Payback without Bookkeeping[1]

The Origins of Revenge and Retaliation

By Isaac Wiegman

(Forthcoming in *Philosophical Psychology*)

1 INTRODUCTION

How complex must a set of instructions be to make a complex product? According to some evolutionary psychologists, complex indeed. For example, Cosmides and Tooby (2000; 2008) call emotions "superordinate programs" with algorithms for detecting situations, assigning priorities, and making inferences. As complex as emotional behaviors may be, one worry is that a simpler set of adaptations may explain them just as well. By analogy, the paper wasp constructs wonderfully complex nests, yet it would be a mistake to think that the paper wasp has an innate blueprint for nest construction. Instead, these wasps coordinate their building behaviors by following very simple rules (Theraulaz, Bonabeau, & Deneubourg, 1999). Here I want to offer the beginnings of a similar account of revenge and payback in humans. Doubtless, the pressures that shaped and maintained payback motives in the human lineage have made it fairly complex: these motives produce a complex set of phenomena, many of which appear to involve learning and record keeping. Nevertheless, I argue that the *origins* of payback motives are in fact quite

[1] This paper would not be what it is without guidance, encouragement and feedback from the following: Justin Bruner, John Doris, Robert Fischer, Robert Kurzban, Edouard Machery, Ron Mallon, Joseph McCaffrey, Adam Morris, Burkay Ozturk, Felipe Romero, Rory Smead, and a generous reviewer for Philosophical Psychology.

simple and ancient. Consequently, it is likely to have a substantially different structure than current evolutionary models suppose (e.g., McCullough, Kurzban, & Tabak, 2012).

Current models focus predominantly on the deterrent value of punishment: individuals that are impelled to retaliate or avenge, even against their immediate self-interest, will achieve fitness benefits (in the form of deterred harm) over the long term. On one such account, the *recalibrational* account, these fitness benefits are *individualistic* and the processes that secure them *cognitively complex* (McCullough et al., 2012; Petersen, Sell, Tooby, & Cosmides, 2010; Sell, 2011): individualistic because the beneficiary of the selected trait is the individual (perhaps extended to include fitness benefits to kin) rather than a genotype or a set of individuals; cognitively complex because revenge involves complex record-keeping on the part of the avenger. If this view is correct, then it is unlikely that revenge originated before the emergence of primates (see below for clarification of this claim). An underexplored alternative is to begin by assuming that revenge originated much earlier and to see what follows. On this hypothesis, the origin of domain specific payback phenomena can and should be explained without any complex machinery for bookkeeping.

This approach becomes attractive in view of payback behaviors in nonhuman animals, particularly those not well explained by a deterrent function. A well-known model of resource competition—the war of attrition—provides a more satisfying explanation. This model identifies selection pressures that existed long before the divergence of primates and evaluates behavioral strategies that benefit populations rather than individuals. Using the war of attrition to understand payback also leads to a revision in our understanding of these phenomena because it casts payback as a much simpler capacity than described by recalibrational models. As a result, complex phenomena of attribution and bookkeeping are an unlikely product of payback motives

on their own and are more likely produced by their interaction with other mechanisms, such as causal learning or norm psychology (e.g. Gopnik, Glymour, & Sobel, 2004; Sripada & Stich, 2007).

2 THE EXPLANANDA: PAYBACK MOTIVES

What are payback motives? These motives include sentiments like anger, rage, resentment, and vengefulness,[2] which naturally lead to payback in a number of guises: retaliation, revenge, and redirected aggression (cf. Barash & Lipton, 2011, Chapters 4–5). In the grip of these motives, one is moved to confront, retaliate, or avenge, often contrary to better judgment.[3] Payback motives can be contrasted with proactive motives for aggressive behavior (Hubbard, Romano, McAuliffe, & Morrow, 2010; cf. Vitiello & Stoff, 1997). For instance, proactive aggression occurs when someone is motivated to harm another to achieve some benefit (whether material or hedonic), as when assault is committed only for the purpose of robbery. By contrast, when anger or vengefulness motivate aggressive behavior, future benefits are not the aim of the action (as understood by the agent).[4] Aggression caused by payback motives is often impulsive

[2] I make no assumptions about the relations between these motives. They may be different names for one and the same trait, or they may be derivatives of a single, ancestral trait. They may also have functions aside from payback, but this depends on how widely one casts payback. For example, I am inclined to categorize "vending machine rage" as part of the phenomenon of payback.

[3] Many cast revenge as an explicitly social and highly intellectual phenomenon. For example, Frijda suggests that it includes appraisals regarding social comparison (Frijda, 1994, p. 274). While I agree that this is an interesting and important phenomenon, it is not my focus here. Rather my interest is in payback motives more broadly, of which revenge is but one species. I also doubt that payback is strongly linked to social comparisons, cf. fn. 4.

[4] This is perfectly consistent with saying that the *function* of revenge is to secure some future benefit for the

in the sense that it requires effortful inhibition to suppress, but when it is not suppressed, there may be no forward-looking aim. For instance, when someone plots revenge, payback may be the only aim, rather than some benefit of payback. The avenger may be insensitive to whether revenge is satisfying, pleasurable, or desirable in any way aside from its fittingness as a response to provocation (cf. Carlsmith, Wilson, & Gilbert, 2008; Lambert, Peak, Eadeh, & Schott, 2014).[5] This is a puzzling feature of payback motives: unless better judgment actively resists or overturns these motives, they impel us to act without full consideration of other costs and benefits.

Why is this so? Many suspect that revenge and payback are evolved imperatives designed to secure benefits that we are unable to reliably anticipate (Daly & Wilson, 1988; Frank, 1988; McCullough et al., 2012). This is the shared starting point for the explanations of revenge that I criticize and the alternative that I offer below.[6] Consequently, I will assume throughout that payback motives are adaptive instincts and that an explanation of their *origins* need not make

---

agent. The distinction between proactive and reactive aggression concerns the agent's motive rather than the evolved function of that motive.

[5] This is not to say that aggression against others cannot be learned. For example, the phenomenon of Shadenfreude consists in pleasure at the pain of another (usually a member of an outgroup). Experiencing such pleasure passively may very well reinforce actions that cause suffering in another (Cikara, 2018; Cikara & Fiske, 2013; Leach & Spears, 2008). In any case, I am uncertain whether this phenomenon should be lumped together with revenge, though it certainly has common elements. If revenge is an instinctual motivation, this kind of "payback" almost certainly has a distinct motivational structure.

[6] See Daly and Wilson (1988, p. 226). See also McCullough et al. (McCullough et al., 2012). A related consideration is that norms prohibiting or restricting revenge are very widespread (cf. Daly & Wilson, 1988, Chapter 10). Frank (1988) draws out the implications of this fact with characteristic elegance (Frank, 1988, p. 39).

any reference to learning, cultural transmission or cultural evolution.[7]

Two final clarifications. First, revenge and payback *behaviors* fit obviously into the broader category of punishment *behaviors*. Nevertheless, my focus is on payback *motives*, which are psychological states that *cause* behavior.[8] Importantly, it is likely that there is a large range of punishment behaviors that are not plausibly caused by a payback motive.

Second, a distinguishing feature of payback motives is that they include an element of spite. By this I do not mean that a spiteful agent is motivated by the suffering or loss of the punishee.[9] Rather, I mean that the benefits of payback (especially the far off or intangible ones) are not usually a part of the individual's calculations, and as a result payback motives reliably produce costly punishment without providing any immediate material benefits, such as food, sex, or shelter.[10] In this spirit, Trivers (1971) posits a payback motive, 'moralistic aggression,' that impels agents to punish those who do not reciprocate favors, even if the agent does not benefit by doing so. A spiteful motive such as this would lead to punishment of a non-cooperator even at a cost and regardless of whether one is likely to interact with the other individual again.[11] Often

---

[7] Nevertheless, learning and culture clearly do explain much about the current function of payback motives and their functions across different cultures (e.g. Nisbett & Cohen, 1996).

[8] Though for ease of expression, I will use "revenge" interchangeably with "payback motives."

[9] In other words, payback motives are not necessarily negative social motives, which focus on satisfying negative other-regarding preferences, e.g., the preference that another person suffer (cf. Jensen, 2010, p. 2643).

[10] Another caveat is that a spiteful motive may be triggered by *imperfect* indicators of far off benefits without an agent having the goal of bringing about those benefits. On my view, spite is about the motivation that guides ongoing behavior, rather than the situation that triggers the motivation. Accordingly, my talk of "calculation" has to do with the ongoing motivation to aggress, rather than its trigger.

[11] For example, Frank (1988, pp. 36–37) criticizes Trivers's notion of moralistic aggression, on the grounds that in many situations this motive appears to be less optimal (and no simpler) than a tit-for-tat strategy, which is directly

enough, this would result in immediate losses that cooperation in subsequent interactions could not recoup. By contrast, many forms of punishment secure immediate benefits greater than the cost of punishment. Take for example the tit-for-tat strategy in the iterated prisoner's dilemma, which involves reciprocal defection. While reciprocal defection does appear to be a form of punishment, a spiteful motive is not necessary (nor optimal) for implementing it, since defection maximizes utility in response to a partner's defection. As a result, material self-interest by itself can motivate these punishment strategies simply and efficiently.[12]

To sum up, humans (and possibly other animals) have certain psychological states such as anger and vengefulness. These and other payback motives are often impulsive and stand in contrast with proactive motives for aggression. Moreover, payback motives are distinct from many other motives for punishment (such as proactive motives) because they are spiteful and can lead to costly punishment. (See table 1 for a summary of these characteristics.) Finally, my focus is on evolutionary explanations of how payback motives originated (rather than just how it was modified or maintained).

Characteristic	**Specification**
Instinctual	Some aspects of payback are unlearned
Spiteful	Can produce costly punishment that is not instrumental for immediate material benefits

in line with self-interest. That is, if material benefits were one's only motivation, one would only punish a defector to achieve a higher payout, to cut losses or coerce the partner to cooperate (in accordance with a tit-for-tat policy).

[12] Hence, I leave aside a great deal of important work on punishment in evolutionary game theory, such as games in which the immediate benefits of punishment do outweigh its costs (cf. the discussion of "loss-cutting" strategies in Nakao & Machery, 2012). I leave aside other work because I do not think it adequate for explaining the *origins* of payback motives as opposed to their subsequent modification (e.g. Clutton-Brock & Parker, 1995; Nowak & Sigmund, 2005).

Multiple instances	Anger, rage, resentment, vengefulness

Table 1 Summary of payback motives.

3 PAYBACK MOTIVES AS ADAPTATIONS FOR DETERRENCE

There have been several attempts in the last 30 years to give revenge and related phenomena an evolutionary explanation (Daly & Wilson, 1988; Frank, 1988; McCullough et al., 2012; Petersen et al., 2010; Trivers, 1971). One family of models, *recalibrational models*, begins by supposing that the function of revenge is to solve an "adaptive problem that is faced by many species: how to change other organisms' incentives to emit benefits and to avoid imposing costs upon oneself" (McCullough et al., 2012). For instance, suppose that Amjad imposes a cost on Baggie and Amjad thereby receives a benefit. Baggie needs a way to prevent Amjad from imposing further costs. Recalibrational models start with an intuitive solution to this problem. If Baggie reacts by imposing a reciprocal cost on Amjad, then if Amjad is rational, she will lower her estimate of the value of gains at Baggie's expense. In effect, Baggie has 'recalibrated' Amjad's dispositions toward him. Thus, it is easy to see that recalibrational models are based on the adaptive value of deterrence. If we inquire of this model, "Why do individuals recalibrate?" the answer is that this tendency benefits those who possess it by deterring the imposition of future costs. Moreover, in keeping with the definition of revenge above, recalibration is spiteful, since the benefits of recalibration need not figure into the avenger's calculations. While a recalibrated disposition is a benefit in some sense, it is not an immediate material benefit, and the material benefits it secures are far off.

This approach manifests two typical starting points for a dominant school of thought in evolutionary psychology (cf. Tooby & Cosmides, 1990, 2008). First, theorizing begins by characterizing a putative adaptive problem and reverse engineering a psychological capacity

required to solve it. Second, the selection pressures that shaped the psychological capacity are conceptualized individualistically. In other words, it is the individual–rather than a group or a genotype–who benefits from the tendency to recalibrate others. [13]

While the adaptive problem here is supposed to be common to many species, the solutions proposed for humans are complex. On these models, it is assumed that one's dispositions concerning another person's welfare are integrated via "internal regulatory variables, stored in memory and continually updated, that humans use to guide social decision making according to appropriate criteria…" (McCullough et al., 2012). In this case, the regulatory variables are called welfare tradeoff ratios (WTRs), and they represent the value that one person places on the welfare of another. Thus, when Baggie recalibrates Amjad, this means that Baggie has increased Amjad's WTR toward Baggie.

So the function of revenge on this model is to recalibrate low WTRs. Indeed, some theorists use this function as a kind of definition for revenge and anger (the dominant motive for revenge):

> Given the substantial impact of other individuals' WTRs on one's fitness, natural selection can be expected to have designed a mechanism to interface with WTR-setting machinery in others and recalibrate them so as to raise their WTR. According to the recalibrational theory, anger is this mechanism. (Sell, 2011, p. 382)

[13] I do not intend "individualistic" to rule out increases in inclusive fitness that sometimes translate to increased reproductive fitness in an individual's kin. Similarly, I do not mean to rule out retaliation or revenge at the level of groups. First, retaliation of a group against another may exist because of pre-existing tendency of individuals to retaliatiate. Second, group retaliation may exist because of fitness benefits conferred on individuals.

On this definition, an organism has a capacity for anger and revenge (McCullough et al., 2012) just in case it has a mechanism that "interfaces with WTR-setting machinery." If we think of WTR machinery as a system for interpersonal bookkeeping, then there is no revenge without bookkeeping according to this theory.

Importantly, on this theory bookkeeping involves a complex attributional machinery:

> …we also posit the existence of cognitive routines for registering that an actor has treated the self with less regard (i.e., that an actor has committed an action that connotes a lower WTR toward the self) than one would have expected based on one's previous estimate of the actor's WTR toward the self. At issue here is not simply whether an individual imposed a cost upon oneself, but whether that cost imposition was *permissible given the victim's understanding of the harmdoer's WTR for the self*… Understanding a harmdoer's intentions is important *because accidental harm does not provide information about the actor's WTR*. Intentional harm, however, implies that the harm was caused by the harmdoer's low WTR for the victim… (McCullough et al., 2012, p. 4, emphasis mine)

So according to this model, monitoring the WTRs of others requires attributing intentions to others. While these attributions need not be conscious or explicit, they are quite complex. This is because they require that the organism respond differently to intentional and accidental harms, since the first, but not the second, indicates a low WTR (as indicated in the quote above). Yet, behavioral sensitivity to the intentions of others has only been evidenced in a handful of species aside from humans (e.g., Phillips, Barnes, Mahajan, Yamaguchi, & Santos, 2009). As a result, the definition of revenge above (which links it to this complex attributional machinery) makes it

unlikely to have arisen before the divergence of primate species.[14]

Perhaps this consequence could be avoided by supposing that complex WTR machinery is nonessential to the model. Instead, one could postulate simple heuristics that other organisms use to compute WTRs. Yet, if these heuristics do not enable the organism to respond differently to accidental and intentional harms, then the simpler machinery is not really tracking WTRs at all (as recalibrational theorists define it). So, if rats get revenge on one another but cannot respond differently to accidental and intentional harms, it would be incorrect to suppose that their revenge is triggered by information about welfare-tradeoff-ratios. It would also be superfluous. Even if we suppose that deterrence is the defining function of revenge, a rat need not be sensitive to WTRs *as such* (e.g., as indicated by intentional but not accidental harms) in order to deter a competitor's behavior. One advantage of my proposal (cf. section 6) is that it can explain payback behaviors in human and nonhuman animals without appeal to complex behavioral sensitivities of this kind.

In sum, recalibrational models begin by postulating a function of deterrence; this function is individualistic, and its psychological implementation requires complex machinery for bookkeeping. On this family of models, the selection pressures that shaped revenge (and which likely explain its origins) were pressures on individuals to deter harms (and withhold benefits) from other individuals *by reacting to low WTRs and recalibrating them* (where reacting to low WTRs *per se* requires behavioral sensitivity to the difference between accidental and intentional

[14] Corvids are an obvious exception. Nevertheless, the mindreading capacities of corvids are probably best explained by parallel evolution rather than common descent. Thus, the evolution of complex bookkeeping capacities in corvids is irrelevant to its origination in mammalian species.

harms). Insofar as this is seen as essential to revenge, this sets a very low bound on how long ago revenge motives originated. Table 2 sums up these elements of the theory.

Element	**Specification**
Fitness beneficiary/ manifestor of adaptation	Individuals (and kin)
Function of revenge	Prevent harms against individual (and kin)
Implementation	Impose cost on low WTR
Elicitor of revenge	Indicator of low WTR
Cognitive complexity (computations necessary to detect elicitor)	High (e.g. must distinguish between intentional and accidental harms)
Ancestral Origins	No earlier than primates

Table 2 Summary of distinctive elements of the recalibrational theory of revenge.

4 Beyond Deterrence: Payback Phenomena in Other Animals

While the foregoing theory of revenge is powerful, there is a wide range of behavioral phenomena in the animal kingdom that it does not easily explain. Specifically, some animal behaviors resemble revenge, and these behaviors are not well-explained as adaptations for deterrence. If they are not adaptations for deterrence, then recalibrational models cannot explain them. Additionally, the existence of payback behaviors in nonhuman animals suggests that payback motives did not originally evolve for recalibration, since, as I have argued, nonhuman animals cannot track WTRs. This argument is not intended to be decisive. Rather, taken together, these considerations motivate a novel proposal concerning the origins of payback motives, one that has distinctive explanatory benefits.

Now consider patterns of resource competition in nonhuman animals—competition for fitness-relevant resources such as physical territories, sexual partners, or position in dominance

hierarchies. While these patterns are widespread in the animal kingdom, they have been most carefully studied in rodents. For instance, once a male rat has established a territory, it will defend it from unfamiliar conspecific males, and certain patterns emerge in these interactions (see, e.g., Blanchard & Blanchard, 1984; Blanchard, Litvin, Pentkowski, & Blanchard, 2009). The 'resident' of a territory tends to exhibit specific behaviors and nonlethal attack strategies aimed at accessing and biting the back of the unfamiliar 'intruder.' By contrast, the intruder tends to take on certain strategies aimed at fleeing the owner or, failing that, preventing access to the back.

Similarly, many mammals exhibit distinct but paired strategies for offense and defense that generally lead to nonlethal competitive encounters (for a review, see, e.g., Blanchard & Blanchard, 1984, pp. 22–28; Blanchard et al., 2009). For instance, Adams (1980) argues that these patterns generalize fairly well to the broader category of muroid rodents, and Leyhausen (1979) observed similar patterns in a wide range of felids (though lethal encounters may be more likely in some species, e.g., Schaller, 1976).

Though there are exceptions to these patterns, a central tendency in many species is that relatively dominant animals or established owners of a resource usually win contests. In many cases, the intruder/challenger assumes defensive postures from the beginning of the interaction and exhibits a motivation to flee. Nevertheless, in rodents if the intruder is cornered so that flight is prevented, then the defensive maneuvers of the intruder do not diminish the attacks of the resident. These attacks can continue without reprieve, almost up to an hour if the intruder is left in the cage of the resident (Potegal & TenBrink, 1984). Leyhausen makes similar observations concerning cats and draws out the importance of this:

> From all this, it is clear that the defensive posture is not a submissive gesture

> in Lorenz's sense… It does not offer up to the superior attacker the object of its attack—the nape of the neck—but seeks to protect it. In addition, it does not necessarily inhibit the attacker, and the attacked animal does not remain passive in the face of further threats but defends itself and, in certain circumstances, proceeds to counterattack. The attacker is inhibited only by the removal of its target and the danger involved in continuing to attack, i.e. the threat being expressed in the defensive behavior – in other words, precisely the opposite effect to that of a genuine submissive posture. (Leyhausen, 1979, pp. 186–187)

The point is that defensive behaviors do not abate offensive attacks. Moreover, this pattern is thought to apply broadly among vertebrates, to the extent that Lorenz's concept of submissive behavior (as an inhibitor of offense) has been challenged as illusory (e.g. Schenkel, 1967).

Importantly, defensive behaviors are strongly associated with *loser-effects* (for a review, see Hsu, Earley, & Wolf, 2006). Loser effects can be induced in rats by pairing the intended loser either with a rat that is slightly larger or one that is from a more aggressive strain and is thus more likely to win the fight (e.g. Lehner, Rutte, & Taborsky, 2011). Over the course of these encounters, the smaller or less aggressive rat will eventually begin to take on defensive postures. In subsequent encounters with rats matched for size and aggressiveness, the loser effect takes hold: the rat who lost the first bout is more likely to lose again (by being the one to take on defensive postures, see e.g. van de Poll, Smeets, van Oyen, & van der Zwan, 1982). This pattern appears to be connected with the formation of stable dominance relationships and appears to be a dominant strategy in game theoretic models of the loser effect (Dugatkin, 1997; Dugatkin, Druen, Dugatkin, & Druen, 2004; Mesterton-Gibbons, 1999). Consequently, once a competitor takes on defensive postures, it is no longer a real threat to a territory owner or dominant rat. This

is beneficial for the loser, because in subsequent competitive interactions, the competition is resolved more quickly and the loser receives less aggression, because they are quicker to retreat and hide (Lehner et al., 2011).

We can draw two conclusions from these facts about offense. The first is that the motivation for offensive attacks resembles payback in that it is instinctive and so also spiteful (cf. section 2). In many experiments, offensive behaviors are exhibited by naïve rats, without prior fighting experience (e.g. Lehner et al., 2011). Thus, they have no experience that would inform them that offensive behaviors are instrumental for a desired outcome. Since the attacks are also costly, they appear to be spiteful rather than being instrumental (from the organism's perspective) for achieving some benefit. Table 3 provides an overview of these similarities to payback motives.

Characteristic	**Specification**
Instinctual	Rats without prior fighting experience defend territory from intruders.
Spiteful	Resident attacks intruder well beyond the intruder's adoption of postures that indicate deference and subsequent deterrence. Thus, there is no clear, immediate, instrumental benefit to the attacks.

Table 3 Similarities between motives for rodent offense and payback motives.

One might object that the prior owner of the resource does in fact receive an immediate benefit from defending it, namely the value of the resource. While it is true that consumable resources like food hold relatively immediate value, patterns of offense and defense do not only concern food, but also territories and positions in dominance hierarchies (Blanchard et al., 2009). Moreover, the value of these resources is extended over time. For instance, in some organisms,

the value of a territory for acquiring food is extended into the future, and there is evidence that aggression is tuned to the *future* value of the territory and not only its immediate value (e.g. Stamps & Tollestrup, 1984). In other organisms, territories serve reproductive purposes. Where reproduction occurs only at certain times of the year, these animals will fight over territories prior to the reproductive phase. For instance, in many avian species, territories are defended before advertising for potential mates and prior to mating, laying nests or parenting young (e.g. Wingfield, Ball, Dufty, & Hegner, 1987). In a sense, these organisms are setting up shop in a territory and the value of defending it is prospective. Therefore, the aggressive actions of a territory owner in this case are spiteful in my sense: they appear to be a reaction to the incursion of a competitor and the reaction is not a calculated response to the immediate value of the resource (since its value is extended over time) or perhaps even its discounted future value (since few animals are capable of representing future states).

Further evidence for this can be found in animals such as rodents and felids. If offensive attacks are motivated to secure immediate benefits, we would not expect animals to continue fighting for long after the challenger starts behaving defensively. Nevertheless, many species do carry on attacks well after the intruder adopts defensive postures (as discussed above). Apparently, the immediate benefit of retaining control of the territory or resource has already been achieved once an intruder takes on defensive postures, so subsequent fighting appears to be spiteful.

The second conclusion is closely related to this point: offensive attacks are not well adapted to the function of deterrence. If we thought their function were deterrence, we would probably predict that the adoption of defensive behaviors by a competitor would inhibit offensive attacks. Once defensive behavior commences, the loser of the fight is established and the loser is already

substantially deterred from challenging the winner on subsequent occasions. Intuitively, it seems a waste of effort to continue attacking an intruder at a cost when the intruder is already inclined to flee and to take on the losing role in subsequent interaction (as indicated by the loser effect). This suggests that the offensive attacks (of a resident or dominant) are not as finely tuned to the purpose of deterrence as they could be.

Of course, there are many avenues of response that a deterrence theorist could pursue here. Among other things, we should not assume that motivational systems will be perfectly optimized to serve their adaptive function. However, there is not space here to tie up all these loose ends. Instead, I point to the apparent lack of optimality vis-à-vis deterrence to raise the possibility that there is another adaptive function of these motivational systems; one that better explains the tendency of the attacker to continue attacking at a cost, even when the contest has been decisively won. I consider such a possibility in the following section.

5 A New Proposal: Adaptation for Enforcing Competitive Conventions

This alternative function for revenge arises out of a game theoretic model that captures the dynamics of frequency-dependent selection. This kind of selection exerts different pressure on a trait depending on the frequency of variant traits in a population. Once a formal model is developed to capture these dynamics, one can evaluate the effects of this kind of selection on social interaction strategies using computer simulations or analytic methods, such as proofs. For instance, given various strategies for interacting in a game like the iterated prisoner's dilemma, one can evaluate the average payoff of a tit-for-tat strategy when played in populations consisting of organisms with various other strategies, like an always-defect strategy or a tit-for-two-tats strategy (e.g. Axelrod, 1984).

In these models, successful strategies are more likely to be present in subsequent generations,

and so these models can tell us which strategies selection is likely to favor given the frequency of other strategies in the population. The concept of an evolutionarily stable strategy (ESS) captures some of the factors that allow a given strategy to persist in a population. Maynard Smith and Price offer this definition: "Roughly, an ESS is a strategy such that, if most of the members of the population adopt it, there is no 'mutant' strategy that would give higher reproductive fitness" (Maynard Smith & Price, 1973, p. 15). Given this definition, we can expect selection to gradually weed out almost all other strategies from a population aside from (one of) the ESS(s). Thus, if frequency dependent models apply to a given species-typical behavior, the evolutionary stability of the modeled behavior can offer a powerful explanation for its species-typicality.

The point of rehearsing these fairly obvious features of widely used models is to make salient what is distinctive about them for my purposes: that the behavioral strategies that are likely to evolve as ESSs usually do not function to benefit an individual organism, but rather to prevent the spread of variant strategies in a population. As Dawkins points out, "An ESS is stable, not because it is particularly good for the individuals participating in it, but simply because it is immune to [invasion by other strategies]" (or the genes that produce them, cf. Dawkins, 2006, p. 72).[15] So, if payback is explained by such a model, it is unlikely to be an adaptation that benefits an individual organism by deterring behavior toward that specific individual. If there is any deterrent function at all, it is instead to penalize variant strategies and deter the spread of those strategies in a population. In this sense, such strategies are selected because they deliver benefits

[15] Dawkins, however, expresses skepticism at the idea of a population-level function. He would also likely balk at calling anything a population-level adaptation, as I do below. Regardless, we can agree that the aggression of owners/residents is not an adaptation that exists to benefit individuals.

of deterrence to the *set* of organisms that possess them (at a certain population state) or the set of genes that produce them (Dawkins, 2006, p. 86). Nevertheless, these strategies may regularly result in costs to many *individuals*: ones that may not be recouped in their lifetime. For traits with this etiology, the adaptive function is not to deliver a fitness benefit (via deterrence) to the individual that possesses the trait, and this is a stark difference from recalibrational models of payback considered previously (in section 2), which are imagined to deliver individual benefits. (I discuss this in greater detail at the end of section 4.2) To be clear, it is the individuals who interact to produce these selection dynamics, but the benefits of the dynamics are manifested at the level of the population. We can mark this distinction by saying that the individual is the *interactor*, but the population is the *beneficiary* of the selective process.[16] We could also say that a population that reaches an ESS *manifests an adaptation* that protects the *population* from invasion by variant strategies (without necessarily protecting *individuals* from losing out to variant strategies).[17]

The model of interest here is called the 'war of attrition' (Bishop & Cannings, 1978; Maynard Smith, 1974).[18] This is because it was developed to explain animal contests in which the costs of fighting build up over time and in which a disputed resource goes to the organism

---

[16] See Lloyd (2007) for an overview of the units of selection debate that sheds light on this distinction.

[17] This is not at all clear for polymorphic ESSs, where there is a stable equilibrium with two or more strategies. In that case, the benefits accrue to more than one trait, and no single, heritable trait is the source of the benefit. So neither trait appears to be an adaptation.

[18] Aaron Sell (2005) has proposed an evolutionary model of anger that also appeals to the war of attrition model. Nevertheless, Sell's model falls squarely within the family of recalibrational models discussed above and falls prey to the criticisms offered in section 3. Gintis (2006) and Descioli and Wilson (2011) have also referred to this model to explain patterns of territorial behavior in humans. Nevertheless, their efforts have been directed at explaining peoples' desire to keep what they have (i.e. endowment effects) or the fact that they are usually successful in doing so, rather than the motivational states that lead one to defend what one has. My contention here is that revenge is one such motivational state.

that persists the longest. Many have argued that these models accurately predict a great deal about the observed structure of animal contests for resources such as food, territories, or mates (see, e.g., Dawkins, 2006, pp. 79–81; Huntingford & Turner, 1987, p. 282; Krebs & Davies, 1993, pp. 156–170; Maynard Smith, 1982, Chapter 3). For instance, they explain why aggressive interactions regarding resource competition (whether for food, potential mates, or territories) are rarely protracted and why the outcome of such an interaction will tend to favor the prior 'owner' of a given resource (Bishop & Cannings, 1978; Haccou & Glaizot, 2002; Hammerstein & Parker, 1982; Maynard Smith & Parker, 1976; Parker & Rubenstein, 1981). Generalizations and special cases of the war of attrition model even explain exceptions to the patterns above: where the owner of a resource stands a real chance of losing, where aggressive contests progress in stages and can escalate to levels where serious injury becomes possible.[19]

In most cases, the war of attrition model predicts that owners will win contests in a majority of cases because ownership is an arbitrary asymmetry, and once introduced, ownership conventions that resolve contests based on such asymmetries are much more successful than strategies that ignore them. Accordingly, I propose in this section that payback motives originated as an adaptation for preventing the invasion of convention-breaking mutant strategies

[19] For instance, Archer and Huntingford (1994) discuss the application of the sequential assessment model to escalated aggressive encounters (e.g. Enquist & Leimar, 1983). I consider this model to be an extension of the generalized war of attrition discussed below in that it assumes that costs build up over time and that assessment of asymmetries in resource holding power are decisive in determining contest outcomes. The main difference is that the sequential assessment model assumes that the costs of fighting (e.g. risk of injury) increase for both contestants (sometimes in stages) as the contest proceeds and that assessment strategies use information obtained during the competition. Something like a reserve strategy (discussed below) remains intact, and that is why 'bluff' strategies (that proceed past a stage at which a contestant assesses its resource holding power to be lower than its competitor) are unstable as a rule.

in asymmetric, temporally extended contests for ownership of resources.[20] I call this the *convention enforcement theory* of payback.

5.1 Contests of Ownership: From Symmetry to Asymmetry

To begin, what is the nature of so-called ownership conventions, and why would such an asymmetry make a difference in animal contests? To see why, consider a symmetric game. Suppose two equally matched contestants are vying for a resource that they value equally, and suppose that the winner will be whoever persists the longest, where persisting in the contest comes with steadily accumulating costs. If every contest for a resource is resolved in this way, which strategies will be most successful?

First, consider pure strategies, ones that persist for the same amount of time, *m*, in every encounter. It turns out that no *pure* strategy for this game is evolutionarily stable. If one assumes that there is such an ESS, one can derive a contradiction by demonstrating the existence of a strategy that has a better payoff. Regardless of the value of *m*, there is always a competing strategy with a better expected payoff in a population of organisms that persist for an interval of *m*.

Instead, the ESS against any pure strategy will be a *mixed* strategy in which organisms choose from a probability distribution of persistence intervals at each encounter. More specifically, the mean of the distribution for the ESS is an interval that accrues a fighting cost equal to the value of the resource under dispute. In a population that consists entirely of this

[20] It may be that payback motives also implement strategies in discrete games like Hawk/Dove. Nevertheless, these games seem to me too idealized to capture important dimensions of animal conflict (cf. fn. 23). Moreover, such models make it more difficult to isolate and explain critical strategic elements (i.e. the reserve strategy).

strategy, no pure strategy can invade. However, the expected value of this strategy is still only zero in a population in which everyone adopts it (see Maynard Smith, 1974). The organism playing this strategy (in such a population) is unlikely to gain anything when the average cost of persisting and the average benefit of winning are summed up.

Maynard Smith (1974) points out that a better strategy would be to decide competitions with a coin toss. In a population dominated by the mixed ESS, the probability that an organism would win any given contest is .5 anyway. So instead of wasting energy determining who by chance happens to persist longest in a given match, everyone would benefit if the contest were instead determined by coin toss. With such a scheme in place, no one would accrue the costs of persisting. By flipping a coin, we introduce an arbitrary *asymmetry* into the contest, and everyone is better off if the asymmetry is used to resolve contests by *convention*.[21] The expected value of adopting a conventional strategy that determines contests by coin toss would be half the value of the disputed resource for each contest, which is far better than any strategy that ignores the coin toss (zero for the mixed ESS that ignores the asymmetry).

If we look to nature, animals use an arbitrary asymmetry in just this way: whoever found the disputed resource first, or in other words, whoever happens to 'own' it. If all such contests are dyadic interactions, then on average, an organism will be the owner of the resource in about half of the contests in which it becomes involved. Thus, ownership can be used in the same way as Maynard Smith's coin toss. If a population of organisms were to decide contests in the favor of resource owners, this convention should have the same effect as deciding contests by a coin toss. Game theorists call this the 'bourgeois convention.' This convention is actually one of two

[21] See also Skyrms (1996, Chapter 4) for a helpful discussion of correlated conventions that break symmetry.

conventions that can break symmetry. The other is called the 'paradoxical ESS' in which the prior owner of the resource gives it up to an opponent. However, it is only rarely observed in nature, perhaps because there are usually some correlated costs associated with ceding ownership.[22] The set of strategies that use the bourgeois convention to settle contests I will call 'bourgeois strategies' (following others). Just like the coin toss strategy, an organism following the bourgeois convention can expect to get half the value of all the resources that it competes for in a population of organisms that follow the convention.

The stability of bourgeois strategies may help explain why owners of resources usually win fights in a variety of species. It may also explain why flank marking, urinating strategically at the boundaries of one's territory, is so common among mammals. Even in absence of strategic flank marking, animals will inevitably urinate and defecate on their territories at a higher frequency than they would elsewhere. Thus, a territory will often end up smelling like its owner, making smell a difficult-to-fake signal, or index, of ownership (Maynard Smith & Harper, 2003). Given the reliability of this index, it is easy to determine which contestant in a territorial dispute is the owner of the territory. Thus, territory ownership is an unambiguous asymmetry that can be exploited to determine the outcome of contests.

5.2 The Reserve Strategy and Spite

Importantly, the stability of bourgeois strategies depends on ownership being backed up by force. The bourgeois strategy must include a 'reserve' component, which involves fighting for a length of time drawn from a probability distribution (the same distribution as the mixed strategy

[22] As Skyrms (1996) puts it, once the correlated costs are figured in 'the basin of attraction of the bourgeois equilibrium will now be larger than that of the paradoxical strategy.' (p. 78) See also Dawkins (2006, p. 81).

described above), in case the convention is not respected. Otherwise, a bourgeois convention will not be stable against a convention-breaking 'mutant' strategy that ignores ownership and fights in every encounter "using the reserve strategy of the rest of the population" (Parker and Rubenstein 1981, p. 225). If the bourgeois convention is not backed up by a reserve strategy on the part of owners (e.g., if they were to relinquish the resource when an intruder attacks), then a certain range of mutant strategies can 'call bluff' and win almost every contest with minimal cost in a population of bourgeois strategists. [23]

The set of bourgeois strategies that include a reserve strategy, I will call "bourgeois reserve strategies." In a population dominated by this strategy, the reserve component will never be observed (unless through some mistake in who is the owner). If everyone in the population respects ownership, then intruders will forfeit the resource to the owner before the owner plays the reserve. Moreover, on this model, the stability of the bourgeois strategy depends on there being a fixed tendency to play the reserve strategy when the convention is violated (as explained above), so the motivation to play the reserve strategy cannot depend on the rewards that accrue to playing reserve. At the very least, a population of bourgeois strategists in which the reserve component is learned via rewards would be more vulnerable to the spread of a convention breaking mutant strategy. Mutants will do no better than bourgeois reserve strategists in a

[23] It will be obvious to some that the war of attrition model is not the only one in which symmetry can be broken by correlated convention, and perhaps not the only model in which the convention must be enforced. By focusing on the war of attrition model as opposed to, for instance, hawk/dove, I have suggested that the origin of revenge derives from temporally extended contests, but does the argument generalize to iterated games in which conventions break symmetry? It may. However, I have a misgiving about explaining the origins of revenge in terms of the iterated hawk/dove game. It is that iterated games simply do not apply to a large class of animal conflicts. The hawk strategy is usually conceptualized as a discrete decision to 'Escalate, and continue until either opponent retreats, or until injured.' (Maynard Smith & Parker, 1976, p. 161) However, almost all animal conflicts have the possibility of temporal extension. Insofar as injury is unlikely to occur in the initial moment of a contest, the space of strategies expands to include decisions about not just whether to escalate but also how long to persist, and so, in these conditions, the interaction may reduce to war of attrition after all. The same is true of various iterated games, such as the retaliation game (cf. Maynard Smith & Parker, 1976, p. 173).

population consisting entirely of bourgeois reserve strategists, but only if reserve is played almost every time a mutant competes (or in other words, only if the population really consists of bourgeois reserve strategists rather than mere bourgeois strategists). In such a population, the mutant may win every fight for a resource, but it will take on considerable costs in about half of its disputes, whereas the bourgeois reserve strategists will never take on costs for persisting (except in the rare encounter with the mutant) but get the resource in about half of their fights. In a population of bourgeois strategists that learn to play reserve via rewards, most of the population would have to learn to play reserve before the mutant strategy received any penalty for violating the convention. Thus, bourgeois reserve strategies are more stable than bourgeois strategies in which the reserve component is learned. This is because the bourgeois reserve strategy requires that organisms play the reserve strategy without assessing the immediate costs of doing so. In other words, what makes the bourgeois reserve strategy stable is that it is instinctual and spiteful: it is unlearned and it imposes costs without immediate material benefits.[24]

Nevertheless, the motive clearly is not selected for *individual* deterrence on this model. To see this, consider a counterfactual. For any population in which the bourgeois convention is

[24] In reality, I am not claiming that one cannot learn to follow a bourgeois convention (cf. Skyrms, 1996, Chapters 71–75). However, I do claim that the bourgeois reserve strategy *as evaluated in war of attrition models* cannot be learned. That is, learning from individual experience would not tend to converge on the bourgeois reserve strategy (as it is ordinarily defined) across all of the conditions required for its stability. Since learning would not reliably produce the relevant phenotype, the bourgeois reserve strategy cannot develop by this mechanism. If one asks why these models should not include reserve learning in their strategy space, there are two reasons. First, the symmetric war of attrition seems to me the most plausible model for resource competition at the outset. Moreover, it is difficult to see how the mixed ESS (to pick a duration interval from a certain probability distribution) in this game could be learned. Moreover, the mixed ESS in the symmetric game is identical to the reserve strategy in the ESSs for the asymmetric war of attrition. Thus, the most plausible evolutionary trajectory from the symmetric game to the asymmetric game is a reserve strategy that is not learned. Second, recent work also suggests that punishment is unlikely to be learned when it is costly and strategies that always punish (without learning) are more likely to evolve if stealing resources (similar to violating a convention) is rewarding (Morris, Macglashan, Littman, & Cushman, 2017).

fixed, that population could have instead broken symmetry with the paradoxical convention. If it had, the reserve strategy would still be in play. The difference is that second-comers would play reserve, whereas the first-comer would give up without a fight. In this case, we would not suppose that the function of the reserve strategy is individual deterrence. Perhaps it prevents a first-comer from holding on to a resource, but it certainly does not prevent other individuals from taking the resource once it is owned (quite the opposite). Any attempt to pin the function as *individual* deterrence in the paradoxical case will thus have to cast the function of reserve as entirely different from its function in a population of bourgeois strategists. By contrast, if we drop the attempt to posit an individualistic function, we can say that the reserve strategy has the same function in both cases: to enforce whichever convention has been fixed in the population or perhaps to deter the spread of alternate strategies. Since the reserve strategy arises when symmetry is broken, and since symmetry can be broken in either of two ways, the non-individualistic interpretation of its function is more accurate in addition to being more parsimonious. Either way, the reserve strategy functions to enforce whichever convention arises, and this is the central commitment of the convention enforcement theory.

5.3 The Generalized Asymmetric War of Attrition

Much of the literature on the war of attrition complicates the background assumptions with which we began. For instance, the models that I have been reviewing so far assume that competitors are equally matched and that a given resource is equally valuable to them. However, these are not safe assumptions for most species. That is, in most species, differences in fighting ability or robustness make it less costly for some individuals to persist in a competitive encounter. These differences in cost influence the structure of the war of attrition in a way that advantages organisms that can accurately assess a competitor's fighting ability (or more

accurately, resource holding power) and desire for the resource (Parker, 1974).[25] These asymmetries introduce interesting changes in the ESS for the war of attrition. The generalized war of attrition model suggest that organisms will decide competitions based on a combination of other variables instead of deciding competitions only on the basis of ownership.[26] Regardless, on these models the reserve component of the strategy remains intact, meaning that when the relevant asymmetry is not respected, the ESS against convention-breaking mutants is for owners to play reserve. For instance, work by Haccou and Glaizot (2002) suggests that the owner and intruder both fight for a certain period of time, with only an infinitesimal probability that the intruder will persist as long as the owner.[27] In experiments with humans, Descioli and Wilson (2011) showed that a similar pattern holds in a virtual environment in which people interact through avatars. While intruders did win fights on occasion, owners tended to persist longer than intruders, even in some cases when the owner was smaller and less capable of damaging the intruder, suggesting that something like the reserve strategy remains in play.

Since I have argued that the reserve strategy is what creates the need for a payback motive, the generalized war of attrition remains a good explanation of the existence of payback motives.

---

[25] Another asymmetry involves the value of a resource to an individual (Grafen, 1987; e.g. Parker & Rubenstein, 1981). If an individual has a greater need for food, for instance, the value of a given food item will be greater to that individual than to an individual who is less hungry.

[26] For a formal description of these models, see Parker and Rubenstein (1981, pp. 223–225).

[27] This model addresses a worry about using the ESS methodology in the war of attrition (cf. S. Huttegger, 2010; S. M. Huttegger & Zollman, 2012): that the bourgeois reserve strategy may not be an ESS against a simple bourgeois strategy, one that respects the convention but does not play reserve. That is, at many possible population states, there will be no behavioral differences between strategies that play reserve and those that do not. If mutant strategies do not invade and role assessment is perfect, then these strategies will look behaviorally identical. If so, this introduces the concern that the convention would not be stable against drift in the persistence time of owners or intruders. Indeed, it is not stable under these conditions (Hammerstein & Parker, 1982). Haccou and Glaizot's (2002) model resolves this worry for the generalized war of attrition by showing that when role perception is not perfect, something like the reserve strategy remains intact, even when a wider range of strategies are at play, at least under certain plausible assumptions (e.g. about the likelihood of mistakes).

At the very least, some combination of war of attrition models is likely to explain the existence and maintenance of payback motives over a large swath of evolutionary history, and these models are consistent with the convention enforcement theory of payback motives.

There are two additional virtues of this theory. The first is that it lays bare what may be the basis for the metaphor of payback that dominates our thinking about revenge and retaliation. Remember that the reserve strategy is to select a duration of fighting from a probability distribution that corresponds to the value of the resource. In other words, the optimal implementation of the reserve strategy will be a *proportional* fighting strategy that extracts a cost that is (on average) equal to the value of what would otherwise be ill-gotten gains (ill-gotten in the sense of "acquired by breaking the convention"). This is payback in its most primitive form. Second, this account does not require any complex cognitive machinery for the strategies to have their adaptive effects or for the strategies to be implemented. For instance, war of attrition models obviously do not require that the recipients of aggression learn from it, nor do they require any kind of bookkeeping mechanism on the part of the aggressor. Table 4 summarizes the elements of this theory in contrast with the recalibrational theory.

Element	**Specification**	
	Recalibrational Theory	*Convention Enforcement Theory*
Interactor in selection	Individual	Individual
Fitness beneficiary/ manifestor of adaptation	Individuals (and kin)	Population at (monomorphic) ESS
Function of	Prevent harms against individual	Prevent spread of variant

revenge/payback	(and kin)	strategies
Implementation	Impose cost on low WTR	Attack convention-breakers
Elicitor of revenge/payback	Indicator of low WTR	Indicator of violated convention
Cognitive complexity (computations necessary to detect elicitor)	High (distinguish between intentional and accidental harms)	Low (e.g. identify intruder/challenger, detect violated expectation of reward or non-punishment)
Ancestral Origins	No earlier than primates	Far earlier than primates

Table 4 Contrasting specifications of competing theories of payback.

6 Co-option for Deterrence and Subsumption of Recalibrational Models

Nevertheless, this model does not entirely rule out other explanations for the structure and maintenance of payback motives (rather than their origins). It is quite possible that the motivational states required for an ESS in the war of attrition were subsequently co-opted and modified for different purposes at different points in evolutionary history. Nevertheless, if the origins of the payback motive are as simple and ancient as I suggest, then this putative trajectory may conflict in important respects with the recalibrational models of revenge offered by many evolutionary psychologists (discussed above in section 2.1). On these models, payback motives such as anger are triggered primarily (and on some accounts, exclusively) by changes in internal regulatory variables such as WTRs that are continuously updated and involve complex inferences from the behavior of others (McCullough et al., 2012; Petersen et al., 2010; Sell, 2005, 2011). As discussed above, determining how much another person values one's welfare requires a complex (though not necessarily conscious) attributional machinery: one that can register the difference between, for instance, harms that occur intentionally and unintentionally

or perhaps even the difference between negligent and non-negligent accidents.

By contrast, according to the model on offer here, payback motives may be triggered by a much wider range of inputs, some of which can involve much simpler inferences (if any at all). This is because the violation of an ownership convention can be detected through very simple indicators. For instance, the presence of an unfamiliar conspecific male in another male rat's territory appears to function as an indicator that an ownership convention has been violated. For another example, a toddler's simple expectation that she will keep what she has found, and her predisposition to get angry otherwise (Alessandri, Sullivan, & Lewis, 1990; Michael Lewis, 1990), may function as an enforcement of an ownership convention.[28]

The latter possibility is especially interesting in connection with a vast literature on the frustration–aggression hypothesis. This hypothesis suggests that aggression can be triggered by frustrated expectations of reward or non-punishment (Berkowitz, 1989, 2012; Berkowitz & Harmon-Jones, 2004; Dollard, Miller, Doob, Mowrer, & Sears, 1939). This hypothesis is interesting because rewarding stimuli are likely to be co-extensive with the fitness-relevant resources governed by ownership conventions, and some punishments (i.e., stimuli marked with a negative valence) are likely to be co-extensive with violation of ownership conventions. As an example of the latter kind, to a resident rat, the smell of an unfamiliar male rat in its territory may be a negatively valenced stimulus precisely because it marks the violation of an ownership

[28] Anger at loss of control develops quite early and only represents an implicit grasp of the first possession convention. However, children develop a more explicit grasp of this convention as early as 3-4 year olds, when they rely heavily on the convention to make judgments about ownership that concern third-parties (see Friedman & Neary, 2008).

convention. If all this is correct, then the frustration-aggression link may have provided a simple way of implementing the reserve strategy. [29] In other words, the adaptiveness of the reserve strategy may help to explain a wide range of frustration-aggression links. It could explain why organisms have default expectations regarding ownership (and perhaps also non-interference) and why organisms might respond to frustrated expectations with aggression. If this is correct, then what recalibrational models interpret as a reaction to a low WTR might be much better understood as a reaction to a broader range of violated expectations of reward (or non-punishment). On this latter account, one would predict that anger could be triggered by frustrated expectations *aside from* low WTRs and perhaps that low WTRs trigger anger *because* they are unexpected punishments or because they frustrate expectations (e.g., a higher WTR was expected). If so, this would mean that the machinery for tracking WTRs is inessential to angry and vengeful behaviors in many cases. Indeed, there is some evidence that anger can be triggered in human infants by frustrated expectations of reward that are not obviously connected to any form of WTR assessment (see, e.g., Michael Lewis, 1990).

This alternate explanation thus can subsume the recalibrational account (in that it can account

---

[29] The frustration-aggression link may very well have been established by a more ancient and general evolutionary problem of overcoming obstacles to one's goals. This evolutionary problem is not essentially a social one, since the obstacle to one's goal could just as easily be a rock as a conspecific. Nevertheless, revenge is a distinctively social interchange. So I would argue that a motive for aggression does not become a payback motive until it begins to be shaped for a distinctively social purpose. A related point concerns the nature of anger qua payback motive: It seems quite possible to me that anger has other functions aside from revenge. By calling anger a payback motive, I mean only that it has been shaped by selection to implement revenge, in addition to whatever function it already had or subsequently acquired.

for all of the complex payback phenomena the recalibrational model explains) while also providing a computationally simple explanation for retaliatory behaviors in developing humans and in nonhuman animals (which the recalibrational model cannot explain). Moreover, it can do this without positing much added machinery to the payback motive, instead focusing on how payback motives interact with other psychological processes.

For instance, recent models of reinforcement learning in rodents and humans suggest fairly ancient (perhaps pan-mammalian) mechanisms for generating expected outcomes, either on the basis of rich inferential models (as in the case of model-based learning mechanisms) or a range of simpler associative mechanisms (see, e.g., Balleine & O'Doherty, 2010). If we suppose that payback motives are triggered by the predictive outputs of such systems, then it is possible to give a unified explanation of several disparate phenomena. It would explain how revenge and retribution can be elicited by low-level frustrations (in accordance with work on the frustration-aggression hypothesis) and also by norm violations (some of which require the use of metarepresentational capacities in constructing the inferential models which guide the generation of social expectations).[30] If this is correct, then the key difference between payback in humans and in other animals is just in the complexity of inferential models (and thus expectations) our minds are capable of constructing to guide learning.

Similarly, the complexity of inferential models can modify the complexity of ownership conventions in humans. In some human cultures, first possession conventions give way to more complex conventions that distinguish between items found on public versus private properties

[30] One of the most prominent accounts of the nature of social norms makes central appeal to social expectations (Bicchieri, 2006).

(see, e.g., DeScioli, Karpoff, & De Freitas, 2017). If ownership conventions are implemented by learning-guided expectations, then increasing complexity of expectations can account for increasingly complex ownership conventions; ones that may also function to prevent costly fights (e.g., DeScioli & Wilson, 2011). Thus, forging connections with emerging literatures on reinforcement learning may offer fertile ground for future research on anger, payback, aggression, and even property rights.[31]

7 CONCLUSION

To sum up, the recalibrational theory proposes that payback is a complex adaptation that employs bookkeeping algorithms to deter harms to individuals. I argued that this theory cannot easily explain certain retaliatory behaviors in nonhuman animals. Such behaviors appear to be instinctual and spiteful, yet they do not obviously function to deter harms. I then offered an alternative theory on which payback motives originated for enforcing ownership conventions used to resolve resource competition. On this account, the original function of payback is not to deter harms to individuals but to prevent the spread of variant strategies within a population. This

[31] I have focused here on how payback phenomena are shaped by modifications to bookkeeping capacities that serve as inputs to payback motives. However, there is much else to explain about the outputs of payback motives, including the many ways that payback is channeled or directed. For example, the aim of revenge can be to restore "karmic balance," to adjust relative status, or even to balance one's pain with the pain of a transgressor. While these are important facts to explain, it is reasonable to leave them for another time. By comparison, were I explaining the ancient origins of hunger, it would be reasonable to leave aside the question of why hunger leads to differentiated food cravings (e.g., for double chocolate fudge ice cream or parmesan spinach gnocchis). Thanks to an anonymous referee for posing this problem.

leads to a much simpler explanation of payback in humans, which is naturally accompanied by a very different hypothesis about the relation between payback and bookkeeping. On the view proposed here, payback need not be triggered by complex mechanisms of bookkeeping, but can also be triggered by simpler indicators of violated ownership conventions. The frustration-aggression hypothesis points to a natural way of tracking these violations: violated expectations of reward and non-punishment. If this is correct, then bookkeeping capacities are not the only inputs that influence payback, because they are only a subset of the mechanisms that influence expectations concerning reward and punishment.

This work may have important implications for ongoing research in the psychology of revenge, punishment, and moral bookkeeping. Concerning revenge, an instinctual impulse for payback could help explain why we pursue revenge even though it rarely gives us the pleasure that we expect from it (see, e.g., Carlsmith et al., 2008). Given how ancient and primitive the impulse is, it may even explain why the feeling of kicking an offending door or vending machine can be so similar to the experience of avenging oneself on another person.

There is a great deal more to say about punishment and moral bookkeeping, both of which involve moral transactions that rely on keeping accounts with others. Moral bookkeeping involves holding people responsible and deciding what they deserve based on their "record," or what they have done. Much of the psychological work on deservingness does not focus on its connection with primitive motivational states like anger and vengefulness (e.g., Lerner, 2003); when it does, the dominant concern is how judgments of deservingness influence emotions rather than vice versa (e.g., Feather, 2006). Nevertheless, if payback motives preceded bookkeeping in our lineage (as I have suggested), then payback motives may provide important evolutionary and developmental constraints on bookkeeping: perhaps bookkeeping develops in the service of

payback motives (among other social motives), especially their extension over longer periods of time (e.g., nursing a grudge against a "deserving" target). For instance, some philosophers have suggested that payback motives are the basis for retributive intuitions concerning moral punishment (Greene, 2008; Nussbaum, 2016; Parfit, 2011, p. 429; Waller, 2015; Wiegman, 2014). According to these suggestions, we became the sorts of creatures who deal out moral punishments in proportion to past offenses because of deeply rooted payback motives: they move us to react aggressively to past harms rather than just aggressing to secure foreseeable benefits. This possibility has important implications for the moral justification of punishment and legal policy. Based on its etiology, we may decide that retribution is an inadequate rationale for punishment (see e.g., Greene, 2008; Nussbaum, 2016; Waller, 2015; Wiegman, 2014).

Possible links between payback motives and moral bookkeeping should also inform attempts to understand the mechanisms of moral punishment. One should not be surprised if genetic or neural predictors of payback are also predictors of moral punishment decisions (e.g. Strobel et al., 2011). Moreover, it may be fruitful to test for other similarities between moral punishment and payback. For example, aggression and revenge seem to depend in large part on people's expectations in a given situation. I am vengeful toward the negligent person who bumped into me because I expected them to watch where they were going. Perhaps moral punishment similarly depends on which expectations are in play, as determined by which norms are salient when the opportunity for punishment arises. Take for example a case where someone's outfit is ruined when a texting pedestrian collides with them. On this hypothesis, punitive judgments would be much harsher if there is a norm against walking and texting and perhaps less harsh when the pedestrian is a tourist, since out-groups are not always expected to comply with local norms (e.g., Schmidt, Rakoczy, & Tomasello, 2012).

Finally, if payback motives contribute to the evolution or development of bookkeeping, then perhaps other evolved motivational states do as well: gratitude, disappointment, shame, and guilt. It may even be that the desire for payback and other ancient social motives partly constitute the cultural practices and psychological capacities of bookkeeping (cf. Strawson, 1963). Properly explored in psychological research, this hunch could open a revealing window onto the furnishings of the moral mind.

A broader, meta-scientific conclusion can also be reached from this work: If evolutionary psychology begins theorizing with psychologically modern humans in mind, cut loose from our more ancient historical moorings, the hypothesis space is too limited. The explanations that seem most plausible may involve unnecessarily complex psychological processes. By contrast, if theory begins with the set of cognitive abilities, behavioral patterns, and selection pressures in our more ancient ancestors, then a range of alternative hypotheses arise. In this case, the alternative is that complex phenomena of bookkeeping are built up from much simpler processes and motivations.

WORKS CITED

Adams, D. B. (1980). Motivational systems of agonistic behavior in muroid rodents: A comparative review and neural model. *Aggressive Behavior*, *6*.

Alessandri, S. M., Sullivan, M. W., & Lewis, M. (1990). Violation of expectancy and frustration in early infancy. *Developmental Psychology*, *26*(5), 738–744. https://doi.org/10.1037//0012-1649.26.5.738

Archer, J., & Huntingford, F. (1994). Game theory models and escalation of animal fights. *... and Social Processes in Dyads and ….*

Axelrod, R. M. (1984). *The Evolution of Cooperation*. Basic Books.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Barash, D. P., & Lipton, J. E. (2011). *Payback: Why we retaliate, redirect aggression, and take revenge*. Oxford: Oxford University Press.

Berkowitz, L. (1989). Frustration-aggression hypothesis: examination and reformulation. *Psychological Bulletin*, *106*(1), 59–73.

Berkowitz, L. (2012). A Different View of Anger: The Cognitive-Neoassociation Conception of the Relation of Anger to Aggression. *Aggressive Behavior*, *38*(4), 322–333. https://doi.org/10.1002/ab.21432

Berkowitz, L., & Harmon-Jones, E. (2004). Toward an understanding of the determinants of anger. *Emotion (Washington, D.C.)*, *4*(2), 107–130. https://doi.org/10.1037/1528-3542.4.2.107

Bicchieri, C. (2006). *The grammar of society: the nature and origins of social norms*. New York: Cambridge University Press.

Bishop, D. T., & Cannings, C. (1978). A generalized war of attrition. *Journal of Theoretical Biology*, *70*(1), 85–124. https://doi.org/10.1016/0022-5193(78)90304-1

Blanchard, D. C., & Blanchard, R. J. (1984). Affect and aggression: An animal model applied to human behavior. In R. J. Blanchard & D. C. Blanchard (Eds.), *Advances in the Study of Aggression* (Vol. 1, pp. 1–62).

Blanchard, D. C., Litvin, Y., Pentkowski, N. S., & Blanchard, R. J. (2009). Defense and Aggression. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences* (pp. 958–974). Hoboken: Wiley.

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316–1324. https://doi.org/10.1037/a0012165

Cikara, M. (2018). Pleasure in response to out-group pain as a motivator of intergroup aggression. In K. Gray & J. Graham (Eds.), *Atlas of Moral Psychology*. Guilford Press.

Cikara, M., & Fiske, S. T. (2013). Their pain, our pleasure: stereotype content and schadenfreude. *Annals of the New York Academy of Sciences*, *1299*, 52–59. https://doi.org/10.1111/nyas.12179

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(19), 209–216.

Cosmides, L., & Tooby, J. (2000). Evolutionary Psychology and the Emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (2nd ed.).

Daly, M., & Wilson, M. (1988). *Homicide*. Transaction Publishers.

Dawkins, R. (2006). *The selfish gene*.

DeScioli, P., & Wilson, B. J. (2011). The territorial foundations of human property. *Evolution and Human Behavior*, *32*, 297–304. https://doi.org/10.1016/j.evolhumbehav.2010.10.003

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., & Sears, R. R. (1939). *Frustration and aggression.* New Haven, CT, US: Yale University Press. https://doi.org/10.1037/10022-000

Dugatkin, L. A. (1997). Winner and loser effects and the structure of dominance hierarchies. *Behavioral Ecology*, *8*(6), 583–587. https://doi.org/10.1093/beheco/8.6.583

Dugatkin, L. A., Druen, M., Dugatkin, L. A., & Druen, M. (2004). The social implications of

winner and loser effects The social implications of winner and loser effects, (December). https://doi.org/10.1098/rsbl.2004.0235

Enquist, M., & Leimar, O. (1983). Evolution of Fighting Behaviour: Decision Rules and Assessment of Relative Strength. *J. Theor. Biol*, *102*, 387–410.

Feather, N. T. (2006). *Deservingness and emotions: Applying the structural model of deservingness to the analysis of affective reactions to outcomes*. *European Review of Social Psychology* (Vol. 17). https://doi.org/10.1080/10463280600662321

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions.* New York: Norton. https://doi.org/10.2307/2072516

Friedman, O., & Neary, K. R. (2008). Determining who owns what: Do children infer ownership from first possession? *Cognition*, *107*(3), 829–849. https://doi.org/10.1016/J.COGNITION.2007.12.002

Frijda, N. H. (1994). The lex talionis: On vengeance. In S. H. M. Van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 263–289). Mahway: Lawrence Erlbaum Associates.

Gintis, H. (2006). The Evolution of Private Property, *7756*, 1–22.

Gopnik, A., Glymour, C., & Sobel, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological …*.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3, The Neuroscience of Morality: Emotion, Disease, and Development* (pp. 35–80). Cambridge: MIT Press.

Haccou, P., & Glaizot, O. (2002). The ESS in an asymmetric generalized war of attrition with mistakes in role perception. *Journal of Theoretical Biology*, *214*(3), 329–349. https://doi.org/10.1006/jtbi.2001.2454

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology*, *96*(4), 647–682. https://doi.org/10.1016/0022-5193(82)90235-1

Hsu, Y., Earley, R. L., & Wolf, L. L. (2006). Modulation of aggressive behaviour by fighting experience: mechanisms and contest outcomes. *Biological Reviews of the Cambridge Philosophical Society*, *81*(1), 33–74. https://doi.org/10.1017/S146479310500686X

Hubbard, J. A., Romano, L. J., McAuliffe, M. D., & Morrow, M. T. (2010). Anger and the Reactive–Proactive Aggression Distinction in Childhood and Adolescence. In M. Potegal, G. Stemmler, & C. D. Spielberger (Eds.), *International Handbook of Anger* (pp. 231–239). Springer New York.

Huntingford, F. A., & Turner, A. K. (1987). *Animal Conflict*. London: Chapman and Hall.

Huttegger, S. (2010). Generic properties of evolutionary games and adaptationism. *Journal of Philosophy*, 1–31.

Huttegger, S. M., & Zollman, K. J. S. (2012). Evolution, dynamics and rationality: the limits of ESS methodology. In S. Okasha & K. Binmore (Eds.), *Evolution and Rationality: Decisions, Cooperation and Strategic Behaviour* (pp. 67–83). Cambridge University Press.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical*

Transactions of the Royal Society of London. Series B, Biological Sciences, *365*, 2635–2650. https://doi.org/10.1098/rstb.2010.0146

Krebs, J. R., & Davies, N. B. (1993). *An Introduction to Behavioural Ecology* (3rd ed.). Oxford: Blackwell.

Lambert, A. J., Peak, S. A., Eadeh, F. R., & Schott, J. P. (2014). How do you feel now ? On the perceptual distortion of extremely recent changes in anger. https://doi.org/10.1016/j.jesp.2014.01.004

Leach, C. W., & Spears, R. (2008). "A Vengefulness of the Impotent": The Pain of In-Group Inferiority and Schadenfreude Toward Successful Out-Groups. *Journal of Personality and Social Psychology*, *95*(6), 1383–1396. https://doi.org/10.1037/a0012629

Lehner, S. R., Rutte, C., & Taborsky, M. (2011). Rats Benefit from Winner and Loser Effects. *Ethology*, *117*(11), 949–960. https://doi.org/10.1111/j.1439-0310.2011.01962.x

Lerner, M. J. (2003). The justice motive: where social psychologists found it, how they lost it, and why they may not find it again. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *7*(4), 389–399.

Lewis, Michael. (1990). Violation of expectancy, loss of control, and anger expressions in young infants. *Developmental Psychology*, *26*(5), 745–751.

Leyhausen, P. (1979). *Cat Behavior: The Predatory and Social Behavior of Domestic and Wild Cats*. (B. A. Tonkin, Trans.) (1St Editio). Taylor & Francis / Garland STPM Press.

Lloyd, E. A. (2007). Units and Levels of Selection. In D. L. Hull & M. Ruse (Eds.), *The Cambridge Companion to the Philosophy of Biology*. New York: Cambridge University Press.

Maynard Smith, J. (1974). The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*.

Maynard Smith, J., & Harper, D. (2003). *Animal signals*. Oxford University Press, USA.

Maynard Smith, J., & Parker, G. A. (1976). The logic of asymmetric contests. *Animal Behaviour*.

Maynard Smith, J., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2012). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

Mesterton-Gibbons, M. (1999). On the evolution of pure winner and loser effects: a game-theoretic model. *Bulletin of Mathematical Biology*, *61*(6), 1151–1186.

Morris, A., Macglashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1704032114

Nakao, H., & Machery, E. (2012). The evolution of punishment. *Biology & Philosophy*, *27*(6), 833–850. https://doi.org/10.1007/s10539-012-9341-3

Nisbett, R. E., & Cohen, D. (1996). *Culture of Honor: The Psychology of Violence in the South*.

Boulder, CO: Westview Press.

Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*.

Nussbaum, M. C. (2016). *Anger and Forgiveness: Resentment, Generosity, Justice*. New York: Oxford University Press.

Parfit, D. (2011). *On What Matters, vol. 2*. Oxford: Oxford University Press.

Parker, G. A. (1974). Assessment strategy and the evolution of fighting behaviour. *Journal of Theoretical Biology*, *47*(1), 223–243.

Parker, G. A., & Rubenstein, D. I. (1981). Role Assessment, Reserve Strategy, and Acquisition of Information in Asymmetric Animal Conflicts. *Animal Behaviour*, *29*, 221–240.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary Psychology and Criminal Justice : A Recalibrational Theory of Punishment and Reconciliation. In *Human Morality and Sociality*.

Phillips, W., Barnes, J. L., Mahajan, N., Yamaguchi, M., & Santos, L. R. (2009). "Unwilling" versus "unable": capuchin monkeys' (Cebus apella) understanding of human intentional action. *Developmental Science*, *12*(6), 938–945.

Potegal, M., & TenBrink, L. (1984). Behavior of attack-primed and attack-satiated female golden hamsters (Mesocricetus auratus). *Journal of Comparative Psychology*, *98*(1), 66–75. https://doi.org/10.1037//0735-7036.98.1.66

Schaller, G. B. (1976). *The Serengeti lion: A study of predator-prey relations*. Chicago: The University of Chicago Press. https://doi.org/10.2307/1296618

Schenkel, R. (1967). Submission: Its features and function in the wolf and dog. *Integrative and Comparative Biology*, *7*(2), 319–329. https://doi.org/10.1093/icb/7.2.319

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young Children Enforce Social Norms Selectively Depending on the Violator's Group Affiliation. *Cognition*, *124*(3), 325–333.

Sell, A. (2005). *Applying Adaptationism to Human Anger: The Recalibrational Theory*.

Sell, A. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior*, *16*(5), 381–389. https://doi.org/10.1016/j.avb.2011.04.013

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.

Sripada, C. S., & Stich, S. (2007). A Framework for the Psychology of Norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind: Culture and Cognition* (Vol. 2). https://doi.org/10.1093/acprof

Stamps, J. A., & Tollestrup, K. (1984). Prospective Resource Defense in a Territorial Species. *The American Naturalist*, *123*(1), 99–114.

Strawson, P. F. (1963). Freedom and resentment. In *Perspectives on Moral Responsibility* (pp. 67–100).

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage*, *54*(1), 671–680. https://doi.org/10.1016/j.neuroimage.2010.07.051

Theraulaz, G., Bonabeau, E., & Deneubourg, J.-L. (1999). The mechanisms and rules of coordinated building in social insects. In *Information Processing in Social Insects* (pp. 309–330). Basel: Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-8739-7_17

Tooby, J., & Cosmides, L. (1990). The past explains the present. *Ethology and Sociobiology*, *11*(4–5), 375–424. https://doi.org/10.1016/0162-3095(90)90017-Z

Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In Michael Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (3rd ed., pp. 114–137). Guilford Press.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*.

van de Poll, N. E., Smeets, J., van Oyen, H. G., & van der Zwan, S. M. (1982). Behavioral consequences of agonistic experience in rats: sex differences and the effects of testosterone. *Journal of Comparative and Physiological Psychology*, *96*(6), 893–903.

Vitiello, B., & Stoff, D. M. (1997). Subtypes of aggression and their relevance to child psychiatry. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(3), 307–315. https://doi.org/10.1097/00004583-199703000-00008

Waller, B. (2015). *The stubborn system of moral responsibility*.

Wiegman, I. (2014). *Anger and Punishment: Natural History and Normative Significance*. Washington University in St. Louis.

Wingfield, J., Ball, G., Dufty, A., & Hegner, R. (1987). Testosterone and aggression in birds. *American Scientist*.