental state inference, the association of temporally distant events, and abstraction across diverse cases. Critically, it can effectively guide behavior in novel, unfamiliar circumstances. Without such a conceptual abstraction, an organism must wait for specific feedback for each category of action it can perform in order to learn optimal patterns of social behavior.

7.3 Planning and choice

Finally, humans have a greatly enhanced capacity to use complex conceptual knowledge to guide behavioral planning and choice. Humans are much more flexible in means-end reasoning than non-human animals, accomplishing large goals by constructing a hierarchy of smaller sub-goals through planning (Badre & D'Esposito, 2009; Conway & Christiansen, 2010; Koechlin & Jubault, 2006). This expands the range of prosocial actions that one person can undertake on behalf of another—not just sharing food, but sharing a plough, so to speak.

Additionally, humans have far greater ability to inhibit impulsive or habitual responses in order to maximize value in the distant future (Rachlin, Raineri, & Cross, 1991). In economic terms, humans have a very shallow rate of temporal discounting: a dollar tomorrow is deemed nearly as valuable as a dollar today. Non-human animals discount future rewards several orders of magnitudes more steeply, often devaluing rewards by more than half within a single minute (Mazur, 1987; Richards, Mitchell, de Wit, & Seiden, 1997; Stevens, Hallinan, & Hauser, 2005). Temporal discounting has important consequences on the stability of prosocial behavior (Stephens, et al., 2002). Even if your dog could understand that fetching the newspaper in May is linked to rewards at Christmas, it would probably not experience those distant rewards as sufficiently motivating; by contrast, your son is likely to weight the prospect of future reward much more heavily. Among humans, the punishments and rewards for social behavior typically occur at a long delay. Our ability to experience the motivational force of delayed reinforcement, and to incorporate it into complex behavioral plans, is critical to the functioning of human social life.

7.4 Cognition and the punisher

So far, I have focused on the way that powerful cognitive mechanisms expand the ability of humans to learn from punishment. At the same time, they can also expand the circumstances in which humans choose to punish. The retributive impulse to punish people who cause you harm is limited by the capacity to identify the relevant causal relationship. The learning and reasoning mechanisms possessed by non-human animals will typically limit inferences of causal responsibility for harm to direct observation. In humans, the assignment of causal responsibility for harm can extend across

[10] For an initial attempt to model the role of individual- versus group-level inferences about social behavior in the context of punishment and prosociality, see Cushman & Macendoe (2009).

miles, years, and long causal chains. For example, consider many Americans' urge for retribution against Osama bin Laden for the September 11th attacks. Thus, the simple rule "punish those who harm you" inherits the tremendously sophisticated ability to assign causal responsibility for harm, affording a powerful motivation to kill a man half a world away whose role in causing the harm was decisive, to be sure, but also very indirect. So, just as uniquely human cognitive abilities afford much greater ability to learn from punishment, they afford much greater ability to assign blame. Our 'taste for retribution' can be a specialized behavioral adaptation and still be profoundly enhanced by very general improvements in cognitive ability.

7.5 Domain generality and the diversity of human social behavior

Each of the three capacities I have considered—linguistic communication, conceptual abstraction and controlled behavioral choice—functions in diverse domains of human thought and behavior. There is no sense in which they are limited to the specific problem of supporting punitive or prosocial behavior. It may be that social demands provided a key selective pressure favoring the emergence of these domain-general capacities (Byrne & Whiten, 1989). If so, they could be regarded as adaptations to social life. But they are clearly not specialized mechanisms in the sense of the fixed action pattern of the goose; rather, their power lies specifically in their flexibility, like the general learning processes of the rat (yet much more powerful still).

Past discussions of cognitive contributions to human prosocial behavior have not emphasized the importance of flexible, domain general mechanisms as opposed to narrowly deployed specialized mechanisms. It has been asserted that the specific problems of punishment, reward and prosociality require cognitive capacities like individual recognition (Trivers, 1971), memory (McElreath, et al., 2003; Stevens, et al., 2005; Trivers, 1971), quantitative representations of value (McElreath, et al., 2003; Silk, 2003; Stevens, et al., 2005), the capacity to track reputation (McElreath, et al., 2003; Mohtashemi & Mui, 2003; Nowak, 2006), and motivations for reciprocity (Axelrod, 1984; Fehr & Henrich, 2003; Trivers, 1971), retributive punishment (Fehr & Gachter, 2002), and the valuation future rewards (Stephens, et al., 2002; Stevens, et al., 2005). Surely, these capacities are critical. Still, what seems to make humans unique is that they can be flexibly engaged across arbitrarily diverse circumstances.

To see why domain generality is so important in the human case, consider a foil: the specialized cognitive mechanisms that support food caching in the scrub jay (Clayton & Dickinson, 1998; Emery & Clayton, 2001). These birds cache thousands of food items in different locations and retrieve them months later, solving a challenging memory task. They retrieve the resources according to their relative nutritional content, and are sensitive to variable rates of decay between food types, solving a challenging valuation problem (Clayton & Dickinson, 1998). They control the impulse to consume the resources immediately and instead cache them for use months in the future, solving a challenging inter-temporal choice problem. They even exhibit sensitivity to the perceptual access of other birds to their hidden caches (Emery & Clayton, 2001), solving a challenging behavioral prediction problem. The problems that food-caching birds solve are much like the problems inherent to social interaction. But, these birds' psychological solutions appear to be specialized to the task of food caching—more like the goose, less like the rat. As far as we know, these specialized mechanisms don't have much of an impact at all on prosocial behavior. In fact, they don't even seem to impact these birds' foraging behaviors beyond the narrow domain of food caching. At the risk of stating the obvious, scrub jays have not learned to plant seeds in order to harvest them later, or to arrange food in ways that promote the growth of nutritionally valuable insects, or (returning to the social domain) to trade food between each other. Yet, humans do all of these things. It is precisely because our reasoning and learning mechanisms are not specialized—because of their extraordinary flexibility—that they so transformational.

7.6 Reward and prosociality

The same powerful and flexible cognitive mechanisms that help people learn from punishment can also help them learn from rewards. As I argued above, the task of using reward to elicit prosociality via general learning processes is particularly difficult because of the basic temporal structure of social interactions. Fists and teeth are always available for immediate punishment; by contrast, opportunities to aid another, or to share resources, are not always available for immediate reward. This may explain why punishment is relatively common among non-human animals (Clutton-Brock & Parker, 1995) whereas evidence for reciprocal reward has been stubbornly difficult to establish (Hammerstein, 2003). But human capacities for linguistic communication, conceptual abstraction and controlled behavioral choice go a long way towards alleviating the constraints of general learning processes. We are able to learn about delayed rewards, comprehend how they are contingent on our own behavior, and motivate our behavior according to them.[11]

[11] Money also serves to eliminate the temporal delay between action and reward, replacing trust for token. There is no

In Section 5.3 I argued that the constraint of salience favors rewarding and punishing actions rather than omissions. This suggests a match between the punishment of harmful actions (versus the reward of omitting harm), and the reward of helpful actions (versus the punishment of omitting help). Combining this observation with the argument that punishment has key advantages over reward as a "teaching" device given cognitive constraints on learning in non-human animals, we can predict that non-human animals will generally avoid harmful actions towards each other (learned from punishment), but will not generally seek out helpful actions towards unrelated others (because of the difficulty of learning from reward). The human ability to leverage language, abstract thought and long-term planning and choice to make strategies of contingent reward a viable strategy thus stands to expand the general boundaries of prosocial behavior from "do not harm" to "provide help"—from a social world of libertarian individualism to a world of collective action. This argument depends on a number of steps and warrants a healthy skepticism. Nevertheless, its implications are substantial. What people will accomplish by aim to help each other vastly exceeds what they will accomplish by simply aiming not to harm each other. If this captures a rough distinction between human and non-human social behavior, then the unique flourishing of human social life can readily be understood.

7.7 Conclusions

I have reviewed three related aspects of human cognition that are radically different from their non-human counterparts: the capacity for linguistic communication, the capacity for reasoning about complex causal relations involving conceptual abstractions, and the capacity for controlled behavioral planning and choice. Collectively, these capacities greatly expand the capacity of humans to learn contingencies between the social consequences of their behavior and the contingent rewards and punishments of social partners. They also greatly expand the contexts in which humans can comprehend the impact of others' behavior on their own wellbeing, potentially expanding the range of circumstances that invoke retributive (or rewarding) motivations. In short, complex and powerful cognition can explain the complex and productive social life of humans. But, the relevant cognitive capacities are not specific to the social domain, much less to morality alone. To the contrary, the very feature of human cognition that explains its transformative role in social life is its domain generality.

obvious analog in the domain of punishment — i.e., a system of symbolic, immediate exchanges of sanctions.

8. Conclusion: The irony of punishment

There is an apparent tension in my argument. On the one hand, human punishment apparently matches some constraints of "general learning processes" possessed by our non-human ancestors. Put simply, our punitive instincts treat people like rats. On the other hand, the stunning complexity of human social behavior derives from new and powerful domain-general processes of learning and reasoning. That is, the foundation of human prosociality and cooperation is our ability to learn very differently, and much more effectively, than rats. Here is the tension: if it is so important to human social life that we learn much better than other animals, why would aspects of our punitive instincts be designed as if we learned just like other animals? Shouldn't punishment be tailored to the constraints—and the possibilities—of human learning?

There is a tension here, but not a logical contradiction. It is possible that some aspects of the psychology of punishment are relics of an earlier social environment, better suited to the learning mechanisms of our pets than our peers. Perhaps they work well enough in modern human life, just not quite optimally. Our taste for sugar and fat are often discussed in similar terms: unquestionably adaptive for our ancestors, but tuned sub-optimally to our present circumstances. At least two aspects of punishment might be similar.

One is the punishment of accidental outcomes. I reviewed several different studies demonstrating that judgments of deserved punishment are strongly affected by the degree of harm caused. Let me add one more favorite example to the mix: the legal penalties associated with drunk driving. Here in Massachusetts, if a drunk driver falls asleep at the wheel, hits a tree, and gets picked up by police, he can expect a fine of several hundred dollars. He might also be forced to enroll in an outpatient treatment program, or even have his license suspended for several months. But if he falls asleep at the wheel, hits a pedestrian, and kills her, he will receive between 2½ and 15 years in prison. These are radically different punishments for exactly the same behavior.

I argued that we punish harmful outcomes, even when accidental, because it is the most effective way to teach social partners what we want them to avoid and how much we want them to avoid it. In an experiment, dart-throwers learned the value of several targets better when rewarded and punished on the basis of outcomes, and worse when rewarded and punished on the basis of intent. This darts game was designed to reflect a time in our evolutionary history when the only way to communicate the value of others' behavior was to reward and punish. But, of course, humans do have language, and the ability to infer what social partners value much better than our nearest primate relatives.

Every drunk driver knows the value of a pedestrian life, though he may disregard it. We do not need to punish drunk drivers who kill in order to teach them that we value of others' lives; we need to punish them to disincentivize future drunk driving by themselves and others. That disincentive could operate as strongly if we punished all acts of drunk driving moderately, rather than punishing cases that cause no harm minimally and cases that cause death maximally.

To be sure, there are instances where intentions simply cannot be known and outcomes are the most reliable proxy. But, just as surely, there are cases where our knowledge of intent is quite reliable, and yet we still have a retributive impulse to grade punishment according to the degree of harm caused. Understanding the origins of that retributive impulse may help us to decide whether, on reflection, to endorse it. If I am right that its origins trace to a social environment very unlike today's, then a skeptical eye is warranted.

A second aspect of punishment that may be poorly adapted to our present situation is, quite simply, punishment itself. Recall that recent experimental and theoretical results that illustrate striking benefits at the individual and group level when prosocial behavior is enforced via reciprocated prosociality, rather than the threat of punishment (Dreber, et al., 2008; Rand, Dreber, et al., 2009; Rand, Ohtsuki, et al., 2009). In these studies, the enforcement of prosociality by punishment tended to devolve into cycles of costly retribution, while reciprocal prosociality tends to evolve into cycles of productive cooperation. Moreover, developmental research suggests that focusing childrens' empathy on the suffering of victims is a far more successful method of promoting prosocial behavior than punishing their transgressions (Hoffman, 2000).

Yet still, we punish—why? I have suggested that punishment was adaptively preferable for much of our evolutionary history because it afforded more immediate and salient responses to antisocial acts than the withholding of reward. This property was critical when social partners used general learning processes to adopt prosociality in the face of punishment, and when those general learning processes were highly constrained. But, humans have the capacity to communicate and comprehend the contingency between behavior and delayed reward, along with the capacity to be appropriately motivated by its prospect. Consequently, in some circumstances, punishment itself may be an outmoded and very costly impulse, compared with the possibilities of reciprocation and reward.[12]

As we have seen, there is an instructive comparison between our "taste for punishment" and our taste for sugar and fat. In both cases, these motivations circumvent the problem of learning a more general associative relationship. We do not need to learn which foods are associated with future energetic states in order to be motivated to consume; likewise, we do not need to learn that the punishment of antisocial behavior can promote prosociality in order to be motivated to act retributively. Moreover, the relative inflexibility of an innate retributive motivation avoids the unsustainable evolutionary dynamic of punishing only those who can, or do, learn from punishment.

But our taste for sugar and fat come with a definite cost in the modern world, where high-calorie foods are more widely available than our evolved tastes anticipated. Much the same is true of our taste for retribution. Cognitive abilities unique to humans that enable strategies of punishment and reward to support vastly more complex and productive forms of prosocial behavior than in non-human animals. Still, in some respects, the structure of our retributive taste—and perhaps even the taste itself—is adapted to those very constraints on learning that the human mind brilliantly exceeds.

References

Alicke, M. (2000). "Culpable control and the psychology of blame." *Psychological Bulletin*, 126(4): 556-574.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Badre, D., & D'Esposito, M. (2009). "Is the rostro-caudal axis of the frontal lobe hierarchical?" *Nature Reviews Neuroscience* 10: 659-669.

Baker, F., & Rachlin, H. (2002). "Self-control by pigeons in the prisoner's dilemma." *Psychonomic Bulletin & Review* 9(3): 482-488.

Batson, C. D., & Shaw, L. L. (1991). "Evidence for altruism: Toward a pluralism of prosocial motives." *Psychological Inquiry* 2(2): 107-122.

Blair, J. (2003). "Neurobiological basis of psychopathy." *British Journal of Psychiatry* 182: 5-7.

Bolles, R. (1970). "Species-specific defense reactions and avoidance learning." *Psychological Review* 77(1): 32-48.

Boyd, R., & Richerson, P. J. (1992). "Punishment allows the evolution of cooperation (or anything else) in sizeable groups." *Ethology and Sociobiology* 113, 171-195.

Byrne, R., & Whiten, A. (1989). *Machiavellian Intelligence*. Oxford University Press.

[12] On the other hand, there are surely cases where punishment is required. Consider a thief: what kind of incentive are your rewards to him? There is nothing you can give him that he can't steal. To stop his behavior may simply require punishment.

Carey, S. (2004). "Bootstrapping & the origin of concepts." *Daedalus* 133(1), 59-68.

Carey, S. (2009). *The Origins of Concepts*. Cambridge: MIT press.

Carlsmith, K. (2006). "The roles of retribution and utility in determining punishment." *Journal of Experimental Social Psychology* 42(4): 437-451.

Carlsmith, K., Darley, J., & Robinson, P. (2002). "Why do we punish? Deterrence and just deserts as motives for punishment." *Journal of Personality and Social Psychology* 83(2): 284-299.

Carpenter, J. P., & Matthews, P. H. (2004). "Social reciprocity." Institute for the Study of Labor.

Clayton, N. S., & Dickinson, A. (1998). "Episodic-like memory during cache recovery by scrub jays." *Nature* 395: 272-274.

Clutton-Brock, T. H., & Parker, G. A. (1995). "Punishment in animal societies." *Nature*, 373: 209-216.

Conway, C. M., & Christiansen, M. H. (2010). "Sequential learning in non-human primates." *Trends in Cognitive Science* 5(12): 539-546.

Cushman, F. A. (2008). "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment." *Cognition* 108(2): 353-380.

Cushman, F. A., & Costa, J. (in preparation). "Why do we punish accidents? An experimental investigation."

Cushman, F. A., Dreber, A., Wang, Y., & Costa, J. (2009). "Accidental outcomes guide punishment in a 'trembling hand' game." *PLOS One* 4(8): e6699.doi:6610.1371/journal.pone.0006699.

Cushman, F. A., & Macendoe, O. (2009). "The coevolution of punishment and prosociality among learning agents." *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). "The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm." *Psychological Science* 17(12): 1082-1089.

Cushman, F. A., & Young, L. (in press). "Patterns of moral judgment derive from non-moral psychological representations." *Cognitive Science*

Darley, J., Carlsmith, K., & Robinson, P. (2000). "Incapacitation and just deserts as motives for punishment." *Law and Human Behavior* 24(6): 659-683.

Darley, J. M., & Shultz, T. R. (1990). "Moral Rules - Their Content And Acquisition." *Annual Review of Psychology* 41: 525-556.

Daw, N., & Doya, K. (2006). "The computational neurobiology of learning and reward." *Current Opinion in Neurobiology* 16(2): 199-204.

Daw, N., & Shohamy, D. (2008). "The cognitive neuroscience of motivation and learning." *Social Cognition* 26(5): 593-620.

Dreber, A., Rand, D., Fudenberg, D., & Nowak, M. (2008). "Winners don't punish." *Nature* 452(7185): 348.

Emery, N. J., & Clayton, N. S. (2001). "Effects of experience and social context on prospective caching strategies by scrub jays." *Nature* 414: 443-446.

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

Fehr, E., & Henrich, J. (2003). "Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism." In P. Hammerstein (Ed.), *The Genetic and Cultural Evolution of Cooperation*. Cambridge: MIT Press.

Fehr, E., & Schmidt, K. (1999). "A theory of fairness, competition, and cooperation." *The Quarterly Journal of Economics* 114(3): 817-868.

Fincham, F. D., & Roberts, C. (1985). "Intervening Causation And The Mitigation Of Responsibility For Harm Doing." *Journal of Experimental Social Psychology* 21(2): 178-194.

Fodor, J. A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Frank, R. H. (1988). *Passion Within Reason: The Strategic Role of the Emotions*. New York: Norton.

Frey, B. S. & Meier, B. (2004). "Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment." *American Economic Review*, 94(5): 1717-1722.

Garcia, J., Ervin, F., & Koelling, R. (1966). "Learning with prolonged delay of reinforcement." *Psychonomic Science* 5(3): 121-122.

Garcia, J., & Koelling, R. (1996). "Relation of cue to consequence in avoidance learning." In *Foundations of animal behavior: classic papers with commentaries*, 4, 374.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). (2001). *The analogical mind*. Cambridge: MIT Press.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). "Explaining altruistic behavior in humans." *Evolution and Human Behavior*, 24, 153-172.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2005). "Moral sentiments and material interests: Origins, Evidence, and Consequences". In Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (Eds.) *Moral Sentiments and Material Interests*, Cambridge, MA: MIT Press.

Goodman, N., Ullman, T., & Tenenbaum, J. (in prep) "Learning a Theory of Causality."

Gutnisky, D., & Zanutto, B. (2004). "Cooperation in the iterated prisoner's dilemma is learned by operant conditioning mechanisms." *Artificial Life* 10(4): 433-461.

Haley, K. & Fessler, D. (2005). "Nobody's watching? Subtle cues affect generosity in an anonymous

economic game." *Evolution and Human Behavior* 26: 245-256.

Hall, J. (1947). *General Principles of Criminal Law*. Indianapolis: Bobbs-Merrill Company.

Hammerstein, P. (2003). "Why is reciprocity so rare in social animals? A protestant appeal." In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*. (pp. 83-94). Cambridge: MIT Press.

Hart, H. L. A., & Honore, T. (1959). *Causation in the law*. Oxford: Clarendon Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York,: Wiley.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic Man' in cross-cultural perspective: behavioral experiments in 15 small scale societies." *Behavioral and Brain Research* 28: 795-855.

Herrmann, B., Thöni, C. & Gächter, S. (2008). "Antisocial punishment across societies." *Science* 319(5868): 1362.

Hineline, P. N., & Rachlin, H. (1969). "Escape and avoidance of shock by pigeons pecking a key." *Journal of the experimental analysis of behavior* 12: 533-538.

Hirschfeld, L. A., & Gelman, S. A. (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Hoffman, H., & Fleshler, M. (1959). "Aversive control with the pigeon." *Journal of the experimental analysis of behavior* 2(3): 213.

Hoffman, M. (2000). *Empathy and Moral Development*. New York: Cambridge University Press.

Kadish, S. H., Schulhofer, S. J., & Steiker, C. S. (2007). *Criminal law and its processes* (8 ed.). New York: Aspen Publishers.

Knack, S., & Keefer, P. (1997). "Does Social Capital Have An Economic Payoff? A Cross-Country Investigation." *Quarterly Journal of Economics* 112(4): 1251-1288.

Koechlin, E., & Jubault, T. (2006). "Broca's area and the hierarchical organization of human behavior." *Neuron*, 50(6): 963-974.

Lorenz, K., & Tinbergen, N. (1938). "Taxis und Instinkthandlung in der Eirollbewegung der Graugans./Directed and instinctive behavior in the egg rolling movements of the gray goose." *Zeitschrift f, r Tierpsychologie* 2: 1-29.

Mackintosh, N. (1975). "A theory of attention: Variations in the associability of stimuli with reinforcement." *Psychological Review* 82(4): 276-298.

Macy, M., & Flache, A. (2002). "Learning dynamics in social dilemmas." *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl 3), 7229.

Marr, D. (1982). *Vision*. New York: Freeman.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, J., & Price, G. (1973). "The logic of animal conflict." *Nature* 246(5427): 15-18.

Mazur, J. E. (1987). "An adjusting procedure for studying delayed reinforcement." In M. L. Commons, J. E. Mazur, J. A. Nevin & H. Rachlin (Eds.), *Quantitative Analysis of Behavior* (Vol. The Effect of Delay and of Intervening Events on Reinforcement Value, pp. 55-73). Hillsdale, N.J.: Lawrence Erlbaum Associates.

McElreath, R., Clutton-Brock, T., Fehr, E., Fessler, D., Hagen, E., Hammerstein, P., et al. (2003). "Group report: The role of cognition and emotion in cooperation." in Hammerstein, P. (Ed.) *Genetic and cultural evolution of cooperation*, 125-152.

McLaughlin, J. A. (1925). "Proximate Cause." *Harvard Law Review*, 39(2): 149-199.

McNamara, J. M., Stephens, P. A., Dall, S. R. X., & Houston, A. (2008). "Evolution of trust and trustworthiness: Social awareness favours personality differences." *Proceedings of the Royal Society of London Series B* 276(1657): 605-613.

Mohtashemi, M., & Mui, L. (2003). "Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism." *Journal of Theoretical Biology* 223(4): 523-531.

Mui, L., Mohtashemi, M., & Halberstadt, A. (2002). "A computational model of trust and reputation." *Proceedings of the 35th Hawaii International Conference on Systems Sciences*

Murphy, G., & Medin, D. (1985). "The role of theories in conceptual coherence." *Psychological Review* 92(3): 289-316.

Nagel, T. (1979). *Mortal Questions*. Cambridge: Cambridge University Press.

Nowak, M. A. (2006). "Five rules for the evolution of cooperation." *Science*, 314: 1560-1563.

Nowak, M. A., & Sigmund, K. (1993). "A strategy of win-stay, lose-shift that outperforms tit-for-tat in a Prisoner's Dilemma game." *Nature*, 364: 56-58.

Palsgraf v. Long Island Railroad Co., 162 N.E. 99 (New York Court of Appeals 1928).

Parkhurst, D., Law, K., & Niebur, E. (2002). "Modeling the role of salience in the allocation of overt visual attention." *Vision Research* 42(1): 107-123.

Rachlin, H., Raineri, A., & Cross, D. (1991). "Subjective probability and delay." *Journal of the experimental analysis of behavior* 55: 233-244.

Rand, D., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. (2009). "Positive Interactions Promote Public Cooperation." *Science* 325(5945): 1272.

Rand, D., Ohtsuki, H., & Nowak, M. (2009). "Direct reciprocity with costly punishment: Generous tit-for-tat prevails." *Journal of Theoretical Biology* 256(1): 45-57.

Randløv, J., & Alstrøm, P. (1998). "Learning to drive a bicycle using reinforcement learning and shaping."

Proceedings of the Fifteenth International Conference on Machine Learning.

Renner, K. (1964). "Delay of reinforcement: A historical review." *Psychological Bulletin* 61(5): 341-361.

Rescorla, R., & Wagner, A. (1965). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement." In A.H. Black & W.F. Prokasey (eds.), *Classical conditioning II: current research and theory* (p. 64-99). New York: Appleton-Century-Crofts.

Richards, J. B., Mitchell, S. H., de Wit, H., & Seiden, L. S. (1997). "Rate Currencies and the foraging starling: the fallacy of the averages revisted." *Behavioral Ecology* 7: 341-352.

Richerson, P., Boyd, R., & Henrich, J. (2003). "Cultural evolution of human cooperation." In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 357-388). Cambridge: MIT Press.

Roberts, W. A., & Mazmanian, D. S. (1988). "Concept learning at different levels of abstraction by pigeons, monkeys, and people." *Animal Behavior Processes* 14: 247-260.

Robinson, P. H., & Darley, J. M. (1995). *Justice, Liability and Blame*. Boulder: Westview Press.

Saxe, R., Carey, S., & Kanwisher, N. (2004). "Understanding other minds: Linking developmental psychology and functional neuroimaging." *Annual Review of Psychology* 55: 1-27.

Schwartz, B., & Reisberg, D. (1991). *Learning and memory*. New York: W.W. Norton & Company.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer-Verlag.

Shiffrin, R. M., & Schneider, W. (1977). "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory." *Psychological Review* 84(2): 127-190.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). "Judgments Of Causation, Responsibility, And Punishment In Cases Of Harm-Doing." *Canadian Journal of Behavioral Science* 13(3): 238-253.

Silk, J. B. (2003). "Cooperation without counting: the puzzle of friendship." In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation* (pp. 37-54). Cambridge: MIT Press.

Sloman, S. (1996). "The empirical case for two systems of reasoning." *Psychological Bulletin* 119: 3-22.

Spelke, E. (2000). "Core knowledge" *American Psychologist* 55: 1233-1243.

Spranca, M., Minsk, E., & Baron, J. (1991). "Omission and commission in judgment and choice." *Journal of Experimental Social Psychology* 27(1): 76-105.

Stephens, D. W., McLinn, C. M., & Stevens, J. R. (2002). "Discounting and reciprocity in an iterated prisoner's dilemma." *Science* 298: 2216-2218.

Stevens, J., Hallinan, E., & Hauser, M. (2005). "The ecology and evolution of patience in two New World monkeys." *Biology Letters* 1(2): 223.

Stevens, J. R., Cushman, F. A., & Hauser, M. D. (2005). "Evolving the psychological mechanisms for cooperation." *Annual Review of Ecology, Evolution, and Systematics* 36: 499-518.

Sutton, R., & Barto, A. (1999). *Reinforcement learning*. Cambridge, MA: MIT Press.

Thompson, R., & Oden, D. (2000). "Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape." *Cognitive Science* 24(3): 363-396.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). "Understanding and sharing intentions: The origins of cultural cognition." *Behavioral and Brain Sciences*, 28, 675-735.

Trivers, R. L. (1971). "The evolution of reciprocal altruism." *Quarterly Review of Biology* 46: 35-57.

Tversky, A., & Kahneman, D. (1981). "The framing of decisions and the psychology of choice." *Science* 211: 453-463.

Waldmann, M., & Holyoak, K. (1992). "Predictive and diagnostic learning within causal models: Asymmetries in cue competition." *Journal of Experimental Psychology: General* 121(2): 222-236.

Warneken, F., & Tomasello, M. (2006). "Altruistic helping in human infants and young chimpanzees." *Science* 311(5765): 1301.

Weiner, B. (1995). *Judgments of responsibility : A foundation for a theory of social conduct*. New York: Guilford Press.

Williams, B. (1981). *Moral Luck*. Cambridge: Cambridge Univeristy Press.

Zak, P., & Knack, S. (2001). "Trust and growth." *Economic Journal*, 295-321.