The Root of Retribution: Coercion or Competition?

1 INTRODUCTION

Everyone has experienced injustices and slights, and so approximately everyone has experienced sentiments like anger, resentment, and vengefulness. These are not just feelings. They are also motives; I call them retributive motives. In their grip, one is motivated to confront, retaliate, avenge, or seek retribution, often contrary to better judgment. These motives can be contrasted with prospective or proactive motives for aggressive or antagonistic behavior. For instance, proactive aggression occurs when someone harms another to achieve some direct benefit (whether material or hedonic), as when assault is committed only for the purpose of robbery.¹

By contrast, when aggression is motivated purely by anger or vengefulness, the aim of the action (as understood by the agent) cannot be described in terms of future benefits. When someone plots revenge, revenge may be the only aim. The avenger may not even care whether revenge is satisfying, pleasurable or desirable in any other way aside from its apparent fittingness as a response to transgression. This is a puzzling feature of retributive motives. Unless these impulses are actively resisted or overturned by better judgment, they impel us to act without full consideration of other costs and benefits.

Why is this so? Many suspect that biological evolution offers the best explanation for retributive motives in humans (e.g. Daly and Wilson 1988; Frank 1988; M. E. McCullough, Kurzban, and Tabak 2012). On many of these accounts, retribution is an adaptation for implementing punishment strategies that are advantageous because, among other things, they coerce audiences of punishment to modify their behavior (usually via reputation). In most cases, it is unclear whether these explanations are meant to be explanations of the origins of retributive motives or merely explanations of its cooption or maintenance. I argue that this adaptationist story cannot plausibly explain the origins of retributive motives, and I offer a new account of the origins of retribution.²

¹ In psychological literatures on aggression, the central contrast is between reactive and proactive (also known as predatory, and instrumental) subtypes of aggression (for reviews, see Vitiello and Stoff 1997; Poulin and Boivin 2000; Hubbard et al. 2010). Reactive aggression is affectively valenced, whereas proactive and predatory forms of aggression (even in other species) are often unaccompanied by the kinds of physiological changes that accompany anger and fear.

² For an argument against a broader range of behavior-modification style explanations of punishment, see Nakao and Machery (2012). See also Elster (1999), pp. 48-50 for an argument against the use of evolutionary models to explain the structure of emotions.
In the following section (section 2), I point out that these adaptationist explanations of retribution must assume that it arose in organisms with capacities for foresight and anticipation, what I call *prospection*. This assumption takes one of two forms, either a) these explanations require *complex prospective capacities* on the part of the punishee or b) they apply only to populations that possess these capacities. There are two key problems for any explanation of retributive motives that employs these assumptions. First, to implement punishment that coerces audiences, retributive motives need to interact with prospective capacities in complex ways. Moreover, the complexity of this adaptation makes it implausible that it first emerged in creatures with prospection. To explain the origins of retribution in this way, one must posit discontinuous evolutionary leaps. Instead, we should prefer gradualistic explanations. According to these explanations, retributive motives gradually entered into more complex interactions with prospection, emerging prior to prospective capacities or coevolving with them. A second problem concerns explanations of retribution that date its origins in distinctively human lineages. On such an account, one cannot provide a unified explanation of retribution-like behaviors in human and non-human animals. By contrast, we should favor theories that can explain retribution in a wider range of lineages, prior to the emergence of distinctively human lineages and prior to (or at least independent from) the emergence of complex prospective capacities.

In section 3, I describe a game theoretic model of resource competition that applies to animals outside the human lineage and also to organisms without complex prospection. I then demonstrate the importance of retributive motives (of some kind) for implementing stable strategies on this model. In section 4, I address potential criticisms and argue that my proposal identifies selection pressures capable of explaining the gradual coevolution of retributive motives alongside prospective capacities in primate and human lineages. I conclude by drawing a lesson from this case: there are clear benefits for understanding the primitive conative adaptations that humans share with a broader range of lineages.

### 2 Retribution as an Adaptation for General Deterrence

There are two families of evolutionary models that explain retribution as an adaptation for coercing audiences of punishment: reputation models and cultural group selection models. The phenomenon on which most of these models focus is actually human cooperation. Nevertheless, on these models, the threat of retribution is necessary to sustain certain forms of cooperation or cultural cohesion. Thus, these models can explain why retributive motives

---

3 While these models are likely to have implications for the evolution of morality, their central concern is not necessarily *moral* punishment. Rather, these models can also explain non-moral punishment, as when someone seeks revenge or retribution for non-moral offenses (e.g. breaches of etiquette) or for actions that they deem to have been morally justified (e.g. taking revenge against someone who killed in self-defense).

Here, I understand punishment as a behavioral strategy by which one entity (e.g. organism or group of organisms) diminishes the fitness of another entity in response to (or proximally caused by) a provoking
would be adaptive for navigating certain social interactions. In this section, I raise the question of whether the adaptiveness of retribution on these models (commitment and cultural group selection models) can explain the origins of retribution (as opposed to its cooption or maintenance).

First, a clarification: the explanations I am interested in are not explanations of punishment per se but rather explanations of retributive motives. Many punishment strategies can be implemented without retributive motives, as when someone prefers larger payouts and believes that punishment will result in better pay outs with minimal immediate costs. By contrast, retributive motives only seem adaptive for implementing punishment strategies that are temporarily spiteful (as I explain in greater detail in section 2.2). Temporarily spiteful punishment is punishment that is costly and the benefits of which are deferred or intangible.

What kind of deferred benefits could make it advantageous to punish under these circumstances? General deterrence is a clear case. Punishments that aim at specific deterrence prevent further harm from specific individuals either by execution, restraint, ostracism or coercion (as when a specific offender is subjected to some form of hard treatment to induce future cooperation). Nevertheless, greater benefits can sometimes be achieved when punishment deters other individuals aside from the immediate target of punishment. General deterrence becomes possible when punishment also coerces the audience of punishment to modify their behavior (whether or not that is the explicit intention of the punisher). For example, avenging (and thereby punishing) personal or familial offenses is sometimes beneficial because it deters future offenses, not only from the current offender but from potential offenders who see or hear about the act of revenge. However, because revenge is costly and
because individuals will not always know about its benefits, the desire for general deterrence (in addition to beliefs about its efficacy) will not necessarily motivate revenge in all the circumstances in which it would be advantageous. Thus, retributive motives may be adaptive for motivating punishment when agents are not otherwise reliably motivated by (or apprised of) its benefits, especially benefits that accrue to general deterrence.

This style of explanations for the adaptiveness of retribution usually involves one of two assumptions: a) they require complex prospective capacities on the part of the punishee or b) they apply only to populations that possess these capacities. This is because punishment achieves general deterrence via the audience’s ability to learn from punishment. The idea is that the audience can acquire information from the punisher, from the act of punishment or from its consequences and use that information to infer how the punisher is likely to behave in various other circumstances. Agents in the audience can then condition their behavior accordingly. For example, self-interested agents with prospective capacities can learn from observing the punishment of others that certain selfish behaviors lead to punishment at least when a known punisher is present. As a result, the likelihood of punishment can offset the material benefits of selfish behavior in the relevant situations and in a way that coerces self-interested agents to modify their behavior.

In the human lineage (and perhaps in a wider range of lineages), prospective processes are among the proximate explanations for this ability. As I understand them, prospective processes select actions based on the internally represented outcomes of a range of alternative actions. Moreover, complex prospective capacities calculate the expected value of actions with reference to a learned causal model (e.g. Gläscher et al. 2010). In humans, causal models allow the postulation of hidden causes of observable behavior in a wide range of domains. As a result, they probably have some integration with abilities to attribute traits, beliefs and desires as hidden causes of human behavior (cf. Gopnik and Wellman 2012; Schaafsma et al. 2014).

enforcement cannot effectively enforce ownership norms or deter certain offenses (Nisbett and Cohen 1996; Sommers 2008).

6 For many contemporary evolutionary psychologists, these assumptions are particularly central. Among evolutionary psychologist, it has become commonplace to argue that evolutionary models require complex, uniquely human adaptations when applied to the human lineage (e.g. DeVore and Tooby 1987). Underlying this trend is an emphasis on the extraordinary power of selection to produce complex adaptations in relatively short periods of time. For instance, McCullough (2008) makes this kind of argument specifically with regard to punishment as a rationale for ignoring phylogenetic constraints on the evolution of punishment. Evolutionary psychologists in the lineage of Cosmides and Tooby often refer to the work of Williams (1966) in this context.
2.1 Reputation Models and Complex Prospection

On reputation models, this kind of integration is necessary for prospection to be a mechanism for audience coercion.\(^7\) Reputation models, in particular, require that agents be able to learn causal relationships to anticipate the outcome or expected utility of their actions.\(^8\) The commitment model is a paradigm case (Frank 1988), and the relevant features of this model are shared by several explicit attempts to explain revenge and punishment in humans (e.g. M. E. McCullough, Kurzban, and Tabak 2012; Petersen et al. 2012). These models attempt to explain how humans are able to solve commitment problems like the following:

**Deterrence.** Suppose Smith grows wheat and Jones raises cattle on adjacent plots of land. Jones is liable for whatever damage his steers do to Smith’s wheat. He can prevent damage altogether by fencing his land, which would cost him $200. If he leaves his land unfenced, his steers will eat $1000 worth of wheat. Jones knows, however, that if his steers do eat Smith’s wheat, it will cost Smith $2000 to take him to court…Smith threatens to sue Jones for damages if he does not fence his land. But if Jones believed Smith to be a rational, self-interested person, this threat is not credible. Once the wheat has been eaten, there is no longer any use for Smith to go to court. He would lose more than he recovered. (Frank 1988, 48)

The problem concerns how Smith can make a credible threat, when Jones knows that it is not in Smith’s immediate interest to go to court. In a case like this, a retributive motive can serve as a commitment device\(^9\), meaning that it would motivate Smith to follow through with his threat and thereby punish Jones, even if it is not in his immediate interest to do so.\(^10\) Smith is likely to

---

7 I am certainly not saying that the role of reputation is essential or ineliminable for explanations of punishment more generally. However, it may be essential or ineliminable for explanations of retribution, and particularly ones that are used to explain retribution in distinctively human populations.

8 Importantly, successful punishment strategies in repeated interactions (such as models of direct reciprocal altruism, e.g. Axelrod 1984; Trivers 1971) often do not require behavior modification (much less prospection), because they involve forms of punishment either that benefit the punisher by cutting their losses or that affect interaction partners without modifying their behavior. For example, an organism can punish another organism by killing it. For a range of biological examples of this form of punishment, see Nakao and Machery (2012). In other words, punishment is not an adaptation for behavior modification on these models, nor does punishment tend to be temporarily spiteful. Thus, the problem I will pose does not apply to many of the models within this broader family (though see section 3.1 for more discussion of these and other models).

9 More specifically, it is a subjective commitment device (Nesse 2001), contrasted with objective or external commitment devices, which might offer, say, material incentives for following through with a threat. For instance, Smith might put the money necessary to sue Jones in an escrow account for his lawyer, the reimbursement of which is contingent on Jones building a fence. In this way, Jones would “tie himself to the mast” with an external commitment device.

10 Interestingly, Sterelny (2012, 110) suggests that some commitment devices are Krebs-Dawkins signals, meaning that their role is not (or not only) informational but motivational. While Sterelny does not
benefit long-term in one of two ways from possessing this commitment device. One possibility is that Smith will retaliate against Jones’s noncompliance and thereby foster a beneficial reputation for following through with his threats. For instance, if Smith’s neighbors Amjad, Baggi, and Castro each subsequently decide to raise cattle on their property, then Smith’s threats will be more credible to them than they were to Jones, providing a net benefit to Smith if they consequently comply with his wishes. The other possibility is that Jones complies because, as a result of Smith’s disposition toward revenge (or some reliable signal of this disposition), he already has such a reputation. Either way, Smith is likely to get reputational benefits from a retributive motive. Moreover, it is easy to see how these benefits could translate to fitness benefits.

This kind of explanation of retributive motivation requires complex prospection because the benefits of punishment only obtain if the audiences of punishments are influenced by the reputation of the punisher and if they can use this information to condition some of their behaviors to the expected behaviors of a specific individual. On this model, the reputational...

发展理念 this proposal with respect to punishment (he focuses instead on trust and hunting), one could imagine at least two ways in which expressions of anger at bullies (or defectors or cheats) could function as Krebs-Dawkins signals. First, they could make failure to punish more painful or unsatisfying for the signaler and thus increase her motivation to punish bullies (et al). Second, they could (also) make the audience of the signal more likely to punish the failure to punish on the part of the signaler. Both possibilities raise the cost of failing to punish by modulating motivational states. This is a very interesting proposal concerning the origin of retributive motives. However, there are several reasons to leave this proposal to the side. First, this signaling system does not obviously require (as do the models I criticize) that punishees modify their behavior as mediated by the reputation of the punisher. Second, this kind of signaling system may also require a more ancient retributive motive to get off the ground. Third, Sterelny explicitly avoids commitment to the claim that many emotional states (some of which support cooperation) are uniquely human, whereas the targets of my argument do not.

11 Frank (1988, ch.5) also goes to great pains to show how reliable signals of moral sentiments like outrage might be achieved, so that those who possess them rarely violate their self-interest despite having a tendency to do so.

12 One might expect that the reputational benefits of punishment would be factored into the decision to punish and that this would make the decision purely prospective in the sense of being guided by internally represented future benefits. However, punishment motives need not be prospective in order to be sensitive to factors that are strategically relevant to the effectiveness of the commitment device. For instance, audience effects, whereby a behavior is mediated by the presence of other individuals, typically can be explained by the eliciting conditions of the behavior rather than by an anticipation of the effects on the audience. Many non-human species display audience effects (e.g. Marler, Dufty, and Pickert 1986) but nonetheless lack the ability to represent the mental states of other organisms. Moreover, there is reason to suspect that some human emotions are unconsciously influenced by the presence of an audience (Fridlund 1991). Thus, there is good reason to believe that reputational effects can be achieved without the significant computational costs of representing their effects in prospective deliberation. Thus, there is good reason to suppose that punishment motives will tend to be sensitive to strategically relevant factors without any internal representation of the strategic effects of those factors (Griffiths 2010; Griffiths 2004).
benefits have their effects by entering into the strategic considerations of those deciding whether to comply with Smith’s demands when threatened.\(^\text{13}\)

Of course, there may be nothing distinctively human about complex prospection or making threats or tracking reputation.\(^\text{14}\) Nevertheless, this depends on how reputation is understood. If it is understood as something that is conveyed through language and that tracks beliefs, desires and traits of other organisms, then reputation tracking is likely to be distinctively human. In any case, one might be tempted to explain the origins of retribution as a response to the adaptive problems posed by distinctively human reputation tracking (Sell 2005a, 21).

2.2 Cultural Group Selection and Complex Prospection

Prominent cultural group selection (CGS) models also seem to depend upon reputation to some degree or another. These models aim to explain the evolution of human cooperation by appealing to the benefits of punishment at the level of cultural groups (e.g. Boyd, Gintis, and Bowles 2010; Bowles and Gintis 2011). These models focus on large, cultural groups and demonstrate that a strategy of retributive punishment can protect cooperation in large groups or prevent group extinction and thus enable cooperation to persist and spread as a result of CGS. On these models, retributive punishment ends up being adaptive because it correlates with group cooperation, which can yield higher fitness benefits to punishers than the relative within-group costs of punishing. The latter would be very low when most people cooperate or when there are low cost forms of punishment such as social gossip and ostracism.\(^\text{15}\)

Many such models depend on prospection because they evaluate the conditions in which purely self-interested agents will cooperate due to the likelihood of punishment. Moreover, on some of these models would-be defectors calculate the likelihood of punishment by the presence (and quantity) of known punishers in a group and not necessarily the number of occasions on which they themselves have been punished in the past. Thus, there is an implicit (and sometimes explicit) assumption that individuals are able to learn about the likelihood of punishment

---

\(^\text{13}\) Reputation models such as these have obvious relevance for understanding cultures of honor (Nisbett and Cohen 1996) in which retaliation is often carried out with the explicit intention of defending one’s reputation. I do not want to detract from these explanatory benefits. I only want to point out that these explanations do not entirely explain the origins of the human motive for retribution.

\(^\text{14}\) Even rats might be able to use causal models to anticipate the outcome of an action (cf. Gläscher et al. 2010). If so, then the problems I raise in section 2.3 would also apply to explaining the origins of retribution in rats (if there is such a thing). Even so, the problem is particularly evident for distinctively human explanations for the origins of retribution, since it is beyond question that humans do have complex prospective abilities that can be applied to almost any decision that humans are capable of making, including decisions about punishment.

\(^\text{15}\) These models cohere well with models of CGS that have been so influential in thinking about cultural evolution. They may help explain how CGS originated (e.g. Richerson and Boyd 2005). Thus it may help explain the human capacity to adapt to the diverse environments in which members of our species have created niches. As I understand it, my argument does not detract from these explanatory benefits.
through other means, like the reputation of punishers or the frequency of punishment in the group. Thus on some of these models, it is necessary for would-be defectors to represent causal relations between reputation or punishment frequency and the likelihood of punishment given defection.

Of course, not all of the models in this family depend on prospection in this way. For instance, Boyd and Richerson (1992) analyze how different punishment strategies fare in competition with a “reluctant cooperator” strategy that defects until punished, then cooperates thereafter. This strategy clearly can be implemented without prospection. Regardless, the explicit target of these models are groups of hunter-gatherers in the Pleistocene, well after the human and chimpanzee lineages had diverged and during which tool use became more common (e.g. Bowles, Choi, and Hopfensitz 2003). Given that crafting tools requires prospection (cf. Bermúdez 2003, 127), these populations already possessed substantial prospective capacities. Moreover, these models aim to explain a kind of extreme cooperation that appears to be distinctively human. Thus, it is likely that all the organisms to whom these models apply had complex prospective capacities that influenced decisions about punishment, cooperation and defection. Insofar as these models are deployed to explain the origins of retribution, they are only likely to explain its origins in distinctively human lineages.

2.3 Temporarily Spiteful Punishment and Complex Retributive Motives

While many of the models above require complex prospective capacities in punishees (or only apply to organism that have them), they indirectly constrain the use of prospective capacities in punishers. In reputation and CGS models, punishment has a cost and the benefits of punishment, if any, are deferred or intangible. This means that punishment in these models is often temporarily spiteful, because it imposes a cost on the punishee at an immediate cost to the punisher.\(^{16}\) Given that prospective capacities benefit organisms in a very broad range of contexts, it is plausible to assume that prospective capacities will be applied to many different kinds of decisions, including the decision to punish. Insofar as organisms do apply their prospective capacities to decisions about punishment, they will be less likely to punish when punishment is temporarily spiteful. As a result, prospective processes will tend to weigh against temporarily spiteful punishment because of its immediate costs and lack of immediate material benefits for the punisher. Nevertheless, insofar as it retains its adaptive value for modifying the behavior of audiences of punishment, the inclination to punish needs to persist despite the

\(^{16}\) As far as I can see, the notion of spite at play here can encompass both Hamilton’s and Trivers’ notions of spite. Whether punishment results in temporary reductions in the direct or inclusive fitness of an organism seems inconsequential to my argument. There is room for punishment to be adaptive for behavior modification insofar as immediate fitness losses, whether direct or inclusive, are recouped later. Since it is not immediately obvious how inclusive fitness benefits could be deferred in this way, consider an example. We can imagine that punishment at a cost benefits my kin (but not me) through deterrence. If so, then the inclusive fitness benefits are not immediate, since the deterrent effect prevents future harms to my inclusive fitness by preventing damage to my kin. Thanks to [revoked] for urging this clarification.
Immediate losses that attend it. Thus, to implement a punishment strategy that requires temporarily spiteful punishment (such as the strategies which are dominant in the models above), the influence of prospective capacities has to be mitigated in the domain of punishment.

This is where retributive motives are thought to be adaptive (cf. Frank 1988): these motives move organisms to punish despite the immediate costs of doing so. As far as I can tell, there are two ways that these retributive motives could function. They either function as incentives, influencing the operations of prospective capacities, or they function as impulses, overriding prospective capacities. As I argue below, both of these possibilities requires a retributive motive with considerable cognitive complexity.

First, consider a possible way in which prospective capacities could implement temporarily spiteful punishment: via long-distance prospection or extreme foresight. For instance, one might weigh the immediate costs of punishment against its far off future consequences, say deterrence, and decide in favor of punishment. Nevertheless, well-established, cross-species work on temporal discounting shows that the prospective capacities of almost any species place exponentially greater weight on immediate rewards than more distant rewards (e.g. Critchfield and Kollins 2001). For example, the far off, intangible consequence of deterrence will as a matter of psychological fact, usually have much less value in the calculation of expected utility than the immediate costs of punishing. Moreover, the dominant strategies in the relevant models punish in precisely this kind of situation, in which its benefits are far off or intangible. For instance, in CGS models, the benefits that make punishment adaptive are benefits of living in a large cooperative group (as opposed to one that dissolves when “the going gets tough”). While these benefits may be large, they are not observable by the punisher or at least, they need not be for temporarily spiteful punishment to evolve. Thus, long-range prospection as it is implemented in most animals could not plausibly function as the proximate motive for temporarily spiteful punishment that these models aim to explain.

So, perhaps the only way punishment could be motivated by prospective processes is if there is an immediate nonmaterial incentive to react to provocation. For instance, if someone has a preference that non-cooperators be penalized for their non-cooperation (or that personal offenses be repaid in kind), then the satisfaction of this preference can serve as an immediate nonmaterial incentive that would favor temporarily spiteful punishment. Call this kind of

---

17 It is important to notice here that prospection is not the same as foresight. Calculating the immediate benefits of an action by itself can be a prospective enterprise. For instance, one could use a causal model to calculate the immediate rewards of one action compared to another without giving any thought to more temporally distant rewards. This doesn’t require a great deal of foresight, but it is a clear case of prospection on my understanding of that term.

18 For a more detailed discussion of these issues, see Frank’s discussion of the matching law (1988, ch. 4).

19 This is how Sterelny (2012, 108) understands the operation of subjective (he calls them “internal”) commitment devices. Neither he nor Frank mark the distinction between incentives and impulses.

Here it is also worth noting that in many cultures of honor nonmaterial incentives for retribution are probably transmitted epigenetically. Thanks to [revoked] for pointing this out. However, norms of revenge
retributive motive a *retributive incentive*. If retributive incentives are to motivate temporarily spiteful punishment in the relevant circumstances, then they must meet two demands: (1) They must be able to outweigh the costs of punishment in a wide enough range of cases to implement successful punishment strategies (as opposed to implementing some other punishment strategy). Moreover, (2) if optimal\(^{20}\), the non-material incentive to react to a provocation must reflect the deferred material benefits of doing so. In other words, the incentive to react to a provocation must outweigh prospective reasons not to punish and do so in the right circumstances, perhaps because it stands as a reliable proxy for some material benefit. Notice that both of these conditions require significant cognitive complexity. To satisfy the first condition, retributive incentives must interface with prospective capacities in the evaluation of the prospective value of different outcomes. To satisfy the second, an incentive must roughly track deferred benefits of punishment. That is, it must be appropriately sensitive to some of the factors that promote the deferred benefit that punishment secures such as general deterrence.

Another way to implement temporarily spiteful punishment alongside prospective capacities would be via a retributive motive that overrides prospective capacities or that retains control precedence (Frijda 1986) in the appropriate circumstances. Call this kind of retributive motive a *retributive impulse*. Unlike retributive incentives, retributive impulses override prospection rather than merely influencing it. Like retributive incentives, retributive impulses require considerable cognitive complexity to do so. In particular, they must override them to the appropriate degree and in the appropriate circumstances.

Consider: Even in a case where punishment is necessary for deterrence, the form of punishment and the timing of punishment could have important consequences. If Jones fails to put up a fence to contain his cattle and Smith’s wheat crop is destroyed, Smith may have several options concerning punishment. Taking some of Jones’s cows might be a reasonable course of action, but only if he has the necessary social support. Because the manner of punishment can be strategically complex, one suspects that prospection has a great deal of value in adjudicating the various options concerning how to punish (if not the decision about whether to punish).\(^{21}\) So retributive impulses would seem to be maladaptive if they inhibited prospection entirely.

So far, I have only considered the output side of retributive motives, specifically, how they interact with other cognitive systems. However, retributive motives also have a considerable

---

\(^{20}\) I leave aside the question of whether or not these incentives must be optimal. The first demand requires complexity enough to mount the argument I make in section 2.4.

\(^{21}\) For instance, this seems to require rational control in many cases, because it requires the imposition of subgoals. See Scarantino (2014) for an account of emotions that employs just this kind of interaction between emotion and rational control.
amount of complexity on the input side.\textsuperscript{22} That is, they must be elicited in the right conditions, and identifying those conditions can be cognitively complex. For instance, on a reputational models, these capacities should be sensitive to fairly abstract factors that impinge on fitness, such as malicious intentions.\textsuperscript{23} However, I do not consider input side complexity to be as problematic at output side complexity. This is because it seems possible to explain how a considerable degree of input sensitivity could be accumulated merely by steady addition of triggers and inhibitors. By contrast, output side complexity requires sensitivity to an indeterminate number of factors that could affect the success or failure of punishment in securing its effects (e.g. deterrence as an indirect aim or more directly giving someone what they “deserve” through retaliation or retribution). This suggests that the output of the retribution mechanism needs to look more like a desire than a fixed action pattern. Desires lead to flexible behavior by interacting with a wide range of beliefs and other informational states (e.g. whether one’s vengeful aims have social support, which actions are likely to harm a provocateur in the most visible and least costly way, etc.), whereas fixed action patterns are largely encapsulated from a wide range of informational states (and thus are insensitive to relevant factors). In any case, the main focus here is on the complexity introduced by prospective capacities of an organism, and that complexity is mainly on the output side (i.e. motivating punishment by influencing the operation of prospection).

We can conclude that temporarily spiteful punishment is a behavioral strategy that probably requires cognitively complex systems to implement it. Evolutionary models that appeal to the adaptiveness of temporarily spiteful punishment for general deterrence offer an intuitive way to explain the origins of retributive motives in humans, since these motives seem to be necessary to implement advantageous strategies.\textsuperscript{24} However, I have shown that (where the relevant models are concerned) these motives must be cognitively complex to implement punishment alongside prospective capacities (whether those capacities are required by the model or whether they merely exist in the populations to which the models apply).

2.4 TWO REASONS TO DOUBT THAT RETRIBUTION ORIGINATED AS AN ADAPTATION FOR GENERAL DETERRENCE

The assumptions that these models make about prospective capacities create two central problems for using the models to understanding the origins of retribution. First, I address a problem arising from the cognitive complexity discussed above (in section 2.3).

\textsuperscript{22} Thanks to [reference revoked] for pointing this out.

\textsuperscript{23} Intentions could be relevant in at least two ways. Either one could detect intentions to harm prior to the harm being done or one could be sensitive to whether or not a harm was inflicted intentionally. While we certainly do punish people for unintentional harms (e.g. Cushman et al. 2009), the punishment is usually less than for intentional harms. In any case, I suspect that both ways of responding to intention reflect input side sensitivities.

\textsuperscript{24} Importantly, such motives are not necessary on prominent models of direct reciprocity (and also many models of indirect reciprocity). I discuss this further in section 2.
To see this problem, consider how these models would explain the origins of retribution. A great deal of the explanatory burden rests on the assumption that there is a range of strategies concerning punishment, including a strategy that can lead to temporarily spiteful punishment. For any selection model, a certain amount of variation needs to exist for selection to operate on it. In this case, the models evaluate the success of different punishment strategies, and in doing so, they presuppose the existence of the range of strategies they evaluate. Of course, it is a well-known limitation of selection models that they assume a certain amount of variation in traits rather than explaining it. Moreover, the ability of a selection model to explain the target phenomenon depends on the plausibility of the relevant assumptions (Sober 2009). However, even if these assumptions are plausible for the selection models I discuss above, the existence of temporarily spiteful punishment as a strategy requires a good deal of explanation in its own right. This is because of the complex capacities required to implement this strategy in organisms with prospective capacities.

To reiterate, one of two forms of cognitive complexity are required: either a retributive impulse must override the prospective motivation to avoid immediate costs or retributive incentives need to outweigh the prospective costs of punishment in the right range of cases. Given their cognitively complex interactions with prospection, it is hard to see how either of these traits could arise de novo (e.g. from a single point mutation or from several simultaneous mutations) in an organism with prospective capacities. It seems unlikely that a small change in genetic material could lead to a complex tendency to punish including a cognitively complex interaction with prospective capacities. Nevertheless, the models above can only be used to explain the origins (as opposed to the mere maintenance or structure) of human retributive motives if the relevant punishment strategies arose in this way. In other words, one might be tempted (as McCullough and his colleagues (2012) seem to be) to use the models above to explain the origins of punishment in humans, but giving in to that temptation would commit one to an origin story that posits a discontinuous evolutionary leap to retributive motives that enter into complex interactions with prospective capacities.

It seems much more likely that in this case evolution had some prior material to work with from which to fashion this more complex trait. One need not look far afield to find traits in other animals that resemble retributive motives. For instance, there is a widespread tendency in the animal kingdom to react aggressively when desires are frustrated (or more accurately, when conditioned rewards are withheld, e.g. Looney and Cohen 1982). Likewise, there is a widespread tendency to respond aggressively to provocation. For instance, in many species, when an animal is attacked by a conspecific, there is a tendency to engage in counter-aggression or to redirect aggression toward another conspecific (see also Cheney and Seyfarth 1989; for an accessible review, see Barash and Lipton 2011; Kazem and Aureli 2005). Finally, aggression research in nonhuman animals suggests that some animals, rodents and cats among others, have neural systems or behavior programs that share a common evolutionary origin with anger in humans and that function in contexts of resource competition (Panksepp 2004; D. C. Blanchard and Blanchard 1984). For each of these phenomena, we see what looks like a retributive motive, namely a motive to react aggressively (and perhaps impulsively) to provocation. In some cases, this motive plausibly occurs without the complex prospective abilities that might explain the
structure of temporarily spiteful punishment in humans (e.g. as an adaptation for general deterrence).

If this is correct, then it gives us a separate reason to doubt that retribution originated as a distinctively human adaptation: this origin story cannot provide a unified explanation of the apparent retributive motives in other animals. While one might suspect that there is a unique and distinctively human tendency to punish norm violations or to punish actions that harm third parties (Bowles and Gintis 2011), even this capacity seems to be closely tied to anger (e.g. Nelissen and Zeelenberg 2009). Moreover, anger is also implicated as a proximate motive in frustration-induced aggression in humans (e.g. Berkowitz 2012) and redirected aggression (e.g. Miller et al. 2003) as well as revenge-like behaviors in the ultimatum game (e.g. Pillutla and Murnighan 1996; Sanfey et al. 2003; Srivastava and Espinoza 2009). Moreover, these phenomena closely resemble the phenomena in nonhuman animals mentioned above. If so, then distinctively human forms of punishment may depend on some of the same neural or psychological mechanisms that underpin tendencies toward reactive aggression in non-human animals (especially frustration-induced aggression and redirected aggression).

If this is correct, then there is some reason to think our ancestors already had something quite like a retributive motive for natural selection to have acted upon prior to the emergence of distinctively human lineages and of strategies for general deterrence. Yet if that is the case, then these explanations are not satisfying explanations of why humans have retributive motives in the first place. Instead they are explanations for why that trait was co-opted, and thereby maintained, to protect cooperation or to solve commitment problems (perhaps among other things). Thus, even if the models make plausible assumptions about ancestral variation in punishment strategies, they still may not offer a satisfying explanation for the existence of retributive motives (as opposed to their mere maintenance), because it seems likely that such motives already existed prior to the contexts in which these other functions became adaptive. In the following section, I describe a prominent model of resource competition that may help to explain the emergence of these more primitive forms of punishment. There are two novel features of this proposal. First, it is one of the first to suggest that this kind of resource competition involves punishment. Second, it is the first to suggest that the psychological motives that implement strategies for resource competition may be the evolutionary root that developed into punishment strategies in a wider range of social interactions (a proposal I consider in the fourth section).

---

25 In humans, this phenomenon goes by the name “displaced aggression” or “triggered displaced aggression”. Barash and Lipton (2011, chap. 3) suggest a connection between anger and redirected aggression in both humans and other animals.

26 See also Jenson (2010) for a detailed discussion of psychological motives for punishment. He suggests that anger may also involve spiteful motives that aim to reduce the welfare of others (as opposed to self-regarding motives) and that either spiteful motives or punitive sentiments may explain a range of punishment behaviors, including third-party punishment.
3  SAFER IDEALIZATIONS: MODELS OF RESOURCE COMPETITION

The kinds of model I consider in this section do not depend on the prospective capacities of organisms. Insofar as models of this kind can explain the adaptiveness of a retributive motive prior to the evolution of complex prospective capacities, they mitigate the problems posed above. Moreover, they make it possible to understand how a retributive motive could have co-evolved with prospective capacities. Thus, these models may also provide a basis for understanding how complex retributive motives might have gradually arisen.

3.1  FREQUENCY-DEPENDENT MODELS

Like some of the models mentioned above, the models in question are game theoretic models. Moreover, these models focus more exclusively on 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. One can then 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 prisoner’s dilemma, one can evaluate the average payoff of a strategy when played in populations consisting of organisms with various other strategies.

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. Thus it allows a clear way of predicting and explaining the existence of some traits. 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 and Price 1973, 15) Given this definition, we can expect selection to gradually weed out almost all other strategies from a population aside from an ESS (if indeed there is an ESS). 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.  

Given that these models do not depend on behavior modification (whether the behavior of the punisher or punishment audience), none of these models face the problem of cognitive

---

27 Evolutionary stability is not the only way to assess or predict evolutionary outcomes. For instance, there is not always an evolutionarily stable strategy for a given social interaction, and in that case, one would instead evaluate the evolutionary dynamics in order to identify equilibria that are resilient against some range of perturbations without being evolutionarily stable (Huttegger and Zollman 2010). In any case, most discussions of the models below, especially the war of attrition model, are cast in terms of evolutionary stability. My purpose here is to draw on this literature rather than to correct or revise it. In any case, I address these criticisms in greater detail below.
complexity discussed above. Moreover, frequency-dependent models already exist which are capable of demonstrating the adaptiveness of retributive motives.28

3.2 Retribution as Enforcement of Ownership Conventions

The model I have in mind was developed to explain the observed structure of animal contests for resources (e.g. food, territory, mates), especially contests in which nonlethal strategies are used. In these contests, the cost of an aggressive encounter builds up over time, and the disputed resource goes to the organism that persists the longest. Thus, game theorists call this kind of game the “war of attrition” (Maynard Smith 1974; Bishop and Cannings 1978). For instance, it explains 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 (Maynard Smith and Parker 1976; Bishop and Cannings 1978; Parker and Rubenstein 1981; Hammerstein and Parker 1982; Haccou and Glaizot 2002). As I discuss below, this is because ownership is an arbitrary asymmetry, and once introduced, ownership conventions that resolve contests on the basis of prior ownership are much more successful than strategies that ignore it. Accordingly, I propose in this section that retributive motives are an adaptation for preventing the invasion of convention-breaking mutant strategies in asymmetric, temporally extended contests for ownership of resources.

28 Aside from the model I focus on, there are a handful of frequency-dependent models capable of explaining the origins of various punishment strategies. This includes models of direct and indirect reciprocity (Axelrod 1984; Nowak and Sigmund 1998; Nowak and Sigmund 2005), partner selection models (Baumard, André, and Sperber 2013), and models of the ultimatum game (Skyrms 1996). Nevertheless, I am skeptical that these models will be helpful in identifying the origins of human retribution. In some of these models, it is unclear whether equilibrium strategies lead to temporarily spiteful punishment. For instance, in direct reciprocity, punishment is usually a loss-cutting strategy, in that it results in the immediate benefit of avoiding losses from agents with low image scores. Other models are inconsistent with a plausible timeline concerning the emergence of retributive motives. For instance, Skyrms (1996, chap. 2) shows that there are equilibrium strategies in one version of the ultimatum game that defy what he calls “modular rationality”. Some of these strategies lead to temporarily spiteful punishment. However, I have two misgivings about this model. First, all the equilibria Skyrms discusses are mixed, so it is not clear whether they can explain the spread of retributive motives in human populations. Retributive practices and the emotion of anger appear to be pan-cultural and possibly species-typical (cf. Daly and Wilson 1988; Ekman and others 1971), so I would conjecture that retributive motives are species-typical. Second, many of our primate relatives do not exhibit temporarily spiteful strategies in ultimatum game scenarios (Jensen, Call, and Tomasello 2007; Kaiser and Jensen 2012). But if, as I suggest above (in section 2.4), retributive motives occur both in primates and possibly in a wider range of mammalian lineages, then prima facie, Skyrms’ model does not offer a plausible explanation of retributive motives. His model seems to makes inaccurate predictions about the behavior of these species in ultimatum game interactions.
3.2.1 Contests of Ownership: From Symmetry to Asymmetry

To start, I briefly explain why asymmetric conventions tend to be evolutionarily stable in contests for resources. To do so, let us begin with 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. Now, 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 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 that has 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 range of persistence intervals at each encounter. More specifically, the ESS chooses a persistence interval from a specific probability distribution. The mean of the distribution is a persistence interval that accrues a cost equal to the value of the resource under dispute. In a population that consists entirely of this 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 was 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. 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, there is an asymmetry that can be used 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

---

29 This is closely related to Aumann’s notion of a correlated equilibrium. However, notice that for Aumann, correlated equilibria are driven by beliefs about the strategies that other agents will adopt. In war of attrition models, this is not so. See also Skyrms (1996, chap. 4) for a helpful discussion of correlated convention.
as a coin toss might be used. 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”. 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 success in evolutionary models 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 and 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.

3.2.2 Preventing Invasion with the Reserve Strategy
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 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 the asymmetry and fights “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. The set of bourgeois strategies that include a reserve strategy, I will call “bourgeois reserve strategies”.

---

30 Some game theorists use this term in discussions of the hawk-dove game, but it is also used in discussions of the war of attrition model (see e.g. Parker and Rubenstein 1981). Some authors attribute the label to Maynard Smith and Parker (1976), where they discuss both the war of attrition model and the hawk-dove game. However, the label does not actually appear in that paper.

31 Notice that the bourgeois convention is one of two possible strategies that resolve contests via ownership. There is also a “paradoxical strategy” in which the prior owner of the resource gives it up to an opponent. This convention is also an ESS in the asymmetric war of attrition, however, it is only rarely observed in nature, perhaps because there are usually some costs associated with ceding ownership. As Skyrms (1994) puts it, once the costs are figured in (along with correlated asymmetries) “the basin of attraction of the bourgeois equilibrium will now be larger than that of the paradoxical strategy.” (p. 78)
mutants until the cost of fighting equals the value of the disputed resource (Parker and Rubenstein 1981).

Notice that in a population of bourgeois reserve strategists, the reserve strategy 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. It follows that the motivation to play the reserve strategy cannot be prospective. If, as explained above, the stability of the bourgeois strategy depends on there being a fixed tendency to play the reserve strategy when the convention is violated, then the motivation to play the reserve strategy cannot depend on learning much less on the prospective value, or anticipated reward for playing reserve. In a population of bourgeois strategist, there would be no anticipated reward for playing the reserve, because most organisms would never have played the reserve strategy (unless by mistake) and thus would have no reason to anticipate a reward for initiating the strategy. Thus, evolutionarily stable strategies for the war of attrition require that organisms play the reserve strategy in the face of the immediate costs of doing so. Since the reserve strategy also imposes costs on the intruder in response to the intrusion, the strategy can lead to instances of temporarily spiteful punishment. It follows that it requires a retributive motive to implement it. That is, it requires a motivation to impose costs that is not conditioned on the immediate material benefits of doing so.

In fact, the very structure of frequency-dependent selection lends itself to a similar conclusion. It provides clear way of explaining why an organism could deploy the bourgeois reserve strategy without any learning from the consequences of playing the reserve. This is because the factors that conduces toward the evolutionary stability of this strategy are not factors that are invariantly present in the experience of all organisms in a species. For instance, an organism with a bourgeois reserve strategy could be nested in populations consisting of combinations of innumerably many strategies. In some of these conditions, for instance in a population dominated by the bourgeois reserve strategy, organisms will not experience any success at all with the reserve strategy. Again, this is because fights will not persist at all in a population of bourgeois strategists. However, in war of attrition models evolutionary stability requires playing the reserve strategy against convention-breaking mutant strategies. Mutants like this will do no better than bourgeois reserve strategists in a 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 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. Thus, the bourgeois reserve strategy is only stable against this kind of mutant strategy because of the tendency to persist in fighting regardless of whether the tendency has been successful in the past. In other words, learning from individual experience would not tend to converge on the bourgeois reserve strategy across all of the conditions required for its stability. Since learning would not reliably produce the
relevant phenotype, it cannot be the mechanism by which the bourgeois reserve strategy develops. 32

This is a kind of “poverty of the stimulus” argument. If the motivation to play the reserve strategy depended on prospective processes (or other learning mechanisms), then it would require learning that reward is contingent on playing reserve. In that case, there would be no reason to expect that individuals would end up using the evolutionarily stable strategy regarding when to play the reserve. Thus the bourgeois reserve strategy has to develop invariantly across variation in fighting success and it has to be non-prospective.

Given the problems of cognitive complexity in the previous section, this is clearly a better explanation for why some organisms have a retributive motive. The fact that the model applies to organisms without prospection eliminates much of the cognitive complexity required to explain punishment as a behavior modification strategy. The only capacities required of an organism are that it needs to identify situations of resource competition and to attack for a specific duration if the intruder does not retreat. While this is still a complex capacity, eliminating the interaction with prospective capacities leaves us far less mystified about how such a variant could have originated. Moreover, it is easy to see how this capacity could be built up from more basic capacities, for instance, the capacity to attack a conspecific with an unfamiliar smell. The model has the added benefit that it would explain why a retributive motive would be constrained by proportionality, since the optimal duration of fighting for the reserve strategy is determined by the value of the disputed resource. This is an explanatory benefit that could yield a large return in explaining the structure of moral punishment, which is also constrained by proportionality.

3.2.3 Why not Hawk/Dove? The Importance of Temporal Extension

Now, 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 also not the only model in which the convention must be enforced. For instance, Maynard Smith and Parker (1976) analyze four different games in which asymmetries can be used to resolve contests for resources, including both the war of attrition and hawk/dove.33

By focusing on the war of attrition model as opposed to hawk/dove, I have suggested that the origin of retribution derives from temporally extended contests, but perhaps the argument generalizes to these discrete interactions in which conventions break symmetry. Such a proposal

32 I am not claiming that one cannot learn to follow a bourgeois convention (Skyrms 1996, 71–75). However, I do claim that the bourgeois reserve strategy as evaluated in war of attrition models cannot be learned.

33 The other two were retaliation (Maynard Smith and Price 1973; Gale and Eaves 1975), and the graduated risk game. In the following I ignore both. The graduated risk game is a special case of the generalized war of attrition (Bishop and Cannings 1978), and in the retaliation game the ESS that seems most relevant (the retaliator strategy) leads to a war of attrition anyway (Maynard Smith and Parker 1976, 173).
would be no less novel than the one I just presented. However, I have a major misgiving about explaining the origins of retribution in terms of the hawk/dove game.

The hawk/dove game is usually modelled as a one shot interaction in which the payoffs are symmetrical with the following payoff structure: \( E_G(D) > E_D(D) > E_D(H) > E_G(H) \), where \( E_G(D) \) is the expected value of playing hawk when the opponent plays dove, etc. In hawk/dove, if the expected cost of defeat is greater than the expected value of victory, contests will also tend to be resolved by an asymmetry like ownership (either in favor of the owner or, in the paradoxical equilibrium the intruder).

The question is whether, in such a contest the convention must be enforced in a way that is similar to the ESSs in the war of attrition. This is not an easy question to answer, because in hawk/dove, there is no possibility of a convention without enforcement. The convention and its enforcement are collapsed into a single decision. When an agent plays the bourgeois strategy in the role of owner, the decision to play hawk both obeys and enforces the bourgeois convention, and there is no way to decouple the two (without changing the strategy space). This in itself is not all that problematic, for the costs that hawk imposes are certainly part of what makes a conventional strategy an ESS against convention-breaking mutants. However, it highlights something that is problematic: that one-shot 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 and Parker 1976, 161) However, most animal conflicts are continuous in that they have the potential to be extended over time. If so, then we have occasion to ask questions about the relative success of various, temporally extended strategies: E.g. subsequent to the decision to play hawk at T1, why should the animal persist in a costly endeavor at any subsequent time point (e.g. until victory is achieved or injury is taken)? Why not instead escalate for brief period of time (perhaps even infinitesimally)? The hawk/dove game cannot answer these questions which are clearly relevant when contests have the possibility of temporal extension. Here is another way of getting at the same point: 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 game is really a war of attrition after all.

Not only do animal contests tend to have the possibility of temporal extension, retributive motives in humans and other closely related mammals appear to be temporally extended. That is, they are not mere dispositions to make a certain kind of decision at a discrete point in time. Consider: History is rife with anecdotes in which revenge is plotted over long periods of time during which the costs of planning and execution steadily increase. Displaced aggression seems to rely on aggressive motives that persist well beyond the moment of provocation and can influence a pattern of decisions made over a period of time (Pedersen, Gonzales, and Miller 2000; Fabiansson and Denson 2012). One expects that the phenomenon of redirected aggression in other primates will be underwritten by similar motives (Cheney and Seyfarth 1989). Ruminating on anger can influence patterns of decisions over time and is costly because it interferes with executive function, as suggested both by first-person experience and empirical research (e.g. Bushman et al. 2005; DeWall et al. 2007; Denson et al. 2011). Making people angry (by having them watch a depiction of merciless bullying) can influence their subsequent...
judgments about punishment in response to unrelated vignettes (e.g. Lerner, Goldberg, and Tetlock 1998; Goldberg, Lerner, and Tetlock 1999). Finally, experimental research on rodents suggests the existence of a kind of aggressive arousal that can lead to aggression that persists over time (Michael Potegal 1992; M. Potegal 1994). The point is this: the phenomena of retribution appear to be manifestations of retributive motives that are extended over time, that can result in steadily increasing costs, and that have the potential to influence patterns of behavior over time. These phenomena are not well understood as a disposition to make a discrete decision. This suggests that retributive motives are better explained by models that include temporally extended strategies in their strategy space. For these reasons, the war of attrition model seems more appropriate than hawk/dove for understanding the evolution of retributive motives.

3.2.4 The Generalized Asymmetric War of Attrition

Importantly, much of the literature on the war of attrition has complexified 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, this is not a safe assumption for many organisms. In many organisms, there are differences in resource holding power (RHP) that introduce asymmetries in the costs of persistence. That is, for some individuals, differences in fighting ability or robustness make it less costly 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 RHP (Parker 1974). Another asymmetry involves the value of a resource to an individual (e.g. Parker and Rubenstein 1981; Grafen 1987). 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. These asymmetries introduce interesting changes in the ESS for the war of attrition. Instead of deciding competitions on the basis of ownership, the generalized war of attrition (GWA) model suggest that organisms will decide competitions based on a combination of RHP-related variables and individual value of the resource. Specifically, the model has a conventional ESS in which competitions are determined by whoever has the highest value for the resource discounted by the relative cost of persisting (Parker and Rubenstein 1981).

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 to play reserve.

Recently, Huetegger and Zollman (2010) have offered an interesting critique of ESS methodology that might threaten this claim. They introduce several other kinds of equilibria that can be biologically significant aside from ESSs. These equilibria arise in part because of behavioral equivalencies between different strategies (at specific states of the replicator dynamics). This suggests a worry about using the ESS methodology in the war of attrition: that the bourgeois reserve strategy is not 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 identical. Thus, it seems possible that the population could drift to a state where
convention-breaking strategies could invade and where the bourgeois reserve strategy does not dominate (cf. Hammerstein and Parker 1982).

Nevertheless, something very much like the bourgeois strategy remains the ESS so long as there are mistakes in role assessment (Haccou and Glaizot 2002). For instance, in the generalized war of attrition model, there are likely to be times when an organism makes a mistake regarding the value of a resource or the relative RHP of its competitor. If so, then reserve strategists will receive a net benefit against its non-reserve counterpart. For instance, every possible mistake concerning RHP assessment has an equivalent net result for the relevant strategies except one, in which a bourgeois reserve strategist mistakenly assesses its RHP as higher than its simple bourgeois opponent. In that case, the bourgeois reserve strategists wins the resource without a fight. The asymmetric benefits of mistakes makes the net payoff for the simple bourgeois strategy less than the payoff for the bourgeois reserve strategy.

Since I have argued that the reserve strategy is what creates the need for a retributive motive, the GWA remains a good explanation of the existence of retributive motives. At the very least, some combination of war of attrition models is likely to explain the existence and maintenance of retributive models over a large swath of evolutionary history. Since the GWA is more applicable to mammalian species (in which abilities to assess relative RHP are widespread), the following discussion will focus on the GWA model. Importantly, the changes in strategy predicted by the GWA do require increased cognitive complexity. Nevertheless, it does not require prospection and the complexity it does require resides entirely in the inputs to retributive motives, or in the capacity to elicit the reserve strategy under the right conditions. Moreover, it is easy to see how these changes could occur piecemeal by the acquisition of novel triggers for the retributive motive (e.g. the perception that a challenger is smaller in size). So this model is not afflicted with the explanatory deficiencies of the models discussed above.

4 CRITICISMS

Nevertheless, there are several other misgivings that one might have about using this model as a way of understanding punishment or punishment motives. One worry is that the reserve strategy just is not a form of punishment. Another is that this kind of competition does not result in temporarily spiteful punishment. A third is that this kind of competition could not be

---

34 On Haccou and Glaizot’s model, the ESS is actually one in which both roles (whether owner versus intruder or higher versus lower RHP) will persist depending on non-overlapping probability distributions. Nevertheless, the reserve component remains intact insofar as the ESS determines that one of the roles will always persist longer than the other (except for an infinitesimally small probability that they persist for the same duration). Thus for our purposes, this ESS resembles the bourgeois reserve strategy in the relevant respects.

35 Though it may also apply more broadly to amniotes or tetrapods or some subset thereof. This is because some species of avians, reptiles and perhaps even amphibians may exhibit capacities for assessment of this kind.
responsible for non-moral punishment in *Homo sapiens*. The first worry can be dispatched with immediately, because by the definition I am working with, this kind of competition clearly leads to instances of punishment whenever an individual plays the reserve strategy. In that case, the “owner” of the resource (or whoever is favored by the relevant asymmetry on a given model) is inflicting a cost on the other competitor at a cost to itself in response to a provoking incident that diminishes the fitness of the owner (specifically, challenging its “claim” to the resource by engaging in competition).

The second worry is slightly more troublesome. This kind of punishment may appear not to be temporarily spiteful, because whoever wins the competition will achieve an immediate benefit, namely ownership of the resource in dispute. While this point may hold concerning consumable resources, it is not so for resources like territories, in which the value of a resource is often extended over time. In that case, part of the value of the resource is deferred, so if an organism plays the reserve strategy to defend such a territory, it will sometimes engage in temporarily spiteful punishment. This kind of circumstance is especially likely to arise in organisms for which territories are valuable for reproductive purposes and for which reproduction occurs only at certain times of the year. In that case, there will be times prior to the reproductive phase in which organisms will have to “set up shop” in a territory and during which defense of the territory can only have deferred value. In fact, there have been several species in which willingness to defend a territory does not entirely depend on the immediate value of the territory (Stamps and Tollestrup 1984). There is no reason to think that the distinction between immediate and deferred value would have any bearing on successful strategies in the GWA, so if an organism plays the reserve strategy to defend such a territory, it will sometimes engage in temporarily spiteful punishment. This kind of circumstance is especially likely to arise in organisms for which territories are valuable for reproductive purposes and for which reproduction occurs only at certain times of the year. In that case, there will be times prior to the reproductive phase in which organisms will have to “set up shop” in a territory and during which defense of the territory can only have deferred value. In fact, there have been several species in which willingness to defend a territory does not entirely depend on the immediate value of the territory (Stamps and Tollestrup 1984). There is no reason to think that the distinction between immediate and deferred value would have any bearing on successful strategies in the GWA, so

However, this is not yet an explanation for why human beings possess a retributive motive. It could only offer such an explanation if this motive were preserved in *Homo sapiens*. So far, I

36 On a related note, Forber and Smede (2014) suggest a mechanism of negative assortment by which spite *per se* can evolve. This may be another model capable of explaining retributive motives. However, similar to Skyrms’ model above, it is unclear to me whether this model applies to other animals with whom we share close common ancestors. Thus, I am not sure whether it applies human ancestral populations prior to the evolution of retributive motives.

Though spiteful motives and punitive motives may be conceptually distinct, I am not sure whether they are psychologically distinct as Jensen (2010) seems to suggest. For instance, to me it seems likely that anger aims at something like “repaying offense” by harming another. If so, the aim is both self-regarding (to repay insults to ME) and other-regarding (to impose a cost on YOU). This is to suggest that spite and punishment may be subserved by the same psychological mechanisms in the lineages to which humans belong. If so, then the evolution of spite and punishment may complement the GWA as an explanation of retributive motives. Misgivings aside, the evolutionary dynamics of Forber and Smede’s model, however, may provide an explanation of how retributive motives were adapted to a different domain of social interactions.
have given no reason to expect that it has been, nor have I given any reason to think that human (or even primate) resource competition is anything like the war of attrition. Nevertheless, there is some evidence that human anger (and perhaps also punishment) has been shaped by the war of attrition.

Aaron Sell (2005) has used the generalized war of attrition model to make surprising predictions about human behaviors and attitudes, and many of these predictions have turned out to be accurate. Some of the predictions are as follows:

"...strength should lead to (i) greater success in resolving conflicts in one's favor, (ii) greater sense of entitlement (i.e., expectation of better treatment), and (iii) greater anger-proneness. The relationship between strength and anger should be specific and functional, not indiscriminate: Strength should predict how easily or frequently one angers, but not necessarily other aspects of anger, such as rumination (how long one remains angry)... Ancestrally, stronger males had more to gain and less to lose by fighting than weaker males did. So stronger men should not only be more anger-prone, but also (ix) have a greater history of actually using physical intimidation and fighting. Aggression was more efficacious the stronger one was; this should have selected for a decision system that judged aggressive responses as more likely to be successful the stronger one is. Therefore, (x) stronger men were predicted to more strongly endorse the view that personal aggression is sometimes a useful or necessary method for dealing with interpersonal conflict." (Sell, Tooby, and Cosmides 2009, 15075)

How does the GWA model lead to these predictions? As Sell and others point out, resource competition in humans is incredibly abstract, and so the bargaining chip in these interactions tends to be abstract cost benefit transactions, rather than immediate material benefits:

Though humans face conflicts of interest over morsels of food (e.g. “that's the last doughnut!”), the majority of resource conflicts between humans do not involve tangible, materially-represented resources, but instead involve conflicts over courses of action (retrospective and prospective), exchanges of information, social alliances, and other abstract cost-benefit tradeoffs between the welfare of two individuals. (Sell 2005b)

Sell claims that the GWA model can be applied to these tradeoffs, presumably because they can be given a value representing the fitness value of the tradeoff. Assuming that strength corresponds with RHP, we can suppose that, ceteris paribus, stronger men will be more willing to play the reserve strategy, since this is an asymmetry that the GWA predicts will be used to...
resolve resource competition. So stronger men will more often resolve these transactions in their favor, they will tend to have a sense of entitlement (expecting others to take on costs to benefit them) and they will be more anger prone, since the relevant asymmetry will more often be in their favor. Sell’s other predictions follow from these fairly straightforwardly.

On this view, anger is an implementation of the reserve strategy to resolve cost-benefit transactions in one’s favor. If so, then temporarily spiteful punishment strategies are one explanandum of Sell’s theory, since the cost benefit transactions have deferred and intangible effects on one’s fitness. However, on this and related models (M. E. McCullough, Kurzban, and Tabak 2012; Petersen et al. 2012), punishment is understood as a behavior modification strategy. The adaptive value of anger is not just to resolve a specific cost benefit transaction in one’s favor, but also to ensure favorable cost benefit transactions involving a given person (or audience) in the future by “recalibrating” the dispositions of that person to value one’s welfare.

This introduces another worry, that these models share the explanatory deficiencies of other reputation models discussed above. However, despite the focus of Sell and others on behavior modification, the adaptive value of punishment on these models need not be understood entirely in these terms. This point is made when we recognize that these models of human punishment and revenge are consistent with two distinct evolutionary stories concerning the temporarily spiteful punishment strategies (manifested in attempts to recalibrate dispositions through revenge and punishment). On the one hand, this recalibration strategy might have evolved primarily because of the punishee’s ability to learn from punishment by recalibrating their dispositions toward the punisher. On the other hand, the relevant strategies might have evolved prior to the evolution of the cognitive systems responsible for negotiating cost benefit transactions. In that case, they would have been shaped by the structure of the war of attrition, then modified as cost benefit transactions became objects of resource competition. Sell’s model of anger and related models of punishment and revenge are not necessarily committed to either explanatory strategy.38

If the arguments of section 1 are correct, then it may be that the latter strategy is the more promising one: that retributive motives evolved prior to the advent of abstract resource competition and was coopted for behavior modification. If so, then the problem of cognitive complexity reasserts itself in a different way. In this case, the problem is not about how such complexity arose de novo but instead how such complexity could have gradually arisen to take

38 While the model is not committed to either explanations, evolutionary psychologists in the tradition of Cosmides and Tooby may be committed to the former explanation. This is because of a long running commitment to individuating modules according to their selected effect functions (e.g. Tooby and Cosmides 1990; Barrett and Kurzban 2006). According to this view, something that has the function of recalibrating dispositions is a fundamentally different psychological kind than something which has the function of preventing the invasion of mutant strategies (without recalibrating dispositions). On this view, functions cannot evolve. To speak of functions evolving, one needs a way of individuating function that is independent of the effects of evolution. Otherwise, it will not be possible to pick out a capacity that has different functions across evolutionary time, since evolution can shape traits to achieve different selected-effects functions.
shape as we find it in humans. Since there are at least two candidate psychological processes capable of implementing temporarily spiteful punishment, a retributive impulse and a retributive incentive, there are also two ways to explain how these processes might have gradually evolved to become cognitively complex. On the one hand, retributive incentives that could gradually evolve to interface with prospective capacities and stand as proxy for deferred or intangible benefits. On the other hand, retributive impulses could gradually evolve to inhibit prospective capacities as the latter capacities begin to encroach on GWA interactions.

Consider first retributive incentives as proxy for deferred benefits. As noted before, the structure of the GWA is the same whether or not a territory has value that is extended over time. So the motive to play the reserve strategy could easily shift from defense of resources or territories that have immediate value to defense of resources that have deferred value (as is the value of territories used for breeding purposes). Likewise, it is easy to see how retributive incentives could stand as non-material incentives to defend resources with more intangible value.

Here is a more concrete story about how this might go. Consider that our common ancestor with rats probably competed for physical territories for breeding purposes. Territorial competition of this kind has been extensively studied in rats (see e.g. D. C. Blanchard and Blanchard 1984; D. C. Blanchard and Blanchard 1988; D. C. Blanchard and Blanchard 2003). These conflicts resemble a war of attrition since lethal strategies are not used and the costs of competing increase over time. Accordingly, these encounters are usually quickly resolved in favor of the rat already occupying a given territory.

While primates do defend physical territories, they do not do so for the same reproductive purposes as rats. Instead, there is some reason to expect that in the primate lineage resource competition became increasingly abstract in tandem with increasing social complexity (Byrne and Whiten 1989). Position in a social dominance hierarchy becomes the main determinant of reproductive success, so defending a position in a hierarchy takes the place of defending a physical territory. Even in rats, when population density increases, the mating system shifts from polygynous (where the alpha rat monopolizes estrus females) to polygynandrous (where males copulate sequentially with estrus females, see Calhoun 1963; Robitaille and Bovet 1976). Since dominant rats in the latter mating system have reproductive priority and probably greater success (Thor and Carr 1979), defending a position of dominance from cohabiting males becomes as important for reproductive success as defending the colony from unfamiliar males. Consistent with this conjecture, the dominant rat in a colony (with a mix of male and female

39 Similar patterns of aggression have also been observed in tree shrews with which we share a closer common ancestor than rodents (Olsen 1969; Walletschek and Raab 1982). While the aggression literature in nonhuman animals focuses mostly on inter-male territorial aggression, female rodents also display some patterns of territorial aggression in defending their nests and their pups.

40 Though there is rarely anything resembling an ordered hierarchy except that one male, the alpha, initiates and wins most fights.
Isaac Wiegman
1/18/16
14160 words

rats) reinforces dominance by the same patterns of aggressive behavior exhibited by resident rats in settings with lower population density. Nevertheless, as social cognition increases in complexity by comparison, a broad range of possibilities arise for defending position in a dominance hierarchy with minimal energy expenditure and minimal risk of injury. Instead of repeatedly biting the back of a subordinate conspecific as rats do, primates can respond to challenges and reinforce dominance with a threatening facial expression or by physically displacing a subordinate with impunity. In humans, the abilities to wield symbols and language make it possible to “put someone in their place” with symbolic or verbal barbs and blows. Certainly, there is the possibility of physical escalation, but it is advantageous even for an unassailable individual to waste as little energy as possible “putting someone in their place”. The increase in available options for defending social dominance and its analogs (perhaps prestige or respect) suggests that the motivational aim of retributive incentives would gradually become more abstract or diffuse, while still having a kind of central direction (e.g. putting others in their place). This is one plausible path by which an incentive to play a reserve strategy could gradually come to resemble the vengeance motive that we are familiar with today.

This “just-so” story has a key virtue. It is a how-possibly explanation of why organisms would evolve to have incentives to satisfy abstract aims (e.g. keeping people in their place) that serve as an immediate proxy for future material benefits. Resident rats are apparently motivated to bite the backs of intruding and subordinate rats. They do so even when intruders are tied down on Plexiglas plates and consequently, when it is difficult to access the intruder’s back (R. J. Blanchard et al. 1977). This motive is “triggered” by subordinate rats or by the identification of a male rat with an unfamiliar scent. Apparently, the incentive to bite the back stands as a proxy for the reproductive benefits of maintaining dominance or ownership of a territory. Nevertheless, it is not difficult to see how this incentive could have evolved (along the lines of the story above) to be triggered by more complex stimuli (input side complexity) and to motivate the achievement of more abstract aims or goals (output side complexity), rather than more concrete ones (like biting the backs of opponents). Subtle indications that one’s dominance status is under threat could come to instate an abstract aim of putting someone in their place (by a broader range of means).  

This kind of explanatory strategy offers some hope for solving the problem of cognitive complexity. This is because it shows how non-material incentives with abstract aims could evolve to “stand in” for deferred or intangible benefits (as discussed in section 2.3). This would explain how these incentives might have become a viable criterion for selection of vengeful

---

41 However, dominant macaques do bite the backs of subordinates and this may have a similar function to back biting in rats (Adams and Schoel 1982). Given the broader behavioral repertoire of macaques, back biting may be derived from the same ancestral behavioral trait as back biting in rats.

42 Griffiths work on Machiavellian emotions (2004, 2010) considers how and why a Machiavellian arms race might have led to this result in the evolution of emotions like fear and anger. However, his focus is on input side complexity.
behaviors through prospective processes as these processes evolve to arbitrate competing incentives (e.g., drinking, eating, avoiding predators, etc.) or inhibit counterproductive aims in more context sensitive ways. Moreover, as mentioned before, selection on the reserve strategy should favor a motive to inflict costs that are on average, proportional to the subjective value of a resources. Thus, they will also tend to track the deferred material benefits of punishment: the value of the resource in dispute.

Now consider the second strategy, according to which retributive motives (or more specifically retributive impulses) gradually evolved to inhibit prospective capacities. The idea is that as prospective capacities achieved greater generality, they would begin to be applied to situations for which retributive impulses were designed. Insofar as those situations require a retributive impulse (to implement an ESS), there would be selection pressure for retributive impulses to inhibit prospective capacities in the relevant context. On this story, the problem of cognitive complexity might be mitigated by explaining how the gradual encroachment of prospective capacities is dealt with by a retributive motive that is slowly modified to retain control precedence in the relevant situations.43

While these evolutionary stories are both highly conjectural, they show why the GWA allows for more gradualistic explanations of the evolution of retributive motives. That is why they overcome the problems of cognitive complexity posed in section 1. If the arguments of that section are sound, this is a key explanatory benefit over and above explanations of retributive motives that place their origins subsequent to the development of complex prospective capacities.

5 CONCLUSION: UNDERSTANDING PHYLOGENETICALLY ANCIENT ADAPTATIONS

Importantly, the war of attrition model does not compete with the other explanations of punishment I discuss in section 2, so long as they are understood merely as explanations for the structure and maintenance of retributive motives in humans (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 for different purposes at different points in evolutionary history. So a retributive motive for resource competition could have been co-opted to serve a reputational function or to support cooperation in large groups. Moreover, given its role in resource competition, it is easy to see how the motive could come to have effects in closely related domains of cooperation or deterrence, both of which concern resources in which an organism

43 A good deal of research in social psychology suggests that angry rumination on a provocation can increase the amount of aggression against a confederate uninvolved in the provocation (e.g., Bushman, Bonacci, Pedersen, Vasquez and Miller, 2005). Thus, rumination may be a mechanism by which the control precedence of retributive motives is retained and extended over time. This idea comes from Barash and Lipton (2011, ch. 3).
has an obvious stake. Organisms cooperate to acquire mutually beneficial resources and they are benefitted when other organisms are deterred from compromising those resources.

While it is compatible with these explanations of the function of the retributive motive, the war of attrition model is clearly a better explanation of how a retributive motive could have originated. I argued that explanations involving the adaptive value of general deterrence via punishment suffer from difficulties when used to explain the origin of retributive motives in humans. The problem of cognitive complexity makes the explanation of the motive incomplete, since it is difficult to see how a trait with such a complex relationship with other traits could have arisen due to simple changes in development or genetics. The more primitive explanation of retributive motives offered by the GWA model helps to mitigate this problem. This case shows that we may sometimes achieve greater insight into the origin of a capacity by looking further back in time at the more primitive selection problems from which it may have arisen.

There is another reason why it is beneficial to look at the patterns of aggression in less cognitively complex organisms with whom we share more ancient psychological capacities. Due to the simplicity of their behavioral repertoire in relation to humans, it is a simpler matter to identify separate behavioral systems and assign them adaptive functions. While humans may have some of the same behavioral tendencies, they are masked by our enhanced abilities to regulate our emotions. Harmon-Jones and his colleagues put it this way:

Basic emotions, such as anger, provide organisms with relatively complex and biologically prepared behavioral potentials that assist in coping with major challenges to their welfare (Panksepp 1998). However, these inherited behavioral tendencies exist only as potential ways of behaving in organisms with larger, more complex brains. Thus, although humans may possess the same emotional instincts as other animals, we may not be as controlled by the dictates of emotions and thus we have more choices (Panksepp 1994). That is, our emotions may be regulated and thus may not directly affect behavior. (Harmon-Jones, Peterson, and Harmon-Jones 2010, 61)

While one might be skeptical that any of the same underlying causes could be involved in human and non-human aggression, it is important to recognize that human psychology did not arise out of an evolutionary vacuum. It is plausible that some dimensions of our emotional responses will be potentiated by phylogenetically ancient mechanisms. The phenomena of punishment in humans may be substantially influenced by such mechanisms.  

44 Similar lessons could be drawn as a criticism of contemporary evolutionary psychology and its approach to the evolution of emotions, cooperation, and altruism in hominin lineages. See e.g. (Ketelaar 2015; Kurzban, Burton-Chellew, and West 2014). Thanks to [revoked] for helping me to see this parallel.
WORKS CITED


Bermúdez, JL. 2003. Thinking without Words. https://books.google.com/books?hl=en&lr=&id=8ycw0eh7EEUC&oi=fnd&pg=PP1&dq=Bermudez+thinking+without+words&ots=K1pik9i&sig=wPgBmH_pg29rJRLjqSOH7kmqd2o.


http://www.springerlink.com/index/p436t045383772x6.pdf.


Schaafsma, Sara M, Donald W Pfaff, Robert P Spunt, and Ralph Adolphs. 2014. “Deconstructing


Sterelny, K. 2012. The Evolved Apprentice. https://books.google.com/books?hl=en&lr=&id=AffzTCFV6zoC&oi=fnd&pg=PR7&dq=%22evolved+apprentice%22&ots=8-gEfj-cGx&sig=qwxWBbm7BsLWYYr2zNT-_97jnVU.


