

TARGET ARTICLE

Evolutionary accounts of belief in supernatural punishment: a critical review

Jeffrey P. Schloss^{*a} and Michael J. Murray^b

^aWestmont College, Santa Barbara, USA; ^bFranklin & Marshall College, Lancaster, USA

Although largely unaddressed by evolutionary theory for more than a century after Darwin, over the last decade a wide range of adaptationist, byproduct, and memetic explanations have emerged for various recurrent features of religious belief and practice. One feature that has figured prominently in adaptationist accounts of religion is belief in the reality of moralizing, punishing supernatural agents. However, there is at present no unified theory of what fitness-relevant feature of the selective environment to which this cognitive predisposition is adapted. We distinguish two divergent and often conflated approaches to supernatural punishment theory which hypothesize the adaptive character of such beliefs arise from the fact that they increase cooperation or decrease the cost of incurring punishment for norm violations. We evaluate these, and group and individual selectionist versions, in view of game theoretic models, experimental studies, and ethnographic data in light of which each proposal is plausible but with which none is fully concordant.

Keywords: supernatural punishment; afterlife beliefs; moralizing gods; evolution of religion; cognitive science of religion; human cooperation; cheater punishment; error management theory

For as long as scholars have studied the phenomenon of religion they have marveled at its power to both unify and divide. Religion is among the most powerful motivating forces in human culture, serving to foster strong cooperative alliances among members of the same religious community as well as severe and even lethal hostility to those outside the group. Many of the most prominent scientific explanations of religion argue that it is adaptive in character because of this capacity to sustain cooperation among individuals within groups in the face of forces that threaten their unity. It is uncontroversial that living in cooperative groups brings significant adaptive benefits both to the individuals in the group as well as to the group as a whole. At all scales of life, from cells working together in a body to social collaboration between individuals, cooperation generates the twin benefits of increased size and specialization of function which allows groups to interact with their local environment in ways that permit them to extract greater benefits, to better resist challenges, and to do both with greater efficiency.

Among humans, cooperators can work together to better defend or expand their territory, groups of hunters can bring down more and larger game, and group members at agrarian stages of cultural development can divide their labor in order to

*Corresponding author. Email: schloss@westmont.edu

improve efficiency of resource extraction and increase carrying capacity for their population. This sets in motion positive feedback, which permits particular individuals to develop further technical and intellectual skills while relying on group members to meet other needs. The result is that human social groups – in a way analogous to multicellular organismic organization (Michod, 2007) – benefit from functional specialization that would not be possible without cooperative interdependence (Maynard Smith & Szathmáry, 2001; Michod & Herron, 2006; Michod, 1997).

While the potential benefits of group life are substantial, these benefits are hard to acquire and sustain in the face of threats of defection. The prospect of defection without loss of reward provides powerful incentives for members of a group to free ride on the efforts of others. This challenge constitutes the central problem of group life at all levels of biological organization from sub-cellular components, to multicellular organisms, to groups of animals living socially, and it is a difficulty that must be solved if groups are going to realize the adaptive power of number.

Such challenges to cooperation are represented at the dyadic scale by the standard Prisoner's Dilemma, and are greatly magnified at group levels as illustrated in models like the Commons game and made famous in "The Tragedy of the Commons" (Hardin, 1968). If all members of the group cooperate, contributing their resources to serve the common good, the group as a whole reaps maximal benefit. However, each member of the group has incentives to cheat others, reaping greater individual benefits for themselves while helping others minimally or perhaps even causing them harm. In a situation where such defection is a tempting strategy, the fragile economy of cooperation threatens to erode rapidly, splintering former cooperators into cheating opportunists, ultimately surrendering the benefits of cooperation. Interacting group members thus face a deep and vexing problem: how to achieve and sustain cooperation in the face of incentives to defect?

One strategy is commonly employed by social organisms and especially in human communities, from small tribal groups to large modern societies: punish cheaters. Threats of punishment for defection can force a revision of the payoff matrix with the result that incentives to cheat may be reduced or eliminated.

However, punishment is not a panacea, and may be a meta-stable means of maintaining cooperation for several reasons. First, it is often costly.¹ In order to sustain a system of punishment, members of the group may have to contribute resources to detecting and punishing those who fail to cooperate. Second, the mechanisms of punishment may be liable to corruption. Third, any system of punishment is only as good as its cheating detection mechanisms. If there is a system of punishment, cheaters will find increasingly sophisticated ways of avoiding detection. This leads to a degradation of effectiveness and/or an escalation of costs, as punishers are forced into a coevolutionary arms race of finding new ways to catch ever more sophisticated cheaters. Finally, any system of punishment is susceptible to the problem of second-order cheating (Johnson, Stopka & Knight, 2003; Fehr, 2004; Panchanathan & Boyd, 2004). The same incentives that present the temptation to defect initially are now in place when it comes to the system of punishment. If one can find a way to keep from making a contribution to the system of punishment, one can enjoy its benefits without paying any of the costs (Milinski & Rockenbach, 2008). As a result, punishing cultures face the further problem of how to deter the regression of "higher-order defection" (Fowler, 2005a; Schloss, 2007).²

Not only is punishment vulnerable to practical liabilities in implementation, but even when it is utilized, recent work suggests that it might not yield important benefits for cooperation. Although empirical studies have clearly demonstrated that punishment increases the amount of cooperative *investment* in public goods games (Fehr & Gächter, 2000, 2002), subsequent experimental work has found that punishment may not increase average *payoff* (Botelho, Harrison, Pinto, & Rutström, 2005; Page, Putterman, & Unel, 2005) or that it actually results in a decrease of average net payoff (Dreber, Rand, Fudenberg, & Nowak, 2008; also Milinski & Rockenbach, 2008; Sefton, Shupp, & Walker, 2007; Egas & Riedl, 2008; Ostrom, Walker, & Gardner 1992). In examining the variation in total payoff to individuals within interacting groups, punishers do worse than non-punishers. Given the ostensible net costs of punishment, there is disagreement over whether it even constitutes an adaptation for maintaining cooperation (Dreber, Rand, Fudenberg, & Nowak, 2008).

Two accounts of the adaptive value of supernatural punishment

This does not mean that punishment cannot function to stabilize cooperation in situations where the costs do not preclude net benefit (Fowler, 2005b; Ohtsuki, Iwasa, & Nowak, 2009). Indeed, an ideal system of punishment would therefore involve minimal or no cost, along with no possibility of corruption and no chance of failure to detect defection. Notice that a system like this would avoid the last problem above as well: if a system of punishment is cost-free, second-order cheating is incoherent since members of the group are not required to make a contribution to punishing in the first place. Yet how could such an ideal system of punishment be implemented? According to some proposals, the optimal implementation would involve belief in moralizing gods or some other form of supernatural sanctioning (Bering & Johnson, 2005; Johnson & Bering, 2009; Norenzayan & Shariff, 2008). Systems of religious belief often feature disinterested gods, spirits, forces, or ancestors who take a deep and abiding interest in the moral behavior of members of human groups.³ These supernatural beings or cosmic forces can take the burden of punishing off of group members by imposing sanctions and rewards via their presumed control of natural or supernatural processes. Furthermore, these agents are conceived of as having abilities to discover our misdeeds infallibly (or with greater acumen than human group members), and they are (in many cases) not construed as liable to corruption or bribery by the defectors whose wrongdoings they oppose.

In spite of what might look like the *prima facie* plausibility of such proposals, there are, however, two quite different accounts of just how belief in supernatural punishing agents actually confers an adaptive advantage. In the first and most straight-forward account (one we will call the “cooperation enhancement” or “CE” account), belief in “supernatural punishment” (SP) is selected for – probably at the group level – because it enhances cooperation and reduces defection under conditions in which other mechanisms for sustaining cooperation break down or become unstable. In the second account (one we will call the “punishment avoidance” or “PA” account), belief in SP confers adaptive advantage to individuals not by enhancing cooperation (though this may be a byproduct), but by preventing them from incurring the high costs associated with having one’s cheating punished. In strategic situations in which the cost of punishment is very high, mechanisms that

deter cheating can be adaptive if they save an individual from ultimate net losses.⁴ In this second account, belief in SP is such a mechanism.

It is important to see that these accounts are genuinely distinct. One might think that they are not, if focusing only on the fact that each works by decreasing the frequency of defection. However, while both proposals invoke the same adaptive phenotype (defection-reducing-belief), they offer very different accounts of the adaptive challenge the phenotype solves and the mechanism by which it secures a benefit. In CE, it is secured by enhancing the benefits of cooperation while on PA it is secured by diminishing the costs associated with being punished for defection.

In what follows we will describe each view in greater detail, consider some empirical evidence that has been offered in favor of both models, and finally consider some theoretical liabilities that will require further articulation of the models.

The cooperation enhancement account

According to the CE account, individuals or groups who postulate punishing supernatural entities do better at forming adaptive, low-cost, stable alliances than those who do not, because defection has a much higher subjective disutility and is therefore avoided (Norenzayan & Shariff, 2008; Shariff & Norenzayan, 2007). Indeed, this proposal takes belief in supernatural sanctions to be a solution to the four-fold problems of stabilizing cooperation by punishment described above.

However, CE seems at face value to confront the following obstacle. Even if we can instill temporary fear of SP, once an individual cheats without supernatural reprisal, the cheater (and perhaps others) will realize that the gods can be fooled after all, do not really care, or more likely, do not even exist (Murray, 2009). For this reason, if the success of religious belief depends on the controlling power of punishment, it would be short-lived.

Two responses can be offered on behalf of this account of SP. First, while religion may facilitate cooperation, its origin and successful transmission may not require this benefit and the credible punishment that underlies it. It may well be that religion arises and endures natively, as a cognitive spandrel, and subsequently comes to confer cooperative benefits. Thus, religious belief is properly understood as an exaptation rather than an adaptation.

Second, belief in the efficacy of supernatural sanctions can be stabilized by additional features of religious systems that deflect the above epistemological problem. For example, religious systems that teach that punishments for cheating will largely or exclusively be dealt out in the afterlife will never be falsified by occasions of “getting away with” defection now.⁵ Also, humans appear to have a native disposition to attribute cosmic significance to events that involve great fortune and misfortune (Murdock, 1980; Pargament, 1997; Swanson, 1960), and are naturally inclined to believe that when good things happen, the gods, spirits or the ancestors are responsible, and when bad things happen, these misfortunes are due to SP or spiritual curse (Pargament, 1997; Bering, 2005).

In fact, one or both of these features are often found in religious systems. It is common for religious systems to invoke instruments like purgatory, hell, and karma which afford mechanisms of SP in the afterlife or future lives (Rappaport, 1999; Bering, Blasi, & Bjorklund, 2005; Astuti, 2007). Comparable trends exist for this-worldly fortunes and misfortunes. Of the 186 societies analyzed by Murdock (1980), every one included ascription of illness to supernatural cause; in a sample of 50

societies, Swanson (1960) found that the majority attributed fortune and misfortune to supernatural sanctioning of good and bad behavior. Moreover, recent work illuminates a widespread cognitive tendency to ascribe cosmic significance to our good and ill fortune (Bering, 2005; Barrett, 2004; Boyer, 2001). Thus, religious systems incorporate a variety of amendments that may sustain belief in the face of what might appear to be undermining evidence, and the apparent epistemological problem with supernatural sanctioning accounts of religious belief, though salient, is not disabling.

The punishment avoidance account

In the second version, the fitness advantage of belief in SP does not derive from increased cooperation, but rather from avoiding the costs of punishment for defecting. This proposal posits that the adaptive benefit of belief in supernatural sanctions accrues to the individual in situations where the actual, externally imposed consequences of being caught at cheating are so sufficiently substantial that additional and strongly internalized modes of cognitive deterrence pay off. Because the cost of punishment to a defector is a function of both the likelihood and the severity of punishment, two types of adaptation have been proposed to increase the costs of being punished and possibly to underwrite the benefit of belief in SP.

First among these are adaptations for cheater detection. A recursive theory of mind (ToM) for imputing and assessing intentionality (Dunbar, 1998; Humphrey, 1992; Ermer, Guerin, Cosmides, Tooby, & Miller, 2006), a sophisticated cognitive system tuned for processing cheating or rule violations (Cosmides, Tooby, Fiddick, & Bryant, 2005; Ermer, Cosmides, & Tooby, 2007; Cosmides, 1989), and an astute ability to both convey and interpret facial and other autonomic signals (Brown, Palameta, & Moore, 2003)—all enhance detection effectiveness (Johnson, 2009; Johnson & Bering, 2009).⁶ Moreover, willingness to punish defection at real cost (though not necessarily net cost) to the punisher (Fehr & Gächter, 2002), punishing in response not only to experienced injury but also in response to third party injury or perceived violation of norms of equity (Fehr & Gächter, 2005; Fowler, Johnson, & Smirnow, 2005; Fehr & Fischbacher, 2004), neurological structures and sentiment facilitating affective reward for punishing (De Quervain et al., 2004; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Fehr, Fischbacher, & Kosfeld, 2005; Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Price, Cosmides, & Tooby, 2002; Wilson & O’Gorman, 2003), and the emergence of coalitional punishment (Boehm, 1999a,b) have been proposed to increase the likelihood that detected defection will indeed be punished. Second are the expanded forms of punishment for and consequences of defection. Unlike other organisms, humans have the ability not only to form judgments about others’ tendencies to cooperate or defect based on direct observations of their behavior, but also to transmit and receive praise and gossip about their behavior. Those who acquire negative reputations may be avoided by other individuals or shunned by the group, thus degrading future cooperative opportunities (and, it is worth noting, providing a cheap form of defector punishment). In this way, reputationally mediated consequences amplify the actual costs (and in promoting wariness, also the likely detectability) of defection.

On this second view of SP then, belief in punishment is adaptive not because it enhances cooperation (which is already presumed to be relatively well-stabilized by the existence of efficacious mechanisms of punishment). Rather, as enhanced

cognitive capacities increase the likelihood of detection and punishment, and as reputationally mediated consequences amplify actual punitive costs, belief in SP enables the individual to avoid or reduce costs that would be incurred when one defects and is detected (Johnson, 2005; Johnson & Krüger, 2004; Johnson & Bering, 2009).

Empirical support for the efficacy of supernatural punishment

While the two theoretical models are speculative, the notion that SP is indeed efficacious in producing relevant behavioral dispositions appears to be supported by a range of convergent empirical data. First, ethnographic data indicate that communities that embrace moralizing, punishing gods or spirits display higher levels of cooperation (Johnson, 2005; Swanson, 1960) and features of belief underwriting cooperation (Boehm, 2008; Wilson, 2005). Second, there is evidence that human beings have a natural tendency to act in accordance with social norms when they are subjected to cues that they are under the watchful eye of even neutral observers. In one study, a collection box was placed in a university lounge in which students were to voluntarily place the money for their purchases. The study included two conditions. In the first condition the collection box was an ordinary wooden box while in the second condition, an identical box was used, with images of two stylized eyes on the front. Even though the “watchful eyes” were not real, the priming effect substantially increased student honesty (Bateson, Nettles, & Roberts, 2006). Similar work (Haley & Fessler, 2005) indicates that students have a higher rate of cooperative behavior when performing a competitive task on a computer screen with stylized human eyes as part of the desktop artwork.

Such studies highlight a general tendency on the part of human beings to cooperate when they are primed with cues that indicate that their behavior is being observed, but does belief in or priming with respect to *supernatural* observers add to the effectiveness of these cooperation-fostering tendencies? There is evidence favoring both yes and no answers to this question.

Bering (2005) has shown that, even from an early age, children tend to follow rules more consistently when they are primed to believe that a supernatural agent is watching them. In one experiment, children were brought into a room and shown a box which, they were told, contained a prize that they could have if they could identify it without looking. After giving the instructions, the experimenter instructs the child that he needs to step out of the room for a short time. In one condition, children were cued with a belief that an invisible princess (“Alice”) was in the room watching the experiment. Children who received the Alice cue cheated by looking in the box significantly less than those who did not, and even among cheaters, it took much longer for cued subjects to cheat than subjects who were not cued.

Adult subjects appear to show similar tendencies. For instance, subjects were engaged in a competitive game requiring them to select the correct answers to a series of questions. In one condition, participants were told that the game was designed by a graduate student at the university, in another they were told that the experiment was designed by a graduate student at the university who had died, and in a third they were told that the experiment was designed by a graduate student who had died and was believed by some to haunt the laboratory. All participants were then told that the computer program had an error which would sometimes cause the correct answer to appear on the screen before they were prompted to give the answer.

They were further told that, to keep the game honest, they must hit the space bar on the keyboard when the glitch occurred to clear the answer. Participants who received the ghost prime hit the space bar, clearing the correct answer, in about half the time of those who did not receive the prime. The results suggest that those who were primed with respect to the supernatural agent were more likely to engage in rule following, cooperative behavior (Bering, 2005).

Still, we may ask to what extent the deterrent effect of the primes in these experiments depends on the supernatural character of the concepts involved. These experiments may merely show that prosociality increases when “detection by agent” primes are used. This is just what one would expect among organisms where indirect reciprocity and reputation management are important for cooperation and survival (Alexander, 1987; Nowak & Sigmund, 2005), but to what extent does the supernaturalness of the concepts impact the relevant behavioral dispositions?

Shariff and Norenzayan found that subjects primed with religious concepts displayed enhanced voluntary giving behavior, as opposed to those who did not, in the context of the Dictator Game (Shariff & Norenzayan, 2007). In the game, a subject (the giver) is given a sum of money which can be shared with a partner subject (the recipient). The game is played only one time, and the giver can choose to give the recipient all of the money, only some of it, or none of it. The giver keeps the balance. Prior to playing the game students are required to read scrambled sentences which either contain supernatural priming words (e.g., spirit, divine, God, sacred, and prophet) or not. Givers were provided with US\$10 to distribute. The results showed that subjects primed with the religious words before the start of the game gave, on average, US\$2 more to the receiver than those without the prime (US\$4.56 vs US\$2.56) (see also Norenzayan & Shariff, 2008). Similar results were found by Randolph-Seng and Nielsen (2007), independent of religious precommitments. However, it is important to note that both Norenzayan and Shariff as well as Randolph-Seng and Nielsen report that priming with “secular, moralizing” concepts (such as “police” or “jury”) are equally successful in promoting prosociality. Randolph-Seng and Nielsen suggest that the most parsimonious explanation for these findings is that “cognitive ideals” (whether religious or not) are activated in appropriate environmental circumstances, thus neutralizing the efficaciousness of the distinctively religious features of the primes (Randolph-Seng & Nielsen, 2008).

Thus, it is not clear that supernatural attributes are necessary for the deterrent effect of moralistic, monitoring agents, nor is there evidence that such attributes increase deterrence when monitoring agents are believed to be present. This does not undercut the plausibility of supernatural sanctioning accounts, since such beliefs may increase the *likelihood* of construing the presence of a monitoring agent. That is, the attributes of hypothesized supernatural agents – special knowledge, presence, or perception – can still in principle accomplish something that natural agents do not: namely, extend the convictions that one is being monitored, thus reducing both the confidence of potential defectors that cheating will go undetected and the actual costs of monitoring, punishing, and being punished.

Unresolved difficulties for supernatural punishment theories

Supernatural punishment theories of divine or cosmic sanction have a great deal of *a priori* appeal, are robust in response to *prima facie* objections, and have suggestive empirical support, but each version remains susceptible to some key concerns.

Cooperation enhancement version

The CE version of SP represents one possible response to the challenge – faced by cooperating groups – of how to prevent individuals from defecting on one another and unraveling the fabric of trust that facilitates the flourishing of its members. The threat of punishment, when properly configured, leads individuals within a group to re-assess the potential costs and benefits of defection and thus provides motivation to resist the temptation to cheat. This model faces at least two unresolved questions.

The first question is why, at the level of the individual, would belief in supernatural sanctions be either helpful or necessary? In some forms of interaction involving cooperative synergy or non-zero sum games, defection is non-advantageous. Thus, the sanctioning power of religious beliefs would be unnecessary. In those interactive contexts in which there is a potential benefit to defecting and exploiting the cooperation of others, a cognitive deterrent to cheating (as opposed to Machiavellian discernment in when to cheat) would not appear to be helpful. There is no current CE proposal for why individual selection would favor a generalized inhibition to defecting such as is posited for belief in SP.⁷

The most straightforward and plausible response to this is that belief in supernatural sanctions is a group-level adaptation to coordinate cooperation and inhibit the destabilizing effects of defection (Wilson, 2002). While individual fitness might be optimized by judicious defection, group function benefits from individuals who refrain from cheating even when it would pay off, but this raises a second question. Leaving aside general debates over group selection, recent theoretical and empirical analyses suggest that punishment is not an efficacious means of increasing net payoff at the group level (Dreber et al., 2008). Moreover, even if it were, it is not clear that the group would benefit from hiring a god to do a human's job. This is because the attribution of moralizing, punishing but non-existent gods (or other supernatural entities) that enforce costly cooperation in those who believe in them, is especially vulnerable to exploitation by those who simply do not believe. Far from solving the problem of defection at the group level, delegating the important task of punishment to fictitious supernatural entities may make the problem worse. Such a system is susceptible to third-order (3°) defectors who do so not by directly betraying a cooperative partner (1° defection) or by free-riding on others' commitment to punishment (2° defection). Rather, the defection consists in being unfaithful to the second-order defection-deterrence mechanism, by disbelieving in the supernatural punishing agents.

One proposed solution to this problem involves supplementing the efficacy of belief in punitive divine agents with costly religious practices that facilitate recognition of those who truly believe and therefore have utilities favoring cooperative fidelity. "Costly signaling" accounts of religion posit that religious behaviors function as reliable signs of commitment (Bulbulia, 2004; Sosis, 2003, 2004; also Cronk, 1994; Irons, 2001), thereby promoting networked or positive-assortative cooperation that avoids defectors (Nowak, 2006). Strong empirical correlations between cohesion of cooperating groups and demanding religious practices lend support to this proposal; moreover, correlations between group solidarity and costly behaviors are much weaker or non-existent in secular communities (Sosis & Alcorta, 2003; Sosis, 2005; Ruffle & Sosis, 2006, 2007).

Notwithstanding its theoretical plausibility and empirical merits, this proposal entails several as yet unresolved questions. First, the arrow of causality underlying

these correlations is not clear: there could be cooperation-relevant selection bias in the kinds of individuals who join religious communities and who are willing to submit to demanding, corporately mandated behaviors. In such a case, religious behaviors would be genuinely reliable signals of cooperative disposition, but they would not be doing any causal work in sustaining community coherence as signals. Rather, they might be concomitant effects or phenotypic spandrels.⁸ Furthermore, costly religious displays may not even be reliable signals of commitment or cooperative intent. In a situation where a “behavioral entrance fee” can be willfully consented to, where the costs and benefits of the fee can be consciously assessed, and where paying it does not foreclose future opportunities for defection there is no reason to assume that payment will be a reliable signal of cooperative fidelity. Nor is there reason that payment should reflect underlying beliefs in the reality of SP, as opposed to hypocritical willingness to manifest the accoutrements of such belief. For this reason, some have proposed that “hard-to-fake” signals – signals that in a situation like this are intrinsically tied to internal disposition, often by autonomically rather than consciously mediated displays – would do a better job manifesting the genuineness of religious beliefs and commitments than would mere costly displays (Schloss, 2007, 2009; Bulbulia, 2009a,b; Gervais & Wilson, 2005 on other kinds of hard-to-fake signals).⁹ Second, the addition of a “cost” for religious systems to function effectively in cheater deterrence reintroduces the very problem they were supposed to redress – that of punishment costs.¹⁰ Indeed, there are theoretical and empirical reasons for concluding that no-cost signals can be equally effective (and thus selectively favored) in this context (Murray & Moore, 2009).

Punishment avoidance version

The PA account of belief in SP avoids virtually all the above problems faced by the CE account, because it is not a group adaptation that requires individuals to relinquish fitness enhancing opportunities, and it is not a cooperation-facilitating mechanism that stands to increase vulnerability to defection or costs in preventing it. Rather, it assumes that cooperation has been stabilized by highly effective, socially mediated punishment. Belief in SP is taken to enhance individual fitness by greatly inhibiting defection and thereby reducing the real and stringent costs of being punished by human agents. However, this account is attended by other questions. Since the optimal interactive strategy is to “be as cooperative as it pays to be, and as selfish as one can get away with,” it is important to determine whether the high potential costs of cheating – caused by elevated detection risks and the reputational consequences of defection in human social groups – make nearly unyielding cooperation the optimal strategy, as the PA account (apparently) maintains. Is this the case?

With precisely this in mind, Johnson and Bering (2009) describe two strategies that one might adopt in managing cooperative relations in reputation-intensive interacting groups. The first is the “god-fearing” strategy. God-fearers (GF) adopt the policy of (near) uniform cooperation, in light of their belief that their defection would be accompanied by inevitable supernatural detection and severe punishment. An alternative strategy is pursued by Machiavellians (M), who seek to defect when they believe they can get away with it, and cooperate otherwise.

Johnson and Bering’s comparison of the inputs to and payoffs of these two strategies is summarized in Table 1 below.

Table 1. Three strategies of defection.

Strategy	IS present?	Can exploit IS for personal gain?	Probability of detection (p)	Cost of punishment (c)	Cost of missed opportunities (m)	Payoff
Ancestral	No	No	High	Same	None	Lowest
Machiavellian	Yes	Yes	High	Same	None	Highest (if $pc < m$)
God-fearing	Yes	Yes	Low	Same	Some	Highest (if $pc > m$)

Note: Three strategies come into competition with the advent of the human intentionality system (IS) and complex language. Grey-shading indicates consequences that act against genetic fitness. Machiavellians outcompete ancestral individuals, and god-fearing strategists outcompete Machiavellians as long as $pc > m$ (table and text from Johnson & Bering, 2009).

Since GF reliably cooperate but M cheat when they believe they can go undetected, GF miss out on whatever opportunities exist to exploit others (m), while M bear the probability (p) adjusted consequences (c) associated with getting caught. If the latter penalties are more severe than the cost of missed opportunities to cheat ($pc > m$), GF follow a more adaptive strategy for the individual in the long run, thus creating selection pressure (genetic or cultural) in favor of belief in supernatural norm-enforcing agents that curb tendencies to cheat.

This proposal is innovative and plausible, but as yet it confronts several unanswered questions. First, to its credit, the model quite fairly avoids arguing that SP must be adaptive, and merely describes the environmental circumstances in which it would and would not be the preferred strategy. However, there are some hidden assumptions that may understate the advantages of an M strategy and obscure the challenges faced by a GF strategy. Specifically, “ P ” (the likelihood of being detected) is represented as being higher for M than for GF. This seems fair since GF are taken not to defect (or to defect less frequently than M), and since no behavior of any kind is perfect, M mistakes will presumably exceed GF. The problem is that the probability of detection (which for this model would more precisely be termed “frequency of detected defections”) is itself a product of f (frequency of defection attempts) and l (likelihood that any given attempt is detected). Asserting that $P_M > P_{GF}$ requires that $f \times l_M > f \times l_{GF}$. This assumes that M and GF defection attempts are equally likely to be detected and/or that GF so infrequently defect that the likelihood of detection approaches zero. Neither assumption is warranted, and there is in fact good reason to question the former: one who consciously deploys a strategy of intentional defection in situations one has assessed and judged to be propitious would seem more likely to employ defection acumen than one who rarely defects, without conscious assessment and in the face of the self-deception or cognitive dissonance that attends violating proscriptive beliefs.

Second, in cases where $pc > m$, cooperation should be selected for regardless of the motives that lead to it (i.e., even in the absence of supernatural belief). While religious belief might, in fact, be one sort of belief that would generate such behavior, so would, or so it seems, a (true!) belief that one’s immediate self-interest is served by cooperation, or a commitment to the singular importance of “obeying the moral law,” or the elimination of urges to cheat altogether. One might argue that there is

something about the threat of SP that is singularly well suited to deterrence of this sort, but not only are the empirical data for this, as we have seen, ambiguous, we also currently lack a proposal for why belief in the supernatural would be necessary to motivate cooperative behaviors that are natural to human beings and that are adaptive in the social environment postulated by this account.

There are a few promising ways in which the PA account of SP can be augmented in the face of the above two concerns. First, it can be argued that fundamental aspects of our capacities for assessing social risk have an innate bias toward overconfidence (Johnson, 2004; Johnson et al., 2006). One reason for this might be adaptive lag, or a temporal disequilibrium between adaptations and the selective pressures of a changing environment.¹¹ Risk-assessing capacities in our cognitive architecture arose and were both accurate and adaptive in a particular social environment, but as the above-described defensive adaptations arose (increasing the likelihood and the cost of being caught at cheating), capacities for assessing risks of punishment would not necessarily have kept up with them. Lagging behind in the evolutionary cognitive arms race would yield a native tendency to underestimate existing risk. Second, one might appeal to pleiotropy or phenotypic variability – the notions that a given phenotype is just one in a cluster of traits associated with a particular genetic endowment, or that a phenotype is adaptive in the average but not full range of environments that mediate it. Thus, it may be that overconfidence is not a maladaptive vestige of a previously adaptive cognitive disposition, but it reflects currently adaptive, general cognitive bias that gives rise to a suite of various behaviors having fitness benefits in a wide range of situations. However, in certain social situations it also manifests itself in behaviors that are detrimental. Because it would not be beneficial to eliminate the bias altogether, it must be selectively over-ridden rather than globally corrected. It is important to note that both lag and pleiotropy scenarios involve a tendency to underestimate the likelihood or cost of being detected in defection, which would benefit from a cognitive counter-weight that introduces an inclination towards caution. Belief in SP is posited to serve as an effective counter-weight of just this sort.

The existence of deeply seated cognitive biases that cannot be eliminated or ought not be globally corrected is an important and plausible adjunct hypothesis to this account of SP, and indeed there are good reasons to think that just such pre-existing tendencies do exist. For example, our amply demonstrated tendencies towards self-biased assessments of our own virtue, our unrealistically optimistic and over-confident projection of outcomes in conflict, and our tendency to believe that we are less susceptible than others to such biases (!) (Pronin & Kugler, 2007), systematically tempt us to believe that our attempts to cheat others will likely meet with success, even in cases where the odds of such success are low. While these tendencies might be adaptive in some conditions, they would not be when defection is subject to costly punishments. As a result, we are forced to find ways to manage our tendencies to err (Johnson, 2009; Johnson & Bering 2009; Haselton & Nettle, 2006; Haselton, 2007).

If other mechanisms prove insufficient for managing this tendency (e.g., Richard Alexander's suggestion that internalizing moral norms through conscience acts as a "reputation alarm," 1987), religion succeeds by positing an all-knowing moralizing agent, which raises our subjective assessment of the likelihood of detection. In this way, religion acts as a better deterrent than assessment of social consequences (better, that is, than either direct, conscious analysis of consequences or reflexive analysis via conscience).¹² This view of SP need not entail that GF cooperate all the time in order

to enjoy the hypothesized benefits (as indeed they do not). Instead, GF need only be more parsimonious in their defection than M, diminishing the instances of cheating and thus the instances of detected cheating.

This promising and considerably more nuanced version of SP addresses the difficulties raised above, but it also presents its own challenges. First, it is at least worth noting that this more nuanced view is significantly different from the view as originally described. In the initial formulation of PA, the theory commended itself simply as a way of helping strategic agents avoid the ordinary costs of defection. However, the present version casts SP as a tool primarily aimed at managing a native tendency to defect on strategic partners owing to an over-confidence in our ability to avoid detection. As a result, the success of this enhanced theory rests on a number of ancillary though as yet undemonstrated (even if initially plausible) hypotheses.

Second, as presently conceived, PA is also liable to the criticism that it fails to take seriously much less exotic mechanisms for managing errors of this sort. For example, others have argued that more mundane strategies of reputation management that appear to have evolved to facilitate and respond to indirect reciprocity may suffice to dampen socially inastute overconfidence (Alexander, 1987; Ohtsuki et al., 2009; Rockenbach & Milinski, 2006, 2009; though see Fehr & Rockenbach, 2003). Indeed, humans have a profound and undeniable capacity for internalizing moral norms, described nicely by Jerome Kagan: “The symbolic private assurance that one is virtuous – given by the self to the self – is an attractive prize humans seek” (1998, p. 164).

One response to this worry would be to claim that cognitive mechanisms for reputation management are not sufficient to assess risks of punishment, because they (along with other aspects of indirect reciprocity) work most effectively in groups that are small enough to reliably mediate reputations among likely cooperation partners. Indeed, Norenzayan has proposed that belief in supernatural, moralizing gods is a recent, cooperation-facilitating, cultural adaptation to life in large societies involving frequent anonymous interactions between individuals whose reputations are unknown to each other. There is considerable empirical support that such beliefs in supernatural sanctions are artifacts of recent, cosmopolitan religions, but far from being a support for the punishment avoiding account of SP, this point entails an additional (third) problem. The proposal that moralizing gods arise where indirect reciprocity and its attendant punitive consequences are no longer effective is the very opposite of positing that belief in moralizing gods is an adaptive response to the distinctive efficacy of reputationally amplified punishment. These are two very different and seemingly incommensurate proposals. The former is really a version of CE: SP is an adaptation for cooperation, which supplements the efficacy of punishment mechanisms in situations where they are not effective at detecting and imposing costs on cheaters. The latter PA account posits that SP is an adaptation for decreasing the risks of punishment in precisely those situations where its detection efficacy and socially imposed costs are very high.

Finally, even if these issues can be successfully addressed, this more nuanced version of SP faces an additional concern. In the preceding paragraphs we have been considering various alternative strategies for securing belief in the goods that SP theorists claim arise from belief in supernatural agents. What has not been raised are the potentially maladaptive costs of religious belief and practice that must be weighed in the balance. As some advocates of nonadaptationist accounts of the evolution of religion point out, religion *seems* to carry with it significant epistemic

and material costs wherever it emerges. Epistemic costs are incurred in virtue of the fact that religious believers must internalize and sustain counterintuitive, counterfactual and sometimes even transparently unreasonable religious beliefs (Irons, 2001; Rappaport, 1999). In addition, sustaining such beliefs requires compartmentalizing them so that the irrationality of that belief does not infect practical and theoretical domains where irrational belief could have catastrophic survival consequences (Bulbulia, 2009b). Furthermore, there are substantial practical costs. Religions routinely require participants to engage in a variety of disparate practices: become celibate for a lifetime, build huge structures with no obvious benefit, sacrifice one's crops or cattle to unseen gods, kill one's healthy offspring, give up the opportunity to work on special "holy" days, give up eating important sources of protein, stop to utter strange words and perform unusual gestures several times a day. In order to perform the M vs GF calculus, we would need to add the costs of religious demands (r) to the costs of forfeited opportunities to defect: GF is stable if and only if $(f \times l)c > m + r$. When these additional costs are added into the matrix, alternative modes of error management may look correspondingly more attractive.

Conclusion

Adaptationist theories of the evolutionary origin and persistence of religion take a venerable intellectual tradition and locate it within a theoretical paradigm that conceptually unifies this human trait with those of other organisms and provides opportunity to empirically test alternative proposals. Functionalist anthropologists and sociologists have long argued that religion plays an important role in coordinating the organismic character of human social life. Evolutionary theory provides an opportunity to formalize the notion of "function" in terms of contribution to an observable and quantifiable entity – fitness – and it offers a proposal for the origin and/or persistence of religious phenotypes in terms of a causal mechanism (natural selection) that operates non-teleologically in all living organisms and at multiple levels of scale.

Notwithstanding the promising proposals currently available, there remain fundamental questions that not only are unanswered but also are frequently unacknowledged or conflated. In this paper we have sought to systematically assess what we take to be one of the most promising theoretical attempts to explain a prominent aspect of many religions – belief in SP – in adaptationist terms. Our goal has not been either to advocate or to dismiss this approach or any one of its variants, but to help move it forward by illuminating several alternative and often inadequately distinguished hypotheses, and by assessing their respective strengths and weaknesses in light of theoretical concerns and empirical data. As a collaborative effort between a biologist and a philosopher with manifestly and at times contentiously divergent perspectives – which may represent in microcosm the interdisciplinary tensions in this expanding field – our aim has been to seek the conceptual clarity necessary for continued commerce and advance in this program of inquiry.

Figure 1 presents a decision tree of different proposals in this field and summarizes many of the issues we consider. Although our focus begins at node #2 with adaptationist accounts related to social life, even the first node distinguishing adaptationist and non-adaptationist proposals reflects issues salient to our discussion in at least two ways. First, while religion often appears fitted to securing

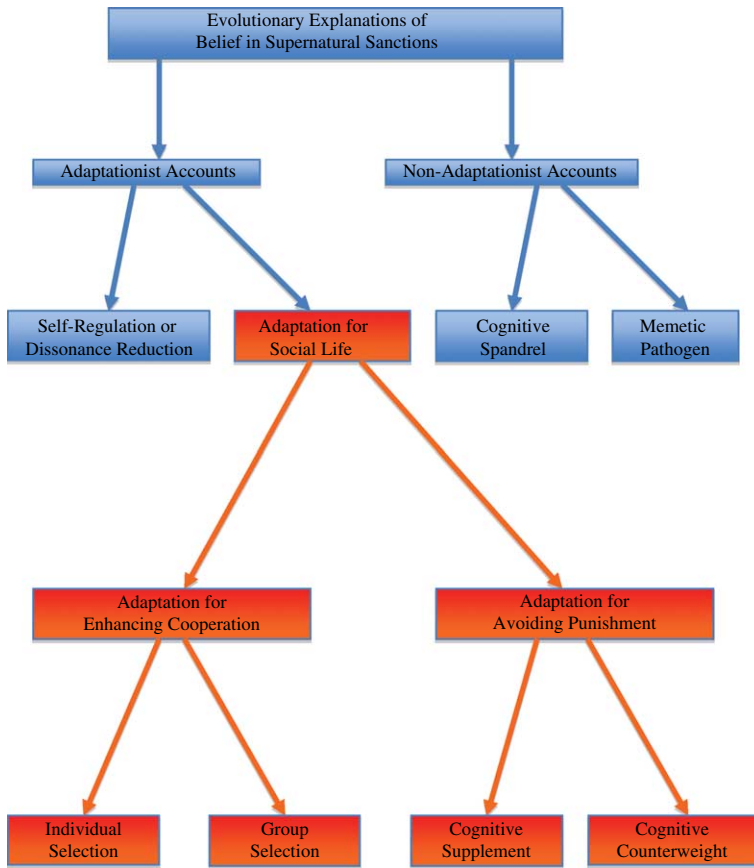


Figure 1. Decision tree for evolutionary accounts of one aspect of religious cognition.

adaptively salient social benefits, it is not yet clear that there is anything about religion in particular that accounts for those benefits. As we have seen, other beliefs and desires might, as far as one can tell, be equally suited to the task of securing the adaptive ends. Thus, it may well be that selective forces have conditioned general dispositions of cognition to facilitate the advantages of cooperation at individual and/or group scales, and major features of religion are structured or “domesticated,” as it were, by these dispositions, but so are the features of manifold other human enterprises from art to sports. In such an instance, evolution would not constitute an explanation of art or of sports or of religion (or of any particular aspect of religious cognition under consideration), but of the general nature of beliefs and practices that unite human groups, and of why religion – for whatever reason it exists – takes the forms it does. Here we are in need of more empirical work that distinguishes religious beliefs, general moral beliefs, and simple prudence in seeking cooperative behavior or in avoiding defection.

Second, even if religion can be shown to promote cooperation, this does not answer the question of whether it is an evolutionary adaptation. Indeed, if it fosters cooperative benevolence of the most altruistically sacrificial kind, it may in so doing contribute to human fulfillment while actually subverting reproductive optimization. In this scenario (and in others less morally benign), it would represent a memetic

pathogen that is maladaptive for the fitness of its hosts while being well-adapted to its own replication. To whatever extent religion may be associated with cooperation, explaining it in evolutionary as opposed to social–functional terms requires that we (a) know the connection to fitness and (b) understand the range and impacts of phenotypic variance in different environments.

Although there are a number of adaptationist proposals for religious belief in an afterlife or various forms of supernatural reward and punishment in terms of self-control, reduction of cognitive dissonance, or other homeostatic functions, we have focused on theories of supernatural sanctioning as having the most promise to both unify and raise questions about game theoretic, experimental, and ethnographic findings. However, this approach has several versions with differing and often wholly conflated explanatory programs – none of which are entirely concordant with existing data. The most fundamental distinction involves differing accounts of the adaptive challenge religion is purported to help solve: facilitating cooperation by supplementing cheater-detection mechanisms that are not fully adequate, or avoiding the severe costs of punishment by cheater-detection mechanisms that are highly efficacious. Empirical work that would help solve this issue entails comparisons of religious belief in punishment-intensive and non-intensive cultures. A historical issue that is highly relevant to this issue and that is not fully resolved is whether punishing gods, and in particular gods whose punishment is connected to social defections, are prevalent early in cultural development, or arise only in later, cosmopolitan societies where exchange takes place between individuals without prior interaction or without knowledge of each others' reputation.¹³

If religious belief in supernatural sanctions is construed as an adaptation for enhancing cooperation, an important question is whether this is an individual or group level adaptation. While in principle these are not mutually exclusive options and selection may operate concurrently on multiple levels (Wilson & Wilson, 2008), the two possibilities are readily associated with quite different and empirically distinguishable means of fostering cooperative benefit. Individual selection accounts have made prominent use of costly signaling theory, where the costs of a behavioral display purchase the actor entrance into a cooperative matrix from which compensatory benefits are received. If the benefits are fully compensatory, the behavior makes adaptive sense (though it would also make sense to pay the costs – if possible – without the dispositional commitments the behavior is taken to signal). However, although such costs may elevate relative fitness for the individual, they decrease average absolute fitness for the group. From the group's perspective, the ideal arrangement would be fully internalized cognitive disincentives to defect – which operate without costs of display, detection, or punishment even in situations where successful defection is likely. Religious belief in supernatural sanctions may provide something akin to this, and the closer beliefs truly function in this fashion, the more likely it would seem that group selection is necessary for their establishment. Although we know that community longevity and philanthropic giving (which can itself be viewed as a costly signal) are higher in religious than secular communities, one thing that would be helpful to know is whether anonymous giving differs between secular communities and religious communities with varying expectations of costly public behaviors.

Alternatively – but not always clearly presented as an alternative – belief in supernatural sanctions may entail an adaptation *to* rather than *for* an efficacious system of social control that stabilizes cooperation. This proposal presumes not the

inadequacy but the impressive adequacy of detection and punishment mechanisms, and posits the inadequacy (unaided by religion) of cognitive mechanisms for punishment avoidance. (This would be the case when less exotic mechanisms of risk assessment and reputation management, including “conscience” (Alexander, 1987), are not up to the task of accurately auditing risks to reputational capital.) One version proposes that relevant religious beliefs serve as supplements to more primitive cognitive mechanisms that do not reflect the special developments of human cheating detection. Why invoke, however, the cognitively costly and seemingly excessive deterrence of belief in all-knowing, all-powerful agents, rather than just improve accuracy or globally reduce confidence in the ability to defect? Another response is that over-confidence is not a failure to adapt that needs to be corrected, but is itself an adaptive cognitive bias that ought not (perhaps cannot) be corrected, thus requiring a cognitive counterweight or complementary bias that operates in specified situations where the risk of a mistake is severe and asymmetrical. These proposals entail different conceptions of the relationship between the disposition to form certain kinds of religious beliefs and other psychological features that may be empirically addressable; e.g., is there a relationship between such beliefs and confidence or risky behaviors within and outside the social domain?

Finally, we should affirm that while these accounts differ in how they construe religious adaptation, in how they comport with existing data, and in what they predict of hopefully yet-to-come empirical studies, they are by no means all mutually exclusive. Memetic, spandrel, and social adaptationist accounts are readily concordant in proposals that plausibly posit exaptation and perhaps even memetic drive (Atran, 2002; Schloss, 2009). Indeed, we think that although CE and PA entail different accounts of the adaptive challenges that described the origin of beliefs in supernatural sanctions, they are not mutually exclusive in the roles they might play in the persistence of these beliefs. With concurrently operative individual and group selection there could even be positive feedback between CE and PA. A promising option that one of us is currently investigating empirically involves the possibility that belief in both supernatural punishment and reward entails hard-to-fake though non-costly autonomic signals based on internalized affective rewards, thus meeting demands of both CE and PA approaches (Schloss, 2007; Bulbulia, 2009b).

Notes

1. The notion of “punishment” suffers from some terminological ambiguity, since it is employed in various ways in the game theoretic, experimental, and anthropological literatures. In some models, punishment by definition simply entails imposing a cost on another at some cost to the punisher, whereas imposing a cost on defection at no cost, e.g., by withholding cooperation, is “defection” (Rockenbach & Milinski, 2009; Ohtsuki et al., 2009). In other accounts punishment involves one party reducing the assets of another, and may be cost free, costly (incurring a cost to the punisher that may be compensated by future benefits), or altruistic (incurring an uncompensated cost). However in practice, even “cost-free” punishment entails costs of vigilance and may incur costs of reprisal when interactions are iterated, as they are in many human social settings. “Costly” punishment – even in games that afford no opportunity for future benefit via increased contributions of others to public goods – may still confer benefit through reputation (Dreber et al., 2008; Sigmund, Hauert, & Nowak, 2001).
2. In modeling punishment and defection under varying social norms that entail reputational benefits and consequences for cooperating, defecting, and punishing,

- Ohtsuki et al. (2009) find that costly punishment can be established, but only under a narrow range of conditions.
3. Although the agents and forces posited by religious belief are often taken to interact with and make behavioral demands on human communities, there is great variation in the nature of these demands, from local and idiosyncratic signals of commitment to more general expectations of cooperative fidelity.
 4. There is, of course, another way to avoid incurring the cost of punishment for defection: do not get caught. If what matters is avoiding the costs of detection and punishment, one can certainly avoid that by not cheating at all, but one might also be able to avoid it by cheating and not getting caught. Below we assess the crucial argument made by defenders of this view that avoiding defection is better than avoiding detection.
 5. There are extensive examples of this in the sacred writings, devotional literature, and analytic discourse of many religious traditions. For example, in the Hebrew scriptures, Psalm 71 begins as an anguished meditation over doubts arising from the temporal flourishing of the wicked. The author then resolves his doubts by affirming flourishing of the righteous and banishment of the iniquitous in the life to come.
 6. Strictly speaking, increasing the likelihood of being detected does not necessarily increase the likelihood of being punished, because some defections are either too trivial or costly to punish, and/or some defectors may escape punishment owing to social status or other factors. However, unless there is reason to posit that enhanced detection is biased toward unpunishable offenses, the first qualification is unimportant: for those defections that are punishable ($n > 0$), increased detection will result in increased costs to the defector. The second qualification is more complicated. On one read (and it is a read consistent with Nietzschean interpretations of religion), belief in SP might be an adaptation primarily or only for those in the social hierarchy most likely to be punished if detected. On another read, an important distinguishing trait of early human groups – in which some accounts propose both moral norms and religious sanctions to have emerged – involves balancing if not leveling dominance structures of primate sociality with a pan-human, egalitarian tendency to punish all defection (Boehm, 1999a,b). This would strengthen the connection between detection and cost of defecting.
 7. One possibility not developed in current proposals is that belief in SP might serve to overcome commitment barriers or cooperative stalemates in Prisoner's Dilemma type games where mutual defection is the Nash equilibrium. This is plausible, though it faces several questions. First, in social organisms where games are iterated it is clear, both theoretically and empirically, that strategies exist to enable cooperation without either punishment or cognition. Second, in situations where overcoming commitment barriers is advantageous – such as pair-bonding or social attachment – there exist affective mechanisms that both facilitate and opportunistically relax commitment (Fisher, Aron, & Brown, 2006; Kendrick, 2006). Third, and consistent with the previous point, although there are proposals for the role of belief – including fictive belief – in marshalling commitment, current theoretical and empirical warrant favors positive over negative illusions (McKay & Dennett, 2009).
 8. Byproduct or spandrel accounts of religious cognition are widely defended (Bloom, 2005; Boyer, 2001). The *prima facie* problem with viewing costly religious behaviors as spandrels rather than as adaptations is precisely that they do appear so costly. In order to be established as a spandrel, the fitness costs of the behavioral phenotype have to be compensated for by fitness benefits of another phenotype emerging from a shared cause. But it turns out that signaling theory requires an analogous cost–benefit balance, where the fitness costs of the signal must be compensated for by its resultant benefits. Thus, there is no *a priori* reason to consider one account more plausible. To resolve this empirically it is necessary to determine (a) actual fitness costs and benefits and (b) (a difficult task) the actual behavioral and dispositional phenotypes from which these fitness consequences accrue (Murray & Moore, 2009).
 9. The terms and the underlying concepts of “costly” and “hard-to-fake” or reliable signals reflect some ambiguity in their current employment. Contrary to some accounts, they are not isomorphic: not all costly signals are reliable and not all reliable signals are costly (Cronk, 2005). Even where hard-to-fake signals are autonomically mediated, they are not, contrary to appearances, always cost free. Indeed, some learning or belief-oriented

behaviors are hard-to-fake precisely in virtue of their costs. For example, linguistic dialects and accents, or sensitivity to culturally variable perceptions of style or humor, are notoriously difficult to fake and reflect the opportunity costs of early exposure. Some other behaviors with direct cognitive or affective costs – including the ability to withstand pain in certain contexts – may be facilitated by or even require, and therefore reliably signal, underlying beliefs. It may be that certain demanding religious rituals, while reflecting conscious choice, require underlying belief or authentic commitment in order to be effective (Sosis, 2004).

10. This assumes that the religious scaffolding necessary to sustain effective belief in supernatural punishment does incur fitness costs (though for a contrary reading on one apparently costly practice, see Larson, 2005). However, there is little in the way of firm data to support this claim (see Kotiaho, 2001). It is obvious that many systems of religious practice and ritual involve resource expenditures, but such expenditures are not equivalent to fitness costs. As a result, more work needs to be done to establish just what costs are involved and to what extent those costs bear on fitness. Searcy and Nowicki (2005) downplay the seriousness of this concern, however.
11. Temporal disequilibrium between environmental challenge and adaptive response is common to the evolutionary dynamics of all species, but may be much more important for human beings, since the genetic basis of adaptive phenotypes changes much more slowly than challenges posed by and the counter-adaptations generated by the cultural environment (Barkow, Cosmides, & Tooby, 1992; Plotkin, 1997, 2000).
12. It is important to note, however, that even if belief in supernatural punishment is more effective in deterring free-riding and managing reputation, the gains must be balanced against what may be significant fitness costs associated with the cultural scaffolding that appears to be required to sustain such a belief. See note 10 for a brief discussion of such costs.
13. This view assumes that in such societies the normal mechanisms of defection prevention would be less effective, though given claims that the move from egalitarian to hierarchical social organization may often typify larger societies, it is not clear that punishment mechanisms are less effective and more in need of supplementing.

References

- Alexander, R. (1987). *The biology of moral systems*. Piscataway, NJ: Aldine Transaction.
- Astuti, R. (2007). Ancestors and the afterlife. In H. Whitehouse & J. Laidlaw (Eds.), *Religion, anthropology, and cognitive science* (pp. 161–178). Durham, NC: Carolina Academic Press.
- Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion*. New York: Oxford University Press.
- Barkow, J., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.
- Barrett, J. (2004). *Why would anyone believe in God?*. Walnut Creek, CA: AltaMira Press.
- Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in real-world setting. *Biology Letters*, 22, 412–414.
- Bering, J.M. (2005). The evolutionary history of an illusion: Religious causal beliefs in children and adults. In B. Ellis & D.F. Bjorklund (Eds.), *Origins of the social mind: Evolutionary psychology and child development* (pp. 411–437). New York: Guilford Press.
- Bering, J.M., & Johnson, D.D.P. (2005). O Lord . . . you perceive my thoughts from afar': Recursiveness and the evolution of supernatural agency. *Journal of Cognition and Culture*, 5, 118–142.
- Bering, J.M., Blasi, C.H., & Bjorklund, D. F. (2005). The development of 'afterlife' beliefs in religiously and secularly schooled children. *British Journal of Developmental Psychology*, 23, 587–607.
- Bloom, P. (2005). *Is God an accident?* Boston, MA: Atlantic Monthly.
- Boehm, C. (1999a). *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.
- Boehm, C. (1999b). The natural selection of altruistic traits. *Human Nature*, 10, 205–252.
- Boehm, C. (2008). A biocultural evolutionary exploration of supernatural sanctioning. In J. Bulbulia, R. Sosis, E. Harris, R. Genet, C. Genet, & K. Wyman (Eds.), *Evolution of religion: Studies, theories, and critiques* (pp. 143–152). Santa Margarita, CA: Collins Foundation Press.
- Botelho, A., Harrison, G.W., Pinto, L.M.C., & Rutström, E.E. (2005). Social norms and social choice. Working Paper No. 05-23, Department of Economics, College of Business Administration, University of Central Florida.
- Boyer, P. (2001). *Religion explained*. New York: Basic Books.

- Brown, M., Palameta, B., & Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology, 1*, 42–69.
- Bulbulia, J. (2004). Religious costs as adaptations that signal altruistic intention. *Evolution and Cognition, 10*, 19–38.
- Bulbulia, J. (2009a). Charismatic signaling. *Journal of Religion and Culture, 3*, 518–551.
- Bulbulia, J. (2009b). Religiosity as mental time travel: cognitive adaptations for religious behavior. In J. Schloss & M. Murray (Eds.), *The believing primate: Scientific, philosophical and theological perspectives on the evolution of religion* (pp. 44–75). New York: Oxford University Press.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187–276.
- Cosmides, L., Tooby, J., Fiddick, L., & Bryant, G. (2005). Detecting cheaters. *Trends in Cognitive Sciences, 9*, 505–506.
- Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon: Journal of Religion and Science, 29*, 81–101.
- Cronk, L. (2005). The application of animal signaling theory to human phenomena: Some thoughts and clarifications. *Social Science Information/Information sur les Sciences Sociales, 44*, 603–662.
- De Quervain, D.J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254–1258.
- Dreber, A., Rand, D., Fudenberg, D., & Nowak, M. (2008). Winners don't punish. *Nature, 452*, 348–351.
- Dunbar, R.I. (1998). The social brain hypothesis. *Evolutionary Anthropology, 6*, 178–190.
- Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of Royal Society of London Series B, 275*, 871–8.
- Ermer, E., Cosmides, L., & Tooby, J. (2007). Cheater detection mechanism. In R.F. Baumeister & K.D. Vohs (Eds.), *Encyclopedia of social psychology* (pp. 138–140). Thousand Oaks, CA: Sage.
- Ermer, E., Guerin, S., Cosmides, L., Tooby, J., & Miller, M. (2006). Theory of mind broad and narrow: Reasoning about social exchange engages ToM areas, precautionary reasoning does not. *Social Neuroscience, 1*, 196–219.
- Fehr, E. (2004). Don't lose your reputation. *Nature, 432*, 449–450.
- Fehr, E., & Fischbacher, U. (2004). Third party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*, 980–994.
- Fehr, E., & Gächter, S. (2002). Egalitarian motive and altruistic punishment. *Nature, 415*, 137–140.
- Fehr, E., & Gächter, S. (2005). Fehr and Gächter reply to “Egalitarian motive and altruistic punishment.” *Nature, 433*, E1–E2.
- Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*, 137–140.
- Fehr, E., Fischbacher, U., & Kosfeld, M. (2005). Neuroeconomic foundations of trust and social preferences: Initial evidence. *American Economic Review, 95*, 346–351.
- Fisher, H., Aron, A., & Brown, L.L. (2006). Romantic love: A mammalian brain system for mate choice. *Philosophical Transactions of the Royal Society: Biological Sciences, 361*, 2173–2186.
- Fowler, J.H. (2005a). Second-order free riding problem solved? *Nature, 437*, E8.
- Fowler, J.H. (2005b). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences, 102*, 7047–7049.
- Fowler, J.H., Johnson, T., & Smirnow, O. (2005). Egalitarian motive and altruistic punishment. *Nature, 433*, E1.
- Gervais, M., & Wilson, D.S. (2005). The evolution and functions of laughter and humor: A synthetic approach. *Quarterly Review of Biology, 80*, 395–430.
- Haley, K.J., & Fessler, D.M.T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior, 26*, 245–256.
- Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.
- Haselton, M.G. (2007). Error management theory. In R.F. Baumeister & K.D. Vohs (Eds.), *Encyclopedia of social psychology* (pp. 311–312). Thousand Oaks, CA: Sage.
- Haselton, M.G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10*, 47–66.
- Humphrey, N.K. (1992). *A history of the mind: Evolution and the birth of consciousness*. New York: Simon & Schuster.
- Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In R. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 292–309). New York: Russell Sage Foundation.
- Johnson, D., Stopka, P., & Knight, S. (2003). The puzzle of human cooperation. *Nature, 421*, 911–912.
- Johnson, D.D.P. (2004). *Overconfidence and war: The havoc and glory of positive illusions*. Cambridge, MA: Harvard University Press.
- Johnson, D.D.P. (2005). God's punishment and public goods: A test of the supernatural punishment hypothesis in 186 world cultures. *Human Nature, 16*, 410–446.

- Johnson, D.D.P. (2009). The error of God: Error management theory, religion, and the evolution of cooperation. In S.A. Levin (Ed.), *Games, groups, and the global good* (pp. 169–180). London: Springer.
- Johnson, D.D.P., & Bering, J. (2009). Hand of God, mind of man: punishment and cognition in the evolution of cooperation. In J. Schloss & M. Murray (Eds.), *The believing primate: Scientific, philosophical, and theological reflections on the origin of religion* (pp. 26–43). Oxford: Oxford University Press.
- Johnson, D.D.P., & Krüger, O. (2004). The good of wrath: Supernatural punishment and the evolution of cooperation. *Political Theology*, 5, 159–176.
- Johnson, D.D.P., McDermott, R., Barrett, E., Cowden, J., Wrangham, R., McIntyre, M., & Rosen, S. (2006). Overconfidence in wargames: experimental evidence on expectations, aggression, gender and testosterone. *Proceedings of the Royal Society of London Series B*, 273, 2513–2520.
- Kagan, J. (1998). *Three seductive ideas*. Cambridge, MA: Harvard University Press.
- Kendrick, K. (2006). The neurobiology of social recognition, attraction and bonding. *Philosophical Transactions of the Royal Society: Biological Sciences*, 361, 2057–2059.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314, 829–832.
- Kotiaho, J.S. (2001). Cost of sexual traits: A mismatch between theoretical considerations and empirical evidence. *Biological reviews of the Cambridge Philosophical Society*, 76, 365–376.
- Larson, D. (2005). Have faith: Religion can heal mental ills. In J. Levin & H. Koenig (Eds.), *Faith, medicine, and science: A festschrift in honor of Dr. David B. Larson* (pp. 215–222). New York: Routledge.
- Maynard Smith, J., & Szathmáry, E. (2001). *The major transitions in evolution*. Oxford: Oxford University Press.
- McKay, R.T., & Dennett, D.C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–561.
- Michod, R.E. (1997). Evolution of the individual. *The American Naturalist*, 150, S5–S21.
- Michod, R.E. (2007). Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104, 8613–8618.
- Michod, R.E., & Herron, M.D. (2006). Cooperation and conflict during evolutionary transitions in individuality. *European Society for Evolutionary Biology*, 19, 1406–1409.
- Milinski, M., & Rockenbach, B. (2008). Punisher pays. *Nature*, 452, 297–298.
- Murdock, G.P. (1980). *Theories of illness: A world survey*. Pittsburgh, PA: HRAF, University of Pittsburgh Press.
- Murray, M.J. (2009). The evolution of religion: Adaptationist accounts. In S. Melville (Ed.), *Science and religion in dialogue* (pp. 439–457). Malden, MA: Wiley–Blackwell.
- Murray, M.J., & Moore, L. (2009). Costly signaling and the origin of religion. *Journal of Cognition and Culture*, 9, 225–245.
- Norenzayan, A., & Shariff, F. (2008). The origin and evolution of religious prosociality. *Science*, 322, 58–62.
- Nowak, M.A. (2006). Five rules for the evolution of cooperation. *Science*, 314, 1560–1563.
- Nowak, M.A., & Sigmund, K. (2005). Evolution and indirect reciprocity. *Nature*, 437, 1291–1298.
- Ohtsuki, H., Iwasa, Y., & Nowak, M.A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, 457, 79–82.
- Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*, 86, 404–417.
- Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: reciprocity, mimicry and efficiency. *The Economic Journal*, 115, 1032–1053.
- Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432, 499–502.
- Pargament, K.I. (1997). *The psychology of religion and coping: Theory, research, and practice*. New York: The Guilford Press.
- Plotkin, H. (1997). *Darwin machines and the nature of knowledge*. Cambridge, MA: Harvard University Press.
- Plotkin, H. (2000). *Evolution in mind*. Cambridge, MA: Harvard University Press.
- Price, M.E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, 23, 203–231.
- Pronin, E., & Kugler, M.B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, 43, 565.
- Randolph-Seng, B., & Nielsen, M.E. (2007). Honesty: One effect of primed religious representations. *The International Journal for the Psychology of Religion*, 17, 303–315.
- Randolph-Seng, B., & Nielsen, M.E. (2008). Is God really watching you? A response to Shariff and Norenzayan (2007). *The International Journal for the Psychology of Religion*, 18, 119–122.
- Rappaport, R.A. (1999). *Ritual and religion in the making of humanity*. New York: Cambridge University Press.

- Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*, 718–723.
- Rockenbach, B., & Milinski, M. (2009). How to treat those of ill repute. *Nature*, *457*, 39–40.
- Ruffle, B., & Sosis, R. (2006). Cooperation and the in-group-out-group bias: A field test on Israeli kibbutz members and city residents. *Journal of Economic Behavior & Organization*, *60*, 147–163.
- Ruffle, B., & Sosis, R. (2007). Does it pay to pray? Costly ritual and cooperation. *The B.E. Journal of Economic Analysis and Policy*, *7*, 1–35.
- Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., & Cohen, J.D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–1758.
- Schloss, J.P. (2007). He who laughs best: Religious affect as a solution to recursive cooperative defection. In J. Bulbulia, R. Sosis, E. Harris, R. Genet, C. Genet, & K. Wyman (Eds.), *The evolution of religion: Studies, theories, critiques* (pp. 205–215). Santa Margarita, CA: Collins Foundation Press.
- Schloss, J. (2009). Evolutionary theories of religion: Science set free or naturalism run wild? In J. Schloss & M. Murray (Eds.), *The believing primate: Scientific, philosophical, and theological perspectives on the origin of religion* (pp. 1–15). Oxford: Oxford University Press.
- Searcy, W.A., & Nowicki, S. (2005). *The evolution of animal communication: Reliability and deception in signaling systems*. Princeton, NJ: Princeton University Press.
- Sefton, M., Shupp, R., & Walker, J.M. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, *45*, 671–690.
- Shariff, A., & Norenzayan, A. (2007). God is watching you: Supernatural agent concepts increase prosocial behavior in an anonymous economic game. *Psychological Science*, *18*, 803–809.
- Sigmund, K., Hauert, C., & Nowak, M.A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences*, *98*, 10757–10762.
- Sosis, R. (2003). Why aren't we all Hutterites? Costly signaling theory and religious behavior. *Human Nature*, *14*, 91–127.
- Sosis, R. (2004). The adaptive value of religious ritual: Rituals promote group cohesion by requiring members to engage in behavior that is too costly to fake. *American Scientists*, *92*, 166–174.
- Sosis, R. (2005). Does religion promote trust? The role of signaling, reputation, and punishment. *Interdisciplinary Journal of Research on Religion*, *1*, 1–30.
- Sosis, R., & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology*, *12*, 264–274.
- Swanson, G.E. (1960). *The birth of the gods*. Ann Arbor: University of Michigan.
- Wilson, D.S. (2002). *Darwin's cathedral*. Chicago, IL: University of Chicago Press.
- Wilson, D.S. (2005). Testing major evolutionary hypotheses about religion with a random sample. *Human Nature*, *16*, 419–446.
- Wilson, D.S., & O'Gorman, R. (2003). Emotion and actions associated with norm breaking events. *Human Nature*, *14*, 277–304.
- Wilson, D.S., & Wilson, E.O. (2008). Rethinking the foundations of sociobiology. *The Quarterly Review of Biology*, *82*, 327.

COMMENTARIES

Affording cooperative populations

Joseph Bulbulia* and Marcus Frean

Victoria University, New Zealand

Free-riding

Much of the kerfuffle about cooperation's evolution has centred on the problem of free-riding, according to which “The prospect of defection without loss of reward provides powerful incentives . . . to free ride on the efforts of others” (p. 47). Schloss and Murray (S&M) consider two adaptationist explanations for religion as solutions to free-riding, finding neither wholly satisfactory.

*Corresponding author. Email: joseph.bulbulia@vuw.ac.nz

The supernatural punishment hypothesis holds that religious inhibition avoids social retaliation, thus benefiting believers. S&M notice that retaliation presupposes the cooperation of the retaliators, so belief in supernatural punishment alone cannot address free-riding. We agree.

Commitment signalling

S&M also consider commitment-signalling models, according to which religion both expresses and authenticates cooperative commitments, enabling religious partners to assort. The authors worry about whether signalling mechanisms are possible or needed. We do not agree.

Against the possibility of commitment signalling, S&M notice that a regress of higher-order defection problems arises from the advantages of signalling without cooperative commitment. However, the authors also notice that a regress may be stopped from signals that reliably identify unwavering cooperative traits. The regress presents an example of the general problem of signal reliability. Solutions, however, are commonplace. The peacock's tail, for example, permits no hypocrisy. Nature breeds much flamboyant honesty (Zahavi & Zahavi, 1997).

S&M worry about the efficiency of religious costs as signalling devices, yet religious signals may be both cheap and reliable – as they also concede. Better that naturalists use “hard-to-fake” signals, in Iron's original sense, which avoids the murkiness of “costly” (Irons, 2001).

Moreover, there are significant virtues of signalling theory that S&M do not discuss, notably its accommodation of the supernatural punishment hypothesis. Fear of punishment may curtail anti-social behavior (tick motivation); the suppression of fear is notoriously hard-to-fake (tick index).

Signalling theory also generalizes to mechanisms other than punishment, including the intrinsic love felt for gods (Bulbulia, 2004), the effects of self-signalling and dissonance (Sosis, 2003) and the effects on cooperation of symbolic markers (Boehm, 1999). Indeed, as Sosis shows, permanent marking may pre-commit partners to cooperation irrespective of belief (Sosis, Kress & Boster, 2007). Life becomes strenuous when one is forever branded with losing symbols. Commitment signalling has its limitations, however.

Risky coordination

Free-riding implies a reliable gain from defection, but what assures this reliability? In large, variable and dynamic social worlds, a more fundamental threat comes from a focal partner's inability to predict cooperative outcomes – that is, from uncertainty.

Consider a simple game (see Figure 1). We flip a coin and select “heads or tails.” Successful coordination WINS BIG; failure LOSES BIG.¹ Suppose there is a third option – “don't play” – which invariably WINS SMALL. Notice, while partners cannot do better than by cooperating successfully their defection may still be motivated from uncertainty and risk-avoidance.² Mutual defection is a cooperative equilibrium: one can do no better than by defecting when another defects. So how can we predict another's cooperation?³

The threat to cooperation from uncertainty is commonplace (Bicchieri, 2006) and it is not always solved (Ostrom, 2005). Yet global uncertainty is a problem for which religious cognition and culture may bring substantial relief.

YOU/THEM	Heads	Tails	Defect
Heads	WIN BIG	LOSE BIG	LOSE BIG/ WIN SMALL
Tails	LOSE BIG	WIN BIG	LOSE BIG/ WIN SMALL
Defect	WIN SMALL/ LOSE BIG	WIN SMALL/ LOSE BIG	WIN SMALL/ WIN SMALL

Figure 1. Cooperation threatened by uncertainty, not free-riding.

Ecological signalling

Natural and cultural selection operate on populations through systems that bend coins, those we flip when acting together for collective advantage, and those that flip in us, when deciding whether to bother with cooperation in the first place.

Ecological signalling conjectures that religious cognition and cultures co-evolve as technologies that bend the internal coins of partners to motivate cooperative behaviors over its doubts (see Figure 2). Differences between ecological signalling systems and commitment signalling systems may be predicted from their distinct functional targets: ecological signals evolve to express and synchronize the cooperative tendencies of populations against defection from uncertainty. They do not evolve for the assessment of partner-specific virtues.

Religious cultures and minds appear nicely co-adapted both to the synchronous *suppression* of strategic rationality, and to the synchronous *expression* of durable cooperative goals at large social scales. This is so because religions function to align the material, emotional and imagined landscapes in which cooperation occurs, generically, from pervasive exposures.

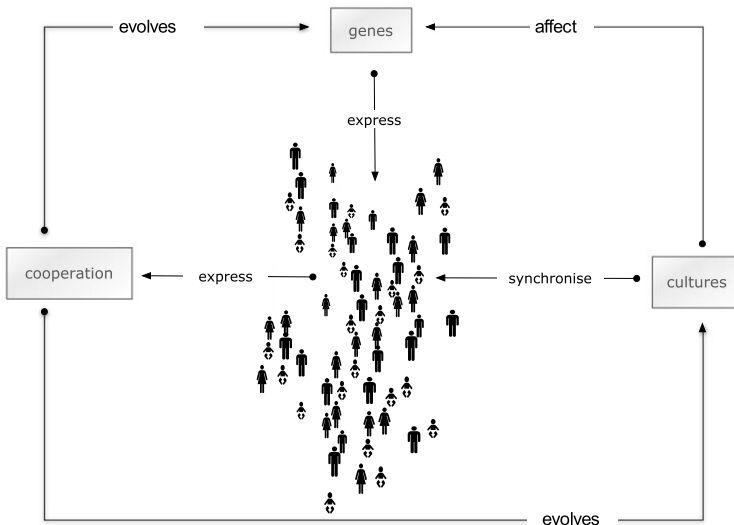


Figure 2. Evolution of cooperative populations.

Ecological signalling makes sense of lingering puzzles in the data on religious cooperation, including:

- (1) The cooperative effects of religious cues among the disbelieving (Mazar, Amir & Ariely, 2008);
- (2) The sensitivity to context of cooperation among the religiously committed (Darley & Batson, 1973; Malhotra, 2010);
- (3) The strong alignment of cognitive states (both declarative and “embodied”) to ritual events in anonymous populations (Konvalinka et al., 2011; Xygalatas et al., 2011).

The manifest alignment of cognitive responses to sacred targets is consistent with the demands of uncertain exchange, but not with those of free-riding. Thus, while commitment signalling remains important, naturalists have largely ignored the problem of defection from uncertainty. Preliminary data suggest that religious ecologies express especially powerful cooperative motives somewhat automatically across large populations, by intricate, diverse and subtle means, which often suppress strategic vigilance (Bulbulia, 2009). Naturalists are only beginning to understand these manifold technologies for collaborative success (Sosis, 2005).

Notes

1. Assuming absolute benefits are desired. We could reconfigure the game with relative benefits, without loss of generality.
2. There are three pure equilibriums, two for cooperation and one for defection.
3. The problem deepens when we consider how uncertainty affects the utility functions of the needy for whom losses matter more than gains, and how the representation of doubts in others may dash confidence for those who are not themselves needy. For again uncertain cooperation’s problem comes from an inability to forecast what others will do. While gambling for cooperation in the game we have imagined is rational wherever half the expected yield of success divided by its failure exceeds the expected yield of defection, nearly all of these variables are too poorly defined in natural human ecologies for selection to entrench a rule. We cannot assign gains and losses in fitness, and it may be difficult to predict how others with varying degrees of risk-sensitivity, information, cooperative habits, etc., will respond.

References

- Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.
- Boehm, C. (1999). *Hierarchy in the forest*. Cambridge, MA: Harvard University Press.
- Bulbulia, J. (2004). Religious costs as adaptations that signal altruistic intention. *Evolution and Cognition*, 10(1), 19–38.
- Bulbulia, J. (2009). Charismatic signalling. *Journal for the Study of Religion*, 3(4), 518–551.
- Darley, J.M., & Batson, C.D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1), 100–108.
- Irons, W. (2001). Religion as hard-to-fake sign of commitment. In R. Nesse (Ed.), *Evolution and the Capacity for Commitment* (pp. 292–309). New York: Russell Sage Foundation.
- Konvalinka, I., Xygalatas, D., Schjoedt, U., Bulbulia, J., Jegindoe, E., Geertz, A., et al. (2011). Synchronous correlates of arousal and emotion in a Spanish fire walk ritual. Manuscript submitted for publication.
- Malhotra, D. (2010). (When) are religious people nicer? Religious salience and the “Sunday effect” on pro-social behavior. *Judgment and Decision Making*, 5(2), 138–143.
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633–644.

- Ostrom, E. (2005). *Understanding institutional diversity*. Princeton, NJ: Princeton University Press.
- Sosis, R. (2003). Why aren't we all Hutterites? *Human Nature*, 14(2), 91–127.
- Sosis, R. (2005). Does religion promote trust? The role of signaling, reputation, and punishment. *Interdisciplinary Journal of Research on Religion*, 1(1), 1–30.
- Sosis, R., Kress, H., & Boster, J. (2007). Scars for war. *Evolution and Human Behavior* (28), 234–247.
- Xygalatas, D., Schjoedt, U., Bulbulia, J., Konvalinka, I., Reddish, P., Jegindoe, E., et al. (2011). Flashbulb memory in a Spanish firewalk ritual. Manuscript submitted for publication.
- Zahavi, A., & Zahavi, A. (1997). *The handicap principle: A missing piece of Darwin's puzzle*. New York: Oxford University Press.

Broadening the critical perspective on supernatural punishment theories

Emma Cohen*

Max Planck Institute for Psycholinguistics, Netherlands, and Max Planck Institute for Evolutionary Anthropology, Germany

According to supernatural punishment accounts, behaviors mediated by beliefs about moralizing, punishing supernatural agents confer fitness advantages on their bearers. Specifically, beliefs in supernatural observers and punishers serve to maintain cooperation by inhibiting defection and/or reducing the costs of punishment of defection.

In any evolutionary biological account of behavior, it is important to distinguish the specific operational details of the mechanisms underpinning the purportedly adaptive behavior from the fitness consequences that ultimately led to its selection in the population (Scott-Phillips, Dickins & West, 2011). The wide-ranging critical review by Schloss and Murray (S&M) suggests that theoretical and empirical detail is considerably lacking on both issues in current adaptationist accounts of the evolution of belief in supernatural punishment. A key contribution is S&M's clarification of the discrepancy between the low cost of acquiring and transmitting the religious belief relative to the costs of putting the belief to adaptive work.

Beliefs in moralizing gods have perhaps been most parsimoniously accounted for as by-products of psychological mechanisms that evolved to solve other adaptive problems having to do with other minds, fairness, teleology and reputation management (Barrett, 2004; Boyer, 2001; Cohen, 2007; Kelemen, 2004). The cognitive cheapness of the belief possibly contributes to the *prima facie* appeal of supernatural punishment theories of cooperation, but the early acquisition in ontogeny and widespread incidence in culture of a by-product belief do not speak to the question of whether the belief reliably guides and maintains cooperation or of whether it sustained a system of punishment that provided inclusive fitness benefits to its bearers and thus was selected for in human evolution. S&M's elucidation of the many costs and conditions entailed in an adaptationist scenario, whether at the level of the group or the individual, raises important questions about the evolutionary viability of belief in supernatural punishment as an explanation for cooperation.

By focusing their review on details and unresolved difficulties of supernatural punishment theories, S&M have helped move these theories toward a point from

*Email: emma_cohen@eva.mpg.de

which we can systematically consider their value and plausibility. There is still considerable work required before we can assess the likely evolutionary fitness costs and benefits of any belief-in-supernatural-punishment adaptation across variable ecological and population-structure environments, and the plausibility of these accounts against competing evolutionary theories. Nevertheless, it seems sensible and prudent that both theorists and critics consider, from the outset, the theoretical plausibility and parsimony of supernatural punishment accounts not only against a scenario in which beliefs in moralizing supernatural entities (or even punishment) are absent, but in light of the broader literature on cooperation. Do we really need supernatural agents? Three general observations are potentially relevant.

(1) *Other (godless) species have punishment*

A wider comparative view on the punishment solutions reached by other organisms, including cleaner fish and ants, bees and wasps, reminds us that the problem of punishment is not exclusively a human dilemma (e.g., Bshary & Grutter, 2005; Ratnieks & Visscher, 1989). Theoretical treatments of the apparent enforcement of cooperation via policing, sanctioning and ostracism across a range of organisms have straightforwardly explained these behaviors in terms of net direct and indirect benefits (e.g., Young & Clutton-Brock, 2006). If one's goal is to account for the evolution of punishment in humans (rather than account for apparent punishment strategies of participants in laboratory games, or cross-cultural patterns of spread of beliefs about moralizing, supernatural entities), a comparative perspective on the mechanisms underpinning punishment in non-humans is a potentially rich source of information and inspiration.

(2) *Costly punishment is rare; alternative evolved mechanisms may suffice*

The cross-cultural ubiquity of (particularly third party) costly punishment has come under question from critical ethnographic reviews and experimental games (e.g., Baumard, in press; Marlowe, 2009). Although there is no doubt that punishment can enhance cooperation, other available options are frequently exploited and provide cost-effective solutions to the problem or threat of defection. These mechanisms include partner-choice and partner-switching, for example, whereby individuals heighten their chances of assorting with the cooperation of others and not with defection. Where assortment breaks down, punitive behaviors such as ostracism or sanctioning may potentially be explained in terms of fairness restoration or reputation preservation (and therefore direct or indirect benefits). Such enforcement mechanisms are required only inasmuch as initial assortment mechanisms are vulnerable to failure. The theoretical necessity of punishment adaptations should therefore be considered in light of the broader ethnographic literature suggesting that costly (third-party) punishment is rare and also in light of the reliability of human assortment mechanisms (such as signalling and reputation) to sustain the successful assortment of cooperators (see Fletcher & Doebeli, 2009).

(3) *Punishment is about fairness, but fair-god beliefs are vulnerable to unstable commitment*

Relevant related research further suggests that punishment is primarily a response to violations of fairness (e.g., Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009). It is not obviously the case that cross-culturally recurrent religious prescriptions, and the concerns of gods, affirm human intuitions about fairness, or that moralizing god beliefs are sufficiently stable to form a robust foundation upon which punishment could evolve. Many gods cross-culturally operate on a non-negotiable hair-trigger response mode to infractions of rules that apparently have nothing to do with fairness, harm, or cooperation. Of course, supernatural punishment accounts do not predict that all gods everywhere will be primarily concerned with human morality. Where it does exist, however, religious commitment to a god who is supposed to be concerned with fairness is especially vulnerable to defection simply by virtue of the fact that the god should reward and punish fairly. The Biblical consolation that “you do not know the works of God who makes everything” is apparently insufficient to sustain the commitment of many “third order defectors” whose god disappointingly turned out to be unfair and unjust by intuitive logic. S&M suggest that this is not a disabling problem for supernatural punishment accounts, but it is not at all clear that commitment to a belief in a fair and punishing god is sufficiently stable to enhance cooperation, or that the costs of theological scaffolding required to sustain these beliefs are viable (against benefits, or against alternative mechanisms). Existing adaptationist accounts demonstrating the sufficiency of standard mechanisms premised on fitness costs and benefits associated with cooperation, forgiveness and (human) punishment merit more careful, and prior, consideration (e.g., Gardner & West, 2004).

References

- Barrett, J. (2004). *Why would anyone believe in God?*. Walnut Creek, CA: AltaMira Press.
- Baumard, N. (in press). Punishment is not a group adaptation. *Mind & Society*.
- Boyer, P. (2001). *Religion explained: the evolutionary origins of religious thought*. New York: Basic Books.
- Bshary, R., & Grutter, A.S. (2005). Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. *Biology Letters*, 1(4), 396–399.
- Cohen, E. (2007). *The mind possessed: the cognition of spirit possession in an Afro-Brazilian religious tradition*. New York: Oxford University Press.
- Fletcher, J.A., & Doebeli, M. (2009). A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society B*, 276, 13–19.
- Gardner, A., & West, S. (2004). Cooperation and punishment, especially in humans. *The American Naturalist*, 164(6), 753–764.
- Johnson, T., Dawes, C.T., Fowler, J., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, 102, 192–194.
- Kelemen, D. (2004). Are children “intuitive theists”? Reasoning about purpose and design in nature. *Psychological Science*, 15(5), 295–301.
- Marlowe, F.W. (2009). Hadza cooperation: Second party punishment, yes: third party punishment, no. *Human Nature*, 20, 417–430.
- Ratnieks, F.L.W., & Visscher, P.K. (1989). Worker policing in the honeybee. *Nature*, 342, 796–797.
- Scott-Phillips, T., Dickins, T.E., & West, S.A. (2011). Evolutionary theory and the ultimate/proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6(1), 38–47.
- Young, A.J., & Clutton-Brock, T.H. (2006). Infanticide by subordinates influences reproductive sharing in cooperatively breeding meerkats. *Biology Letters*, 2(3), 385–387.

Stratification and supernatural punishment: cooperation or obedience?

Rolando de Aguiar* and Lee Cronk

Rutgers University, USA

Schloss and Murray (S&M) have provided an insightful and important contribution to our understanding of the role of supernatural punishment in the evolution of religious systems. Future researchers will need to pay particular attention to their refinements of the cooperation enhancement (CE) and punishment avoidance (PA) approaches. While S&M acknowledge “considerable empirical support that . . . belief in supernatural sanctions [is associated with] recent, cosmopolitan religions” (p. 57) their approach could be further refined through greater attention to the role of social, economic and political stratification in the shaping of religious doctrine (Cronk, 1994). We argue that stratification and hierarchy are the critical elements in producing a cognitive ecology and social structure in which punishing gods can thrive.

Humans are famously obedient to authority (e.g., Milgram, 1963), and there is a great deal of empirical evidence that males in particular possess cognitive adaptations to assess dominance status and modify behavior accordingly. A number of visible traits including stature (Hensley, 1993), eye color (Kleisnera, Kočnara, Rubešová, & Flegra, 2010), and facial structure (Mueller & Mazur, 1996) have been shown to signal dominance, and humans seem to use auditory clues as well. Subordinate men, for instance, unconsciously adjust their vocal pitch to that of a dominant conversation partner (Gregory & Webster, 1996; see also Gregory & Gallagher, 2002; Puts, Gaulin, & Verdolini, 2006; Puts, Hodges, Cárdenas, & Gaulin, 2007). Thus, humans seem to have a number of psychological adaptations that allow us to perceive and navigate status hierarchies effectively.

Hierarchy and stratification are important but not ubiquitous in human societies (Dubreuil, 2010), and there is considerable variation even among types of societies that are often painted with broad strokes. For example, hunter-gatherers are frequently labeled as egalitarian, but many such groups include some stratification (Kelly, 1995). Status differences based upon sex and age are particularly common (e.g., Hart and Pilling, 1979). Stratification may have its greatest incidence in larger human societies, but its seeds are present even among the smallest, most homogeneous groups.

Stratification is maintained through mechanisms of social control. Coercion is one obvious way to maintain control, but it can be costly. Manipulation through the use of signals is often a less costly and less risky alternative. As Schloss and Murray (2011) note, judgmental gods and judgment-based afterlife beliefs are not universal. Considerable evidence exists that such beliefs are rare among hunter-gatherer, small-scale, and egalitarian societies, and common among food producing, large-scale, and hierarchical societies. Swanson (1960) may have been the first to note an association between stratification and the belief in supernatural powers that reward and punish individuals according to how well they behave (see also Peregrine, 1996). Similarly, Roes and Raymond (2003) found an association between social complexity and the belief in moralizing gods. Most recently, Dickson, Olsen, Dahm, and Wachtel (2005) found an association between subsistence type (a common proxy for degree of

*Corresponding author. Email: rdeaguiar.anth@gmail.com

stratification) and the belief that the quality of one's experience in the next life is contingent upon how one behaves in this one. While only 10% of food collecting societies maintain such beliefs, nearly 90% of plow agricultural societies have them. As societies become more socially, economically, and politically stratified, punitive, judgmental gods and judgmental afterlife beliefs become much more common.

Hierarchies can also serve to protect individuals from those lower in rank. If a worker objects to something her boss is telling her to do, the boss can always appeal to the hierarchy: "I, too, am just following orders." When the top of the hierarchy is occupied by a capricious, omniscient, incorporeal being whose primary concern is obedience, a ruler's accountability is reduced even further. By enforcing the divinely prescribed order of things, the ruler is merely doing his or her job.

The hierarchical approach creates a framework in which the CE and PA approaches can be seen as working together. The CE viewpoint suggests that the threat of supernatural punishment enhances cooperation among all members of religious groups. An unstated assumption is that this cooperation benefits all participants. While the hierarchical perspective does not contradict that argument, it suggests that costs and benefits may be distributed unequally – those nearer the top of the hierarchy may benefit much more than those at the bottom. The PA account suggests that individuals subscribe to beliefs that include supernatural punishment in order to avoid real world punishment. In the hierarchical view, elites are using the threat of supernatural punishment as an inexpensive means of encouraging non-elites to follow the rules, but real-world punishment is, of course, a fallback option.

One of the predictions of the hierarchical perspective has already been supported: there is indeed a cross-cultural association between social stratification and belief in judgmental gods. We also predict a relationship at the individual level between the degree to which people believe in the hierarchical system and the strength of their beliefs in supernatural punishment.

References

- Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon*, 29(1), 81–101.
- Dickson, D.B., Olsen, J., Dahm, P.F., & Wachtel, M.S. (2005). Where do you go when you die? A cross-cultural test of the hypothesis that infrastructure predicts individual eschatology. *Journal of Anthropological Research*, 61(1), 53–79.
- Dubreuil, B. (2010). *Human evolution and the origins of hierarchies: The state of nature*. Cambridge: Cambridge University Press.
- Gregory, S.W., Jr., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology*, 70(6), 1231–1240.
- Gregory, S.W., Jr., & Gallagher, T.J. (2002). Spectral analysis of candidates' nonverbal vocal communication: Predicting U.S. presidential election outcomes. *Social Psychology Quarterly*, 65(3), 298–308.
- Hart, C.W.M., & Pilling, A.R. (1979). *Tiwi of North Australia*. New York: Holt, Rinehart & Winston.
- Hensley, W.E. (1993). Height as a measure of success in academe. *Psychology*, 30(1), 40–46.
- Kelly, R.L. (1995). *The foraging spectrum: Diversity in hunter-gatherer lifeways*. Washington, DC: Smithsonian Institution Press.
- Kleisnera, K., Kočnara, T., Rubešová, A., & Flegra, J. (2010). Eye color predicts but does not directly influence perceived dominance in men. *Personality and Individual Differences*, 49(1), 59–64.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.
- Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces*, 74(3), 823–850.
- Peregrine, P. (1996). *The Birth of the Gods* revisited: A partial replication of Guy Swanson's (1960) cross-cultural study of religion. *Cross-Cultural Research*, 30(1), 84–112.

- Puts, D.A., Gaulin, S.J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, 27(4), 283–296.
- Puts, D.A., Hodges, C.R., Cárdenas, R.A., & Gaulin, S.J. (2007). Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, 28, 340–344.
- Roes, F.L., & Raymond, M. (2003). Belief in moralizing gods. *Evolution and Human Behavior*, 24, 126–135.
- Swanson, G.E. (1960). *The birth of the gods: The origin of primitive beliefs*. Ann Arbor: University of Michigan Press.

Supernatural punishment: what traits are being selected?

Helen De Cruz^{a*} and Johan De Smedt^b

^aResearch Foundation Flanders, Centre for Logic and Analytical Philosophy, Katholieke Universiteit Leuven, Belgium; ^bDepartment of Philosophy and Ethics, Ghent University

In their critical analysis of supernatural punishment (SP) theories, Schloss and Murray (S&M) tease apart two distinct but often conflated adaptationist approaches to religion: those that argue that religious belief enhances cooperation (cooperation enhancement, CE), and those that say that it helps people to withstand the temptation to cheat, helping them avoid the costs of being punished (punishment avoidance, PA). They also make a distinction between individual selection and group selection. However, they pay little attention to the traits that are the targets of selection: evolved psychological dispositions, flexible behavioral strategies, or culturally transmitted norms. These three types of traits roughly correspond to three styles of evolutionary approach to human behavior: evolutionary psychology, behavioral ecology, and dual inheritance theories (Smith, 2000). Each has a different expected temporal scale in which adaptive change takes place – this constrains the plausibility of particular hypotheses and their mutual compatibility.

Evolutionary psychologists explain our behavioral repertoire as a result of evolved psychological adaptations, which were shaped in ancestral environments. Let us examine PA accounts from this perspective. If belief in SP is an ancient adaptation, we would expect it to have emerged in small, egalitarian groups, where there is a tendency to punish all defection. The strong connection between detection and defection costs in such societies is said to favor the evolution of God-fearing psychological mechanisms. However, the purported egalitarianism of ancient human groups is to some extent an idealization, because evidence for social inequality, in the form of lavish beadwork in children's burials, dates back at least to the Upper Paleolithic (Vanhaeren & d'Errico, 2005). It is unlikely that in such societies everyone would have faced the same risk of being punished. Also, ethnographic parallels show that at least some small-scale egalitarian societies (e.g., Ju/'hoansi, Kalahari hunter-gatherers) practice mainly *low-cost* forms of punishment, including gossip and jokes at the expense of the offender (Wiessner, 2005). The low cost of punishment (for both offender and punisher) seems incompatible with PA. As Wiessner (2005, p. 135) says: “[this] informal means of punishment would be ineffective or insufficient in a larger-scale society with less mobility.” Taken together, it seems unlikely that PA emerged in ancestral environments. For the evolutionary psychologist, CE thus remains the only viable option, but it then still remains unclear, as S&M aptly point out, why religion, rather than other forms of

*Corresponding author. Email: Helen.DeCruz@hiw.kuleuven.be

group-level activities, would enhance cooperation. Indeed, the plenitude of Upper Palaeolithic mobiliary art, such as Magdalenian Venus figurines, and their geographic clustering in distinct styles suggest that art may have been used to affirm group identity and group commitment (De Smedt & De Cruz, in press).

Behavioral ecology takes the current environment as the relevant context of adaptation, and builds models on the basis of expected utility of different behaviors. As organisms are expected to optimize their behavioral repertoire to their ecological and social contexts, this approach predicts little or no mismatch between expected and actualized fitness benefits. Forms of CE that see religion as costly signaling are compatible with this approach, because it expects individuals to make flexible, optimal choices, which could include a choice for a religious affiliation that allows for costly signaling. However, this places limitations on the explanatory scope of SP theories. They cannot explain, for example, why hunter-gatherers or medieval villeins would be religious, since there is little point in the costly signaling of one's membership in a religious community if there is no freedom of religious choice. After all, a "free market" of religious groups is a relatively recent and not globally widespread phenomenon, mainly restricted to northern America (Finke & Stark, 1989). Many European countries have state-funded churches with low levels of expected commitment, which makes competition by smaller high-commitment religious groups harder. Prior to the eighteenth century, religious choice was quasi nonexistent, as the treatment of religious minorities in Europe exemplifies – one need but think of the massacre of the Cathars in southern France, or the historical persecution of Protestants. There is still no freedom of religion for the present-day Iraqi housewife. So, although behavioral ecology allows for CE, it seems to be limited in explanatory scope.

Dual inheritance theories examine human behavior as a product of culturally transmitted norms that have effects on genetic fitness, and that can thus become part of a feedback loop. This approach is the most congenial to the possibility of group selection. In particular, groups must be distinct from each other and form cohesive wholes for group selection to occur. Group selection also requires that the fitness benefits of altruistic groups over selfish groups must outweigh the fitness benefits of selfish individuals over altruistic individuals within mixed groups (Sterelny, 1996). Human cultures, with their ethnic markers and distinct languages, do exhibit high between-group variation, and considerable within-culture homogeneity, allowing for group selection to occur. It is within this context that we can situate Norenzayan and Shariff's (2008) argument that belief in supernatural sanction is a group-level adaptive cultural response to life in large societies, where interactions between unrelated and unacquainted agents become increasingly important. However, as they themselves point out, the presence of large, cooperative and not very religious groups indicates that secular institutions like the police can be equally successful in instilling cooperation. From the perspective of dual inheritance theory, PA seems thus not very likely, since people in at least some societies (e.g., agnostic Scandinavian countries) can withstand the temptation to cheat when effective punitive mechanisms are present *without* belief in divine punishment.

While different SP theories are not all mutually incompatible, some of them may be so because of their divergent assumptions about the temporal scale on which selection acts and which traits are the targets of selection. Depending on the style of evolutionary approach one chooses, contrasting SP theories can be fleshed out. However, like most SP theorists, S&M remain inexplicit about whether psychological mechanisms, behavioral strategies or cultural traits play the most prominent role in their review of the evolution of religious behavior.

References

- De Smedt, J., & De Cruz, H. (in press). Human artistic behaviour: Adaptation, byproduct, or cultural group selection? In K. Plaisance & T. Reydon (Eds.), *Philosophy of behavioral biology: Boston Studies in Philosophy of Science*. Heidelberg: Springer.
- Finke, R., & Stark, R. (1989). How the upstart sects won America: 1776–1850. *Journal for the Scientific Study of Religion*, 28, 27–44.
- Norenzayan, A., & Shariff, A.F. (2008). The origin and evolution of religious prosociality. *Science*, 322, 58–62.
- Smith, E.A. (2000). Three styles in the evolutionary analysis of human behavior. In L. Cronk, N. Chagnon & W. Irons (Eds.), *Adaptation and human behavior. An anthropological perspective* (pp. 27–46). New York: De Gruyter.
- Sterelny, K. (1996). The return of the group. *Philosophy of Science*, 63, 562–584.
- Vanhaeren, M., & d'Errico, F. (2005). Grave goods from the Saint-Germain-la-Rivière burial: Evidence for social inequality in the Upper Palaeolithic. *Journal of Anthropological Archaeology*, 24, 117–134.
- Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi Bushmen. A case of strong reciprocity? *Human Nature*, 16, 115–145.

Why God is the best punisher

Dominic Johnson*

University of Edinburgh, UK

In this article I briefly: (1) clarify the supernatural punishment hypothesis (SPH) and its components; (2) propose why supernatural agents offer more effective punishment than alternatives; (3) re-examine the conditions under which a fear of supernatural punishment can evolve; (4) explain the role of overconfidence in the original SPH; (5) outline new empirical support for SPH; and (6) reconcile contradictory elements of the “punishment avoidance” and “cooperation enhancement” versions of SPH.

1. SPH and its components

Schloss and Murray (S&M) make a distinction between two adaptive accounts of supernatural punishment: “punishment avoidance” (PA) and “cooperation enhancement” (CE). Other authors have also recently distinguished a “supernatural monitoring hypothesis” from SPH (Atkinson & Bourrat, 2011). These distinctions are useful in teasing apart different cognitive and behavioral mechanisms, but it is important to recognize that: (1) they are overlapping, not alternative concepts; and (2) all of these concepts come under the framework of SPH (see Figure 1). PA can evolve through individual selection and causes CE. By contrast, CE on its own is vulnerable to free-riders, relies on group selection, and does not necessarily cause PA. “Supernatural monitoring” is not an adaptive hypothesis, because surveillance in itself does not matter unless there are consequences.

2. Why God is the best punisher

S&M end their article by asking why, if selfish behavior became especially costly in human history, evolution did not favor a simpler solution for suppressing selfishness other than God: “why invoke...the cognitively costly and seemingly excessive

*Email: dominic.johnson@ed.ac.uk

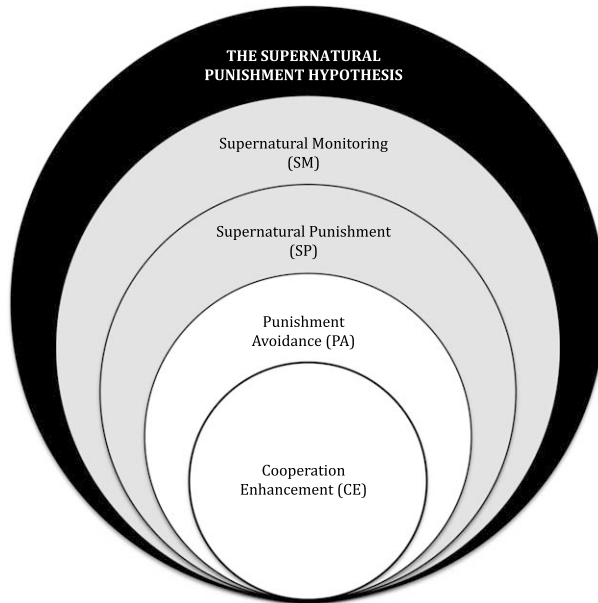


Figure 1. SPH (black circle) proposes that a fear of “supernatural” detection and punishment for selfish actions increases Darwinian fitness by reducing the “real-world” costs of norm transgressions. SPH requires two core “mechanisms” (in grey): supernatural monitoring (SM) and supernatural punishment (SP). These mechanisms have two core “consequences” (in white): punishment avoidance (PA) and cooperation enhancement (CE). They are nested, because each concept is a pre-requisite of the next: CE results from PV, which requires SP, which requires SM.

deterrence of belief in all-knowing, all-powerful agents, rather than just improve accuracy or globally reduce confidence in the ability to defect?” (p. 61). This is exactly the right question to ask, it presents the greatest challenge to SPH, and it remains unanswered by their analysis. I propose that there are several good reasons why God – or other supernatural agents – make the best deterrents against selfishness.

First, the cognitive science of religion suggests that supernatural beliefs are not “cognitively costly.” They may be behaviorally costly, but beliefs in supernatural agency and supernatural consequences are the cognitive default, and are most easily accessed (Atran, 2004; Boyer, 2001; Purzycki et al., 2011). Other solutions, if they are (or once were) rivals, might therefore have been trumped simply by ease of supernatural cognition (whatever the reasons for it).

Second, God works *because* God is “excessive.” Error management theory (EMT) shows that assessment mechanisms that aim for the true probability of an event (e.g., the true probability p of detection for cheating) will be suboptimal because, given some distribution of mistakes (centered around estimates of p), half the time we will overestimate p (and get away with it), and half the time we will underestimate p (and get caught). Therefore, simply “improving accuracy” (in estimating p), as S&M suggest, will not necessarily help. Indeed, if false negative errors (assuming secrecy and getting caught) are more costly than false positive errors (assuming detection and missing a reward), then only *exaggerated estimates* of the probability of detection (e.g., a false belief that supernatural agents are watching

you) will avoid the worst errors. EMT thus suggests that the best solution to avoiding the problem of detection is a mechanism that *overestimates* the true probability of detection. If humans are *already* overconfident about their likelihood of getting away with cheating, as S&M argue, then the counter-balancing bias would need to be greater still. Their concern that God is a “seemingly excessive” deterrent is prescient – the threat of punishment *may need to be excessive*, such as an omniscient and omnipotent God, as this is the only way (or at least one good way) to make overconfident humans avoid dangerous mistakes (Johnson, 2009b).

Third, one response to the EMT argument above is to accept the point – we need biased, not accurate, assessment – but to invoke a *non-religious* cognitive bias to correct the problem instead of God. However, as S&M already point out, the problem *arises* in part because of *overconfidence* in avoiding detection, so we cannot simultaneously postulate a bias in *underconfidence* in avoiding detection as well. God offers an alternative tool to solve the problem in a different way.

Fourth, cognitive biases might act as cautionary mind-guards, but they suffer two weaknesses: (1) a cognitive bias may suppress selfishness but has no *consequences*, whereas God suppresses selfishness *and punishes* if cheating is discovered; (2) a cognitive bias only affects individuals, with no external validation, whereas a belief in supernatural punishment is a shared aspect of culture that is reinforced by the whole community and bolstered by explanations of multiple events.

Fifth, the empirical evidence suggests that people do expect supernatural punishment for selfish behavior (even atheists in some cases), while there is a lack of evidence for alternative psychological mechanisms (only overconfidence, as noted, which works in the opposite direction).

Sixth, some empirical evidence suggests that religious beliefs are *more effective* at suppressing selfish behavior and promoting cooperation than equivalent non-religious beliefs (e.g., Sosis & Bressler, 2003). So even if alternative mechanisms exist now or in the past, religion may have had the competitive edge in selection.

Seventh, God offers *better detection*. Since humans are limited by the laws of nature, real-world detection is by no means certain to occur. By contrast, supernatural agents, though variable in power and motives, are peculiar precisely because of their ability to be in many places at one time and to have access to people’s thoughts and actions. No human can match the detection abilities of God.

Eighth, God offers *better punishment*. We might fear detection by other humans, but how bad can the consequences be? Typical punishments are some form of sanction, which may deter some, but not all. Even major punishments such as death are somewhat finite. The thing about God is that God’s punishments can be significantly worse than any earthly punishments that humans could inflict: they are certain, possibly worse than death, and infinite. No human can match the punishing abilities of God.

Ninth, supernatural punishment may trump other solutions, even good ones, because of a range of evolutionary constraints (Johnson, 2009a; McKay & Dennett, 2009): (1) *economics* – a fear of supernatural agency may have been biologically cheaper or more efficient to select or sustain than alternatives; (2) *history* – a capacity for supernatural beliefs may have been more readily available, especially given the recent evolution of theory of mind (itself necessary for beliefs in supernatural agents and agency), which gave rise to the very problem we are trying to solve (the increased costs of selfishness with social transparency); (3) *adaptive landscape* – fear of detection and punishment by supernatural agents may have been a small step up the

local fitness peak from fear of detection and punishment by human agents, even if better solutions from, say, corrective cognitive biases occupy higher fitness peaks beyond uncrossable valleys in the adaptive landscape.

To summarize this section, supernatural agents may be the only way (or at least one good way) of scaring people into avoiding costly social mistakes, especially given the human overconfidence in avoiding detection that S&M identify. “Simple prudence” (S&M, 2011), “cautious action policies” (McKay & Dennett, 2009), or even relying on our “conscience,” may not be good enough. As long as people believe it (an important caveat, of course), deterrence by supernatural agents is vastly superior to any alternative. In the ideal type (an all-powerful God), supernatural punishment solves several tricky game theoretical problems: (1) cheats are automatically detected (God is omniscient); (2) cheats are automatically punished (God is omnipotent); (3) there are no “second-order free riders” (God does the punishing); (4) there are no reprisals against punishers (no vigilantes are needed); and (5) there are fewer first-order free riders (reducing the necessity and thus the costs of real-world monitoring and punishment). The ideal type may be rare because in traditional societies supernatural agents typically do not have the full complement of powers as above, but the theoretical logic remains even if there is variance in the extent of supernatural agent capabilities and the extent to which people believe in them. Despite variation, supernatural agents are much better than humans at solving the game theoretical problems of cooperation.

3. Conditions for the evolution of God-fearing

Jesse Bering and I laid out a framework for when a God-fearer (GF) strategy could evolve in comparison to an alternative Machiavellian (M) strategy (Johnson & Bering, 2006). The broad condition for the evolution of GF is whenever the probability of detection of selfish actions (p) multiplied by the cost of punishment (c) exceeded the costs of missed opportunities (m). That is, when $pc > m$. S&M suggest that this should be modified to $f \times l(c) > m + r$ (where f is the frequency of defection attempts, l is the likelihood that any given attempt is detected, and r is the cost of religious belief and behavior). Although useful for clarification purposes, I suggest below that this does not alter the fundamental conclusions, for several reasons.

First, our model considered $pc > m$ (or $pc < m$) as an environmental condition, which would lead to selection for strategies that alter p or m (individuals cannot alter c). Differences between GF and M do not lead to the conclusion that $pc > m$ (as S&M imply); they result from it. Given variation in the frequency of cheating in the population, $pc > m$ will select for GF (or a strategy like it).

Second, what is the effect of r ? In the original model, a GF belief is not, in itself, costly (all it does is alter p and m , the costs of which are already included in the model). Nevertheless, let us say that we allow that r invokes additional consequences of religious beliefs and practices. This would be important because GFs have positive values for r , tipping the balance against the evolution of GF. However, allowing additional consequences associated with religious beliefs and practices means that we should include the benefits (b) that result from religious beliefs and practices, as well as the costs (r). The value of b may offset or, as many evolutionary theories suggest, exceed r due, for example, to higher payoffs for joint activities among believers (e.g., Sosis, 2003), further facilitating the evolution of GF. The original model effectively assumed that $r = b$, which may or may not be true in any given socio-ecological

setting, but making this assumption allows us to focus on the independent role of other parameters.

Third, how do f and l differ between GFs and Ms? The original model assumes that GFs make fewer attempts to cheat ($f_{GF} < f_M$), otherwise GF = M and there would be no difference in strategies to compare. By contrast, the original model stressed that both GFs and Ms shared the same cognitive capacities for reputation management (intentionality system and complex language; GFs and Ms both had these while the ancestral state did not). In other words, it assumed that GFs and Ms are equally likely to be detected when cheating ($l_{GF} = l_M$). Given these assumptions, $l_M \times f_M > l_{GF} \times f_{GF}$ (equivalent to $p_M c_M > p_{GF} c_{GF}$ in the original model), tipping the balance in favor of the evolution of GF.

However, while S&M agree that GFs cheat less often ($f_{GF} < f_M$), they suggest that Ms are more likely to get away with cheating when they do so ($l_{GF} > l_M$). This is debatable, since although one could argue that Ms are more skilled deceivers, GFs perceive detection under more circumstances, which may make them especially cautious whenever they cheat. Nevertheless, let us say that we allow that $l_{GF} > l_M$; this does not affect the outcome of the model. Why? Because if $f_{GF} < f_M$ and $l_{GF} > l_M$, then as long as they are similar in magnitude these differences will cancel out ($p = f \times l$). For example, if GFs are half as likely to cheat as Ms, but they are twice as likely to be detected when they do so, then $p = 0.5 \times 2 = 1$ for GFs and $p = 1 \times 1 = 1$ for Ms, so the overall probability of detection would be the same for each strategy ($l_M \times f_M = l_{GF} \times f_{GF}$).

The upshot of all this is that if S&M are right that Ms are more likely to get away with cheating than GFs ($l_{GF} > l_M$) then the probability of detection ($p = f \times l$) and thus the associated cost of cheating is the same between GFs and Ms. Without fitness differentials, neither strategy could be selected for by evolution. However, if $f_{GF} < f_M$ and $l_{GF} = l_M$, as in the original model, then GFs will evolve wherever $pc > m$. One possible source of confusion is that S&M situate f as a component of p , whereas f is already included in the original model as m (less cheating, f , is equivalent to a higher number of missed opportunities, m). Similarly, l is already included in the model as p (the likelihood that any given attempt to cheat is detected, l , is the same as the overall probability of detection across all attempts, p).

4. The role of overconfidence in SPH

S&M argue that their “more nuanced version” of SPH is “significantly different” from the original. Rather than simply avoiding the costs of defection, they see supernatural punishment as a tool to manage “a native tendency to defect on strategic partners owing to an over-confidence in our ability to avoid detection” (p. 57). However, original descriptions of SPH already included the problem that people underestimate their probability of detection (or, equivalently, are over-confident about their ability to avoid detection), along with evidence in support. Indeed, *without* an assumption that people underestimate the probability of detection, there is no puzzle, and no adaptive advantage of a fear of supernatural punishment.

First, we focused on the fact that selfish behavior is evolutionarily ancient (Johnson & Bering, 2006). Indeed, it is often motivated by deep-seated parts of the limbic system that we share with many other animals and is hundreds of millions of years old. Many behaviors from aggression to impatience can be traced to limbic

brain areas (Davidson, Putnam & Larson, 2000; McClure, Laibson, Loewenstein & Cohen, 2004). This problem is implicitly but widely recognized in, for example, legal allowances for crimes of passion that arise from immediate, emotional motives rather than premeditated motives (Goldstein, 2002). Human biology leads to behavior that undervalues the potential costs of selfishness.

Second, we cited empirical studies supporting the idea that people underestimate the probability of detection and punishment. For example, potential offenders have been shown to systematically downplay the likelihood of capture, as well as the costs of punishment (Robinson & Darley, 2004).

The ancient bio-psychological mechanisms underlying selfish behavior mean it was not easy for evolution to simply “retune” these motives to avoid displaying selfish behavior in certain social settings. Rather, they must coexist with, and sometimes contradict, motives and rationales from higher brain areas. This is why selfish behavior is potentially so costly and why we need corrective mechanisms.

5. New studies supporting SPH

S&M review the empirical support for SPH and point out some key weaknesses: “it is not clear that supernatural attributes are necessary for the deterrent effect of moralistic, monitoring agents, nor is there evidence that such attributes increase deterrence when monitoring agents are believed to be present” (p. 52). I would add another empirical weakness: we do not know if negative supernatural punishment is better at suppressing selfishness or promoting cooperation than positive supernatural rewards. However, new studies are shedding light on these deficiencies: (1) supernatural primes have been shown to significantly increase sanctions on unfair behavior “over and above secular punishment primes” (McKay, Efferson, Whithouse & Fehr, in press); (2) cheating is less likely among those who see God as “punishing rather than loving” (Shariff & Norenzayan, in press); (3) people are much quicker to attribute God’s knowledge of “ill deeds rather than good deeds” (Purzycki et al., 2011); and (4) a very large cross-cultural sample from the World Values Survey found that a range of moral transgressions were rated as significantly less justifiable “among those who believe in the afterlife” (Atkinson & Bourrat, 2011). It is also worth noting two supportive general principles recently identified in psychology: a bias towards “attributing agency in negative events more than positive events” (Morewedge, 2009), and a bias for people to be more attentive to, and more affected by, “negative rather than positive information and events” (Baumeister, Bratslavsky, Finkenauer & Vohs, 2001).

The latter studies imply that even where religious doctrine or theology emphasize reward as well as punishment, human beings are likely to be particularly prone to concerns about punishment. Punishment also has an intrinsic leverage in game theory: rewards can encourage many to cooperate, but they cannot deter people from cheating. The former studies offer initial steps in demonstrating that “supernatural attributes” are more effective than secular equivalents, and that such attributes increase deterrence.

6. Reconciling PA and CE

S&M point out a major problem in reconciling PA and CE accounts: the PA account invokes a moralizing God as a solution to “abundant real-world punishment”

(supernatural punishment spreads because the costs of real-world punishment are high); the CE account invokes a moralizing God as a solution to a “*lack of real-world punishment*” (supernatural punishment spreads because the costs of real-world punishment are low). They appear, therefore, to offer two contradictory proposals.

However, I would like to suggest that they are actually complementary mechanisms that “come to the fore in different ecological settings” (see Table 1). PA is primarily a theory about how a fear of supernatural punishment emerged in human evolution, along with theory of mind and complex language – sometime in the Pleistocene (1.8 million to 10,000 years ago). CE is primarily a theory about how a fear of supernatural punishment maintains cooperation in more recent, large, anonymous societies where the reputations of strangers cannot be tracked – sometime in the Holocene (10,000 years ago to the present). CE also fits with Alexander’s hypothesis that moralizing gods were necessary to maintain social cohesion and collective action under the threat of severe intergroup conflict as human societies became large (Alexander, 1987; Roes & Raymond, 2003). With this perspective, PA and CE are not mutually exclusive at any time, but become more or less important in different socio-ecological settings.

Note some interesting features of these different scenarios: (1) CE suffers from the free-rider problem and requires group selection – but this is something that would be much more likely in situations of severe inter-group conflict and inter-group fitness differences (Bowles, 2006); (2) PA requires gods concerned with egalitarian “morals,” since the objective is to suppress selfish behavior that negatively affects other people. However, the objective of CE is collective action, which can be achieved through dictatorial leadership, strong social hierarchies, and capricious gods (like the Romans). It does not necessarily require moral gods, just ones that effectively achieve cooperation.

Conclusion

S&M have done a great service to evolutionary theories of religion by carefully teasing apart a diverse range of disorganized theoretical proposals and empirical

Table 1. Key components of PA and CE versions of SPH.

Era	Evolutionary problem	Real-world punishment	Level of selection	Adaptive function of supernatural punishment	Requisites
Pleistocene	Maintaining reputation within small groups	Effective	Individual	Punishment avoidance (PA)	Moral gods
Holocene	Maintaining cooperation within large groups Inter-group competition	Not effective	Group	Cooperation enhancement (CE)	Leadership

findings. In this article I have argued that: (1) the various components they describe fit well within the general framework of SPH; (2) supernatural agents offer more effective punishment than alternatives; (3) the conditions under which a fear of supernatural punishment can evolve remain favorable, if in need of empirical validation; (4) the role of overconfidence is already implicit in the original SPH; (5) new empirical work offers important new support for SPH; and (6) contradictory elements of the PA and CE versions of SPH can be reconciled by distinguishing their effects in the Pleistocene and Holocene epochs. To end, while supernatural agents vary in the extent of their omniscience and omnipotence, and the extent to which people believe in them, they carry a pack of theoretical ace cards: powers of deterrence, detection, and punishment that no human individual or organization can match. It looks as though evolution worked this out long before we did.

References

- Alexander, R.D. (1987). *The biology of moral systems*. Aldine, NY: Hawthorne.
- Atkinson, Q.D., & Bourrat, P. (2011). Beliefs about God, the afterlife and morality support the role of supernatural policing in human cooperation. *Evolution and Human Behavior*, 32, 41–49.
- Atran, S. (2004). In *Gods we trust: The evolutionary landscape of religion*. Oxford: Oxford University Press.
- Baumeister, R.F., Bratslavsky, E., Finkenauer, C., & Vohs, K.D. (2001). Bad is stronger than good. *Review of General Psychology*, 5, 323–370.
- Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314, 1569–1572.
- Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.
- Davidson, R.J., Putnam, K.M., & Larson, C.L. (2000). Dysfunction in the neural circuitry of emotion regulation: A possible prelude to violence. *Science*, 289, 591–594.
- Goldstein, M.A. (2002). The biological roots of heat-of-passion crimes and honor killings. *Politics and the Life Sciences*, 21, 28–37.
- Johnson, D.D.P. (2009a). God would be a costly accident: Supernatural beliefs as adaptive. *Behavioral and Brain Sciences*, 32, 523–524.
- Johnson, D.D.P. (2009b). The error of God: Error management theory, religion, and the evolution of cooperation. In S.A. Levin (Ed.), *Games, groups, and the global good* (pp. 169–180). Berlin: Springer.
- Johnson, D.D.P., & Bering, J.M. (2006). Hand of God, mind of man: Punishment and cognition in the evolution of cooperation. *Evolutionary Psychology*, 4, 219–233.
- McClure, S.M., Laibson, D.I., Loewenstein, G., & Cohen, J.D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306, 503–507.
- McKay, R.T., & Dennett, D.C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–561.
- McKay, R.T., Efferson, C., Whitehouse, H., & Fehr, E. (in press). Wrath of God: Religious primes and punishment. *Proceedings of the Royal Society B: Biological Sciences*. doi: 10.1098/rspb.2010.2125
- Morewedge, C.K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology*, 138, 535–545.
- Purzycki, B.G., Wales, N., Finkel, D.N., Shaver, J., Cohen, A.B., & Sosis, R. (2011). *What does god know? Supernatural agents' perceived access to socially strategic and nonstrategic information*. Manuscript submitted for publication.
- Robinson, P.H., & Darley, J.M. (2004). Does criminal law deter? A behavioural science investigation. *Oxford Journal of Legal Studies*, 24, 173–205.
- Roes, F.L., & Raymond, M. (2003). Belief in moralizing gods. *Evolution and Human Behavior*, 24, 126–135.
- Shariff, A.F., & Norenzayan, A. (in press). Mean Gods make good people: Different views of God predict cheating behavior. *International Journal for the Psychology of Religion*.
- Sosis, R. (2003). Why aren't we all Hutterites? Costly signaling theory and religious behavior. *Human Nature*, 14, 91–127.
- Sosis, R., & Bressler, E.R. (2003). Cooperation and commune longevity: A test of the costly signaling theory of religion. *Cross-Cultural Research*, 37, 211–239.

Disbelief in the gods and the “cooperation enhancement” account of supernatural punishment theory

Ryan Nichols*

Center for Philosophy of Religion, University of Notre Dame and Department of Philosophy, Cal State Fullerton, USA

This insightful, comprehensive paper teases out several implications that follow from the adoption of one of two different accounts of the evolutionary operation of supernatural punishment.

Though the punishment avoidance account is preferable, I want to reflect on its alternative. Schloss and Murray (S&M) discuss a problem for the cooperation enhancement account of belief in supernatural punishment (SPT-CE): that agents fail to, and will believe and act as though they have failed to, receive supernatural punishment from deities after performing anti-cooperative behavior, which causes disbelief in the gods. S&M develop two responses. On the first, “while religion may facilitate cooperation, its origin and successful transmission may not require this benefit and the credible punishment that underlies it. It may well be that religion arises and endures natively, as a cognitive spandrel, and subsequently comes to confer cooperative benefits. Thus, religious belief is properly understood as an exaptation rather than an adaptation” (p. 49). On the second, “belief in the efficacy of supernatural sanctions can be stabilized by additional features of religious systems that deflect the above epistemological problem” (p. 49), like teachings about hell.

First, SPT-CE is an explanation of how belief in supernatural punishment confers an adaptive advantage owing to the cognitive effects of this belief on the believer’s behavior, but in response to the “problem of disbelief” the authors suggest that belief in supernatural punishment is an exaptation. This sows uncertainty in those readers who thought that the authors were arguing that SPT-CE is an account of the fact that belief in supernatural punishment is adaptive. More important, to respond to the problem of disbelief by saying that “religion arises and endures natively, as a cognitive spandrel” suggests that the authors are interested in explaining the origins and transmission of religion rather than whether belief in supernatural punishment is adaptive. This impression is reinforced by a somewhat woolly discussion of cooperation and by an indecisive treatment of group selection. Supernatural punishment may be responsible for the endurance and transmission of religion through culture; it may explain the memetic success of some religions, but those claims are orthogonal to what I thought was the central issue surrounding discussion of SPT-CE: the adaptive benefits of belief in supernatural punishment.

Second, according to the second response to the problem of disbelief, supernatural punishment theory has available additional resources that show belief in supernatural punishment is adaptive after all. (Do the first and second responses to this problem thus sit awkwardly together? Perhaps the gap between them was intended to be finessed via the uses of modal terms like “may/may not” and “can” here.) “Additional resources” refers to components that cause fear in in-group members, like imagery of hell and God’s punishing authority over it. But how

*Email: rnichols@exchange.fullerton.edu

do these facts bear on the evidential question raised by the problem of disbelief for SPT-CE? I thought this is the problem: a disbelieving free-rider, Fred, thinks there is no supernatural punishment because he believes God does not exist. He is party to a religious in-group and has incentive to extract resources from the group without paying costs of cooperation. In-group members wary of disbelievers deploy fear-inducing features of their memplex in order to cajole Fred to believe in nasty punishments of powerful deities so that he will change his behavior and not take advantage of in-group altruists. This appears beside the point. *Ex hypothesi* Fred disbelieves in God, on which hell depends. So why would supplemental forms of punishment in the religious memplex have the effect on Fred that S&M suggest it would? The authors have not justified the inference from the presence of memes like hell in religious systems to the fact that those memes successfully inhibit cheating behavior in those persons who already disbelieve. (I am not suggesting that the presence of hell in religious systems has no effect on religious populations; their presence might make religious believers more cooperative.)

Third, the description of cooperation in the opening of the paper lacks a definition of “cooperation.” This contributes to uncertainty about the role of group selection in the presentation of SPT-CE. The SPT-CE account needs to explain how belief in supernatural punishment encodes a generalized inhibition to defection when that inhibition appears not to be advantageous to individuals’ fitness. S&M say that the “most straightforward and plausible” solution to this problem is that SPT-CE is a “group-level adaptation to coordinate cooperation and inhibit the destabilizing effects of defection” (p. 53). Earlier they write that SPT-CE implies belief in supernatural punishment is “selected for – probably at the group level . . .” (p. 48). The punishment avoidance account is described as preferable to SPT-CE on the grounds that it “is not a group adaptation that requires individuals to relinquish fitness enhancing opportunities” (p. 54). S&M understandably intend to “leave aside general debates over group selection,” but for a paper devoted to theoretical discussion of two evolutionary accounts of belief in supernatural punishment, to leave the status of one of the two ambiguous in this way is surprising.

Fourth, the authors open the paper describing belief in supernatural punishment as a helpful means humans used to create and sustain cooperation in the face of challenges of first- and second-order defection and free-riding. The primary explanandum in the target paper is the adaptive benefits of belief in supernatural punishment (a form of in-group cooperation), for which S&M present two competing explanations (CE and PA), but I lack a sense of how much adaptive human cooperation, and what kinds of adaptive human cooperation, are in need of explanation via appeal to the effects of belief in supernatural punishment. After all, S&M cite research indicating secular word scrambles are correlated with as much cooperative economic behavior as are religious word scrambles. The present point is important since humans are not as uniquely cooperative as many – especially proponents of “strong reciprocity,” like Fehr and Fischbacher, frequently cited by S&M – have claimed (see target article for references). This is emerging with further research on economic games. Flipping standard economic games on their head, Kümmerli, Burton-Chellew, Ross-Gillespie, & West (2010) show that in games where 100% cooperation would maximize participants’ personal financial gain, they still did not fully cooperate. This and other new research calls into question the logic of the utility functions of previous economic games and corollary conclusions about

uniquely human prosociality that were cited as motivating the target article in the first place.

Perhaps the form and amount of cooperation that belief in supernatural punishment could explain is not especially interesting. If religion was ancestrally practiced within human bands and migration was sufficiently infrequent, then relatedness (r) is going to be high. Limited dispersal and increased genetic similarity resulting from population structure leads to increases in average relatedness in the group so that one's first cousin might have an r of much more than an eighth. This might be helpful in improving an understanding of the origins of religion-based cooperation (provided it is accompanied with data from the anthropology of religious groups) by replacing talk of group selection with a more prudent discussion of mechanisms of kin selection that arise in groups. Could it be that the behavioral tracks of religious practice and supernatural punishment might boost in-group cooperation in part because they have been laid upon well-worn rules for inclusive fitness behavior?

My small-minded comments aside, the conceptual trailblazing Schloss and Murray accomplish in the paper should be looked on as a huge favor to those of us in the Cognitive Science of Religion and allied fields who have struggled to understand where hypotheses from supernatural punishment theory might lead. This paper ought to be read by people working in the area for years to come.

Reference

- Kümmerli, R., Burton-Chellew, M.N., Ross-Gillespie, A., & West, S.A. (2010). Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences USA*, 107, 10125–10130.

Imagine there is no religion

Ilkka Pyysiäinen*

The Study of Religions, Helsinki University, Finland

Schloss and Murray (S&M) have written an important and well-argued target article. They provide a perceptive analysis that helps us understand the many issues and problems involved in the differing evolutionary approaches to religion. Here I argue that the by-product view of religion (Boyer, 1994; McKay & Dennett, 2009; Pyysiäinen & Hauser, 2010) that S&M by-pass all too quickly offers the best way of explaining some of their insights.

S&M write that religion may have arisen as a cognitive spandrel, not as an adaptation, and that it may then have conferred cooperative benefits (as also argued in Pyysiäinen & Hauser, 2010). Religious behaviors might be signals of a cooperative disposition but they do not necessarily do any causal work in sustaining community coherence. However, “(t)he *prima facie* problem with viewing costly religious behaviors as spandrels rather than as adaptations, is precisely that they do appear

*Email: ilkka.pyysiainen@helsinki.fi

so costly” (n. 8), but appearing and being costly are two different things, and S&M present no compelling evidence for their claim about the costliness of all religion.

Although it certainly is possible to find examples of materially and cognitively costly forms of religion, not all religion is costly; simple forms of folk belief rather “come naturally” to humans (McCauley, 2000, in press). Sperber and Wilson’s (1986) relevance theory thus yields the following predictions:

- (1) When the processing costs of two interpretations are equivalent, an inferentially richer interpretation is favored;
- (2) When the inferential potential is the same, the less costly interpretation is favored (Boyer, 2003, p. 352).

Theological systems, for their part, are cognitively costly elaborations of folk beliefs. They also necessitate material and social infra-structures that are costly (see Pyysiäinen, 2009). If this is the case, the by-product argument cannot be that easily dismissed.

First, “religion” cannot be explained by any one theory alone. The category of “religion” is simply too vague (Boyer, 1994; Saler, 2010). Second, if various kinds of natural beliefs and the related behaviors we group under the umbrella term of “religion” are by-products of evolved cognitive mechanisms and emotions, they cannot be the ultimate foundation of cooperation and morality (Boyer, 1994; Pyysiäinen & Hauser, 2010; see Sinnott-Armstrong, 2009). As S&M point out, we may well ask “why belief in the supernatural would be necessary to motivate cooperative behaviors that are natural to human beings and that are adaptive in the social environment postulated by this account” (p. 56).

Boyer (2006) observed earlier that the human mind, indeed, is endowed with numerous non-religious, prosocial, cognitive mechanisms, all of which evolved independently of supernatural or religious beliefs, and operate in similar ways in people with or without such beliefs (see also Pyysiäinen & Hauser, 2010). “Religious” concepts and beliefs are used to legitimize and explain prosocial and moral intuitions because treating these as norms laid out by invisible agents has been the most cognitively cheap strategy. Gods, ancestors, and so forth are “interested parties” in human social life (Boyer, 2001, 2002). As their cognitive representation relies on the same cognitive mechanisms as any agent representation, they do not form an essentially distinct category of agency (Boyer, 2001; Pyysiäinen, 2009; see Hari & Kujala, 2009).

Also, S&M argue that “it is not clear that supernatural attributes are necessary for the deterrent effect of moralistic, monitoring agents” (p. 52). Referring to Bering’s experimental research, they ask “to what extent the deterrent effect of the primes in these experiments depends on the supernatural character of the concepts involved” (p. 52). There is not necessarily “anything about *religion* in particular that accounts” for adaptively salient social benefits (p. 59). Yet religious beliefs “may increase the likelihood of construing the presence of a monitoring agent” (p. 52). Thus, feeling or experiencing a vague presence of an invisible agent seems to trigger moral intuitions and to increase prosocial behavior, even if the agent is not religiously understood.

In the standard model of the cognitive science of religion, supernaturalness is understood as counterintuitive to one’s expected and innate ontology (Boyer, 2001). Thus, for example, a personal agent without a physical body is counterintuitive (see Pyysiäinen, 2009). “Religious” counterintuitiveness is but one general

example. In this view, an agent that is present but cannot be perceived is counterintuitive quite irrespective of whether it has any religiously understood supernatural attributes. Feeling the presence of such counterintuitive agency seems to increase individual prosociality and morality (e.g., Bateson, Nettle, & Roberts, 2006; see Boyer, 2002). “Religious” interpretations of such agency are merely a (perhaps, causally impotent) add-on to various natural human inclinations.

Thus, the most parsimonious strategy is to explore the evolution of these mechanisms and their significance for the evolution of cooperation and morality. We cannot legitimately reason from the present functions of religious traditions, as we now know them, to the evolutionary causes of prosociality and morality. Religious traditions presuppose rather than explain the human tendencies and abilities to read other minds, sympathize, and act cooperatively. But once in place they may contribute to human cooperation (see Pyysiäinen, 2010).

References

- Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in real-world setting. *Biology Letters*, *22*, 412–14.
- Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Berkeley, CA: University of California Press.
- Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.
- Boyer, P. (2002). Why do gods and spirits matter at all? In I. Pyysiäinen & V. Anttonen (Eds.), *Current approaches in the cognitive science of religion* (pp. 68–92). London: Continuum.
- Boyer, P. (2003). Science, erudition and relevant connections. *Journal of Cognition and Culture*, *3*(4), 344–358.
- Boyer, P. (2006). Prosocial aspects of afterlife beliefs: Maybe another byproduct. A commentary on Bering. *Behavioral and Brain Sciences*, *29*(5), 466.
- Hari, R., & Kujala, M. (2009). Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, *89*, 453–479.
- McCaughey, R.N. (2000). The naturalness of religion and the unnaturalness of science. In F.C. Keil & R.A. Wilson (Eds.), *Explanation and cognition* (pp. 61–85). Cambridge, MA: MIT Press.
- McCaughey, R.N. (in press). *The naturalness of religion and the unnaturalness of science*. New York: Oxford University Press.
- McKay, R.T., & Dennett, D.C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*(6), 493–561.
- Pyysiäinen, I. (2009). *Supernatural agents: Why we believe in souls, gods, and buddhas*. New York: Oxford University Press.
- Pyysiäinen, I. (Ed.) (2010). *Religion, economy, and cooperation*. Berlin: Mouton de Gruyter.
- Pyysiäinen, I., & Hauser, M. (2010). The origins of religion: Evolved adaptation or by-product? *Trends in Cognitive Sciences*, *14*(3), 104–109.
- Salter, B. (2010). Theory and criticism: The cognitive science of religion. *Method & Theory in the Study of Religion*, *22*(4), 330–339.
- Sinnott-Armstrong, W. (2009). *Morality without God*. New York: Oxford University Press.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Big gods were made for big groups

Azim F. Shariff*

Department of Psychology, University of Oregon, USA

Schloss and Murray’s (S&M’s) thoughtful map of the supernatural watcher hypothesis’ current landscape sizes up its two theoretical peaks and empirical

*Email: shariff@uoregon.edu

valleys. The cooperation enhancement (CE) and punishment avoidance (PA) accounts both recognize the prosocial effects of religious systems but diverge in their analysis of who directly benefits from this prosociality – individuals or groups. As important as the distinction between whether supernatural punishment is a group- or individual-level adaptation is the distinction as to whether these adaptations are genetic or cultural in origin. Each combination of factors (see Figure 1) requires different theoretical commitments and makes different empirical predictions that can be tested against existing evidence. I want to take this opportunity to briefly review both debates and explain why I believe the CE account rests on firmer empirical ground.

Cultural vs genetic origins

S&M are for the most part mute on the issue as to whether the belief in supernatural punishing agents represents a genetic adaptation or a cultural one. However, this debate is worth briefly engaging, as it can constrain the viable set of solutions to the other debates that are raised.

Before fully wading into the CE territory, S&M make a short note about the long-standing debates over group selection, warning their fellow explorers that, however attractive the arguments that lie ahead may be, “here be dragons.” The authors are right to be wary of explanations of human cooperation that heavily rely on genetic group selection. The strict preconditions (small group size, intense selection pressures and very limited between-group migration) required for genetic group selection are not met by either early or modern human societies. However, though often conflated with its genetic cousin, the cultural group selection argument requires fewer and less strenuous situational preconditions and provides a very plausible mechanism for the selection and transmission of cooperative norms (Henrich, 2004). As a result, cultural – and not genetic – group selection provides a more viable explanation for the origins of supernatural punishment beliefs.

		MODE OF SELECTION	
		Genetic	Cultural
BENEFICIARY	Individual	<i>Does not explain cultural variability in god-beliefs</i>	<i>Does not fit the pattern of ethnographic evidence</i>
	Group	<i>Neither early nor modern human societies meet the necessary preconditions for viable GGS of cooperation</i>	<i>Empirically sound and plausible</i>

Figure 1. Evaluating the different scenarios in the evolution of supernatural watchers. Theoretical and empirical constraints suggest that cultural group selection is the most likely explanation.

Individual-level genetic explanations of supernatural punishment beliefs also run into a number of challenges (see Shariff, Norenzayan, & Henrich, 2009). Most prominent is the marked variability in god-beliefs across both cultures and time. If beliefs in omniscient and punitive gods were genetic adaptations rooted in our Pleistocene past, we would expect these beliefs to be psychological universals, or, at the very least, more prevalent in hunter-gatherer societies. Neither is true. Zuckerman (2007) estimates that there are 700 million atheists in the world, and there are many millions more who do not believe in the type of “big gods” that meet the requirements for establishing credible deterrence. Indeed, anthropological studies of foraging societies tend to reveal beliefs in local gods with limited monitoring and punitive powers (Boehm, 2008).

Big gods with omniscient scope and ultimate punitive abilities tend to be relatively recent Holocene innovations and ones that developed only in large, complex societies (Henrich et al., 2010; Roes & Raymond, 2005; Snarey, 1996). This pattern of evidence is most consistent with these beliefs originating as byproducts of existing cognitive adaptations, emerging relatively late in human history (5,000–10,000 years), and spreading culturally, rather than genetically.

Individual- vs group-level selection

Thus, focusing solely on cultural evolution, the PA and CE accounts can be distinguished as to whether supernatural punishment beliefs benefit the individual and are transmitted via conformity biases, prestige biases, etc., or whether the beliefs confer benefits on the group and are transmitted via cultural group selection.

The ethnographic evidence just discussed is of note here, as well. The PA account predicts that as the benefits of defection rise in comparison to the costs, the defection-suppressant effects of big gods would become less adaptive and the gods themselves would wane. The CE account instead predicts that as defection becomes more tempting for individuals, and thus more liable to destabilize groups, the need for big gods increases.

The evidence supports the latter account. The big gods most effective at enforcing norm-following behavior have been shown to emerge as societies grow larger and more complex, more market integrated or more reliant on cooperation to address scarce resource allocation (Henrich et al. 2010; Roes & Raymond, 2005; Snarey, 1996) – that is, as societies grow more vulnerable to being undermined by individual defection, and individuals see more opportunity and benefit for gaming the system. These examples feature societies where defection becomes increasingly valuable for the individual, increasingly costly to the society, or both. They thereby comprise (to my knowledge) the best existing evidence pitting the individual and group-level explanations against each other.

Thus, my reading of the empirical work on this issue currently favors the cultural, group-level selection account (again, see Shariff, Norenzayan, & Henrich, 2009, for more detail).

Addressing issues with the cultural group selection account

S&M spread their criticism judiciously between the various positions and raise a number of outstanding issues for this account. For one, S&M question whether belief in supernatural punishing agents can really serve as an actual human punishment

replacement when such a belief deviates so starkly from reality. I refer readers to Johnson's (this issue) remarkably thorough response explaining why such agents are not only highly effective at establishing cooperation via the threat of punishment, but also cognitively 'cheap' (making them, I would add, ideal candidates for culturally selected evolutionary byproducts).

S&M also raise the important issue of disbelief, which presents a significant threat to any cooperative system that relies on sustained widespread belief for its function. However, the dire need for mechanisms within religions to minimize disbelief can provide a mutually revelatory explanation for the power and peculiarity of anti-atheist prejudice (AAP). Though research on the topic is thin, recent studies show AAP to be more robust than "standard-fare" outgroup hostility, despite the fact that atheists form neither a coherent nor an especially visible group (Gervais, Shariff, & Norenzayan, in press). Moreover, consistent with predictions made by the discussed theories, experiments show that the negativity directed towards non-believers is driven primarily by distrust, rather than dislike, and this distrust is powerfully predicted by endorsement of the belief that a supernatural monitor encourages good behavior.

Numerous aspects of religions can be seen as mechanisms aimed at deflecting or disincentivizing doubt (see Dennett, 2006), but AAP represents a particularly necessary and overt one. Admittedly, to sustain cooperation, these mechanisms would have had to be very effective at minimizing disbelief within those societies that relied on supernatural sanctioning to sustain cooperation.¹ Keeping levels of defection-by-disbelief low enough to prevent overwhelming free-riding has in all likelihood been one of the primary selective challenges in the evolution of religious systems. The religions of today bear the marks of that legacy; more than a few researchers have noted the remarkable effectiveness of religions (more so than any other cultural institution) at preventing defection under even the most extreme circumstances (e.g., Berman, 2009). Indeed, one could argue that is what they were built to do.

Note

1. When we look at the world today, it at least seems that those societies that rely least on religious beliefs to do the heavy lifting of cooperation are also those that are most tolerant of religious disbelief – though this awaits proper empirical testing that takes into account the numerous possible confounds.

References

- Berman, E. (2009). *Radical, religious, and violent: The new economics of terrorism*. Cambridge, MA: MIT Press.
- Boehm, C. (2008). A biocultural evolutionary exploration of supernatural sanctioning. In J. Bulbulia, R. Sosis, C. Genet, R. Genet, E. Harris, & K. Wyman (Eds.), *The evolution of religion: studies, theories, and critiques* (pp. 143–150). Santa Margarita, CA: Collins Foundation Press.
- Dennett, D.C. (2006). *Breaking the spell*. London: Penguin.
- Gervais, W.M., Shariff, A.F., & Norenzayan, A. (in press). Do you believe in atheists? Why distrust is central to anti-atheist prejudice. Manuscript submitted for publication.
- Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53, 3–35.
- Henrich, J., Ensinger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327, 1480–1484.
- Roes, F.L., & Raymond, M. (2003). Belief in moralizing gods. *Evolution and Human Behavior*, 24, 126–135.

- Shariff, A.F., Norenzayan, A., & Henrich, J. (2009). The birth of high gods: How the cultural evolution of supernatural policing agents influenced the emergence of complex, cooperative human societies, paving the way for civilization. In M. Schaller, A. Norenzayan, S. Heine, T. Yamagishi, & T. Kameda (Eds.), *Evolution, culture and the human mind* (pp. 117–136). Mahwah, NJ: Lawrence Erlbaum Associates.
- Snarey, J. (1996). The natural environment's impact upon religious ethics: A cross-cultural study. *Journal for the Scientific Study of Religion*, 80, 85–96.
- Zuckerman, P. (2007). Atheism: Contemporary numbers and patterns. In M. Martin (Ed.), *The Cambridge Companion to Atheism* (pp. 47–65). Cambridge: Cambridge University Press.

RESPONSE

How might evolution lead to hell?

Jeffrey P. Schloss*^a and Michael J. Murray^b

^aWestmont College, USA; ^bFranklin & Marshall College, USA

We are grateful to the respondents for thoughtfully illuminating points in our analysis they view as correct and important, for graciously pointing out the ambiguous or unpersuasive, and for suggesting new ways in which research on belief in supernatural punishment (SP) might be carried forward. The various responses represent a range of agreement and disagreement about both our own and the other respondents' analyses. Significantly, they also situate themselves at different branch points of evolutionary inquiry as described in our paper, thus focusing on different questions. First, is a trait an adaptation and if so, what is the challenge for which it provides a functional solution? Second, what is the selective regime that gave rise to the trait, including the level of selection? Third, what is the nature of the replicator informing the trait: genetic, cultural-memetic, and/or (a distinction made by some, e.g., Plotkin (1997, 2000)) learned strategies? We will organize our reply along this sequence.

Both Pyysiäinen and Cohen endorse a by-product account of religious belief and question the adequacy of adaptationist explanations of religion, though with differing arguments. Pyysiäinen acknowledges that we propose religion may have arisen as a spandrel, but he suggests that we bypass too quickly this option which “offers the best way of explaining” our observations. We agree that we both endorse it and then pass it by. It is not undeserving of consideration, but (a) what it purports to explain is not the specific feature of religion SP addresses (however adequately or inadequately), and (b) it is not necessarily an alternative to SP accounts, since arising as a byproduct admits to the possibility of exaptation. In response to the suggestion that highly costly religious behaviors may constitute a *prima facie* problem for exclusively spandrel accounts, Pyysiäinen claims that we do not present compelling evidence for our “claim that all religions are costly.” However we do not intend to make this claim, nor does SP assert or require this. What is salient to the adaptationist point is only what Pyysiäinen himself consents to: there are “examples of materially and cognitively costly forms of religion.” Finally, we strongly agree with

*Corresponding author. Email: schloss@westmont.edu

Pyysiäinen's important conclusions that "religion cannot be explained by one thing alone" and that once (and however) it arose, religion "may contribute to human cooperation." We explicitly propose the former point in our conclusion, and the latter point is precisely the exaptation position we favor. However both seem to be at variance with Pyysiäinen's initial suggestion that by-product accounts constitute the best explanation.

In her erudite comments, Cohen urges caution in adopting SP accounts, and rightly points out that early cognitive acquisition and wide cultural prevalence do not resolve the adaptationist questions of whether religious belief stabilizes cooperation or establishes punishment systems. We agree with the caution and the qualification. However, her ensuing three worries are more mixed. First, Cohen points out that punishment is by no means restricted to humans, and a comparative perspective that includes godless species should inform any attempt to explain the evolution of punishment in humans. This is true, but accounting for punishment is not itself the goal of SP. At least in the case of punishment avoidance (PA), it *assumes* effective punishment has already arisen, presumably by means that – even if they employ unique cognitive capacities – nevertheless reflect the fitness tradeoffs that stabilize punishment in other species. SP is understood as a response to, not a source of a well established system of punishment. Second, she points out that costly third party punishment may actually be quite rare. We agree, but for PA, even cheap third party punishment (e.g., shunning) can be costly for the punished; and for cooperation enhancement (CE), it is precisely the absence of effectively stringent punishing mechanisms that occasions the rise of belief. Thus, for neither option does a high frequency of costly punishment constitute a necessary precondition. Third, Cohen points out that belief in a consistently fair God is vulnerable to subversion by the actual allocation of observable rewards and punishments. We agree, and this is a point we emphasize in the paper. Although we describe various ancillary components of religious belief that may buffer such insults, an adequate explanation may also entail the fact that religious beliefs are underwritten on a coarse-grained level by native cognitive dispositions that are difficult to set aside.

Nichols picks up on the issue of unbelief by raising the question of how the ancillary hypotheses we mention (like believe in afterlife punishment) could help solve the problem. If the problem is the cooperation-destabilizing effects of unbelieving defectors, scaring defectors with threats of eternal torment will not work, since they are the very ones who do not believe in something like hell anyway. The simple response is that afterlife punishment is not posited here to constrain the behavior of unbelievers but merely to help insulate belief against empirical disconfirmation. The more nuanced but important response is that there are actually two very different challenges posed by unbelief, having to do with the two different versions of SP. For PA accounts, believing has individual adaptive value because it reduces the likelihood of severe punishment upon defection. The problem is stabilizing belief in a moral god in the face of countervailing evidence. One solution is to add propositions that are not subject to empirical falsification, thereby securing the benefits to the believer of continued belief. If some choose defection over belief, in an environment where punishment is very costly, so much the worse for them. This poses no problem for the PA account of belief. In contrast to this, for CE accounts, disbelieving may benefit the individual at the group's expense. Nichols is right that the proposition of afterlife consequences may not be a solution to this problem, at least not without strong cultural group selection. (However, in the logic of Pascal's

wager it may be.) The important fix here is facilitating assortative interactions by recognition of defectors in belief, and we cited – and critiqued – various versions of signaling theory that have been proposed for this. Yet another response that works for both PA and CE accounts is to say that dispositions to believe are innate and may have arisen initially as spandrels. Here Nichols objects that this contrasts with the adaptationist approach of SP, but this seems to miss the fact that an expected trait comes to function adaptively.¹

Bulbulia and Freat advocate commitment signaling as an alternative account to SP. Although Bulbulia and Freat agree with our critique of SP, they disagree with our purported worry about whether signaling models are “possible or needed.” We are not aware of having such worries. We argue here and elsewhere (Schloss, 2007) that some form of signaling is probably needed to solve the problems of tertiary defection by facilitating detection of genuine commitment. Nor do we claim that signaling proposals are impossible. What we do claim is that costly signals are not necessarily reliable signals in virtue of their costs alone, and that the conditions that make a peacock’s tail reliable (contrary to Bulbulia and Freat’s citation of this icon) may not apply to costly religious practices. Moreover, we agree with and explicitly claim in the paper that unfakeable signals need not be costly. The question needing empirical resolution is which kinds of religious displays function adaptively under what circumstances (if any) – either as reliable signals of commitment or as manipulative signs (see Cronk, 1994). Finally, although we find commitment signaling models very plausible, Bulbulia and Freat have not given grounds for considering them an *alternative* to SP. First, it is not clear that their particular account of ecological signaling is an adaptation to defection. The twin challenges to all social exchange are coordination of cooperators and control of defectors. Ecological signaling may contribute chiefly to the former, in which case it is not an alternative to SP. Second, even if it is an adaptation to defection, it may supplement rather than substitute for SP.

De Aguiar and Cronk astutely propose that hierarchy and social stratification are environmental variables that should be considered in SP accounts. We agree this is theoretically plausible and empirically assessable. They also suggest that this has the potential to unify CE and PA approaches. Indeed, hierarchy may contribute to PA, since elites may manipulatively employ SP. They do not claim, but we would further propose, that powerful social elites may increase the actual costs of defection, which drives the need for PA. As we note in our paper, this generates a Nietzschean empirical prediction: divinely mandated cooperative other-regard should typify the religion of slaves, not masters. Their claims about CE constitute a more complicated issue. De Aguiar and Cronk point out that although CE enhances group cooperation, it need not be equal: perhaps elites benefit more. This is possible, but so is the opposite: especially under conditions of group selection, CE may attenuate variance in fitness related to stratification. In fact, in oscillations between the confrontation of religious elites by prophetic, revivalist, and reformationist movements and the routinization of charisma described in Weberian accounts of religious sociality, both dynamics may be at work. Finally, De Aguiar and Cronk note that the prominence of moralizing deities in large, hierarchical societies is consistent with what their account would predict. Although this is true, it is also concordant with standard CE accounts that make no reference to stratification and for which hierarchy may simply be a concomitant variable.

The next two responses – De Cruz and De Smedt and Shariff – move the conversation forward to the scale of selected replicator, with comparable concerns but very different arguments. De Cruz and De Smedt begin by noting that SP theories “pay little attention to the traits that are the targets of selection,” i.e., innate dispositions, learned strategies, or culturally transmitted norms. This is problematic on two counts: first, any account should begin not by asking which, but by asking whether religious traits are targets of selection at all, i.e., they may be spandrels. Second, the “trait” that theories of SP seek to explain in evolutionary terms and which may well be a phenotypic target of selection is simply belief in SP. The primary question is whether it serves an adaptive function and if so, what this function is. The proximal etiology of and replicational entity underlying this belief are secondary.

Now of course it could be that the nature of the replicator constrains the range of adaptive accounts that are plausible. Shariff argues this persuasively, but the three examples De Cruz and De Smedt give do not demonstrate this. First, De Cruz and De Smedt claim PA accounts are “not credible” in a nativist context since “at least some” hunter-gatherer societies practice low cost punishment. We too claim that PA suffers challenges from ethnography, but De Cruz and De Smedt’s point is not an example. “At least some” societies does not provide a picture of the prevailing selective environment, and even if it did, low cost to punisher does not equal low cost to punished. Second, they claim a flexible-strategy context renders some accounts of hunter-gatherer religion implausible because signaling membership in a religious community has no value if there is no religious choice, but neither version of SP theory purports to explain belief in SP in terms of a costly signal. Moreover, even if this were relevant, the adaptive value of a signal is not to declare membership in a particular religious community, but to convey commitment to the community one happens to be a member of – whether or not there are alternative membership options. Third, De Cruz and De Smedt claim that PA is not likely from the perspective of dual-inheritance theory because people in some societies can withstand the temptation to cheat without believing in supernatural sanctions. No advocate of SP claims that belief in punishing deities is the only route to reducing defection. The hypothesis is that it supplements other routes and that, under some selective regimes, it is the favored route. Indeed, far from making PA less plausible in light of variation between populations, we would expect such variation if belief in SP is culturally transmitted as dual-inheritance theory suggests.

Shariff provides a very well-argued defense of the CE thesis and a strong case for its link to selection at both cultural and group levels of scale. Although we agree with his three main conclusions, their underlying rationales merit brief comment. First, although the large cultural variability in religious beliefs argues convincingly for cultural over genetic sources, these are not mutually exclusive. The latter could provide coarse-grained dispositions that are both more finely-tuned and differentially transmitted at the cultural scale. Second, we agree that the CE account is rendered more plausible by group selection (in fact, it requires it). Indeed, the model of selective environments that Shariff uses to predict where PA and CE would prevail is entirely concordant with what we and Johnson propose, where PA emerges when defection costs > gains. We agree that the ethnographic evidence of SP’s prominence in later, larger cultures supports CE, and posit this to be a significant challenge to PA in our paper. However, an advocate of PA has two recourses. (a) The arrow of causality between moralizing gods and group size may run in the reverse direction:

moralizing gods is a prerequisite for rather than adaptation to large group size. Thus, although belief in big gods or other forms of cosmic moral sanctioning may have emerged in small-scale societies, those in which such beliefs arose grew and those without this adaptation to large scale sociality did not. (b) The likelihood of being detected (and hence the costs of defection) may be greater rather than less in larger cultures, because the risk-assessing mechanisms by which potential defectors assess the likelihood of detection may be attuned to the face-to-face context of small societies in which social cognition arose. Thus, the costs and the timescale of extensively mediated reputational consequences may be consistently underestimated, for which SP provides a correction. Johnson develops the issue of overconfidence, which may comport with, rather than be challenged by, the ethnographic data Shariff cites. Third, Shariff acknowledges the potentially destabilizing effects of disbelief for CE, but notes that religions try to deflect unbelief, through mechanisms that include well-documented anti-atheist prejudice (AAP). This is true, but the problem with unbelievers is not just the avowed atheists who are then prejudicially marginalized, it is also the religious posers who claim to believe but do not, and exclusionary mechanisms like AAP may exacerbate rather than ameliorate this. Moreover, the most potentially destabilizing issue is not even fakes, but the hypocritically self-deceived: those who really believe they believe, but shirk the costs entailed by consistent belief. It is precisely at this point that the commitment signaling models provide not an alternative, but a supplement to SP accounts of religious belief.

Finally, Johnson not only offers a credible defense for the plausibility of PA but also develops an expansive proposal for how PA and CE need not be mutually exclusive and may be complementary modes of adaptation. Although a detailed reply to his extensive analysis is out of reach, we can respond briefly because we are in substantial agreement. Each of the six major points he develops in the paper are plausible, and we accept (with some quibbles we will not indulge) numbers 1, 2, 4, and 6. In fact, many of the arguments he advances to support his points, we ourselves put forward in our paper. Here we suggest three qualifications to other points.

First, although we agree with Johnson's analysis of the role that overconfidence plays in PA accounts, we need to correct an important misunderstanding. Johnson claims that we argue "our more nuanced version," which incorporates cognitive bias, is significantly different from the original. We do believe that versions of PA that include notions of error management represent an important nuancing of the initial approach, but this refinement is not our insight – it is Johnson's (2009). We apologize if it appears we proposed this refinement as our own. That said, we offer the following qualification. Johnson claims that unless people overestimate the likelihood of getting away with defection, there is "no adaptive advantage to a fear of supernatural punishment." This is true and not true. While the individual selectionist perspective of PA does assume excessive confidence is the problem for which belief in omniscient punishers is a solution, CE approaches do not have this requirement. This difference constitutes an opportunity to empirically assess the accounts.

Second, we fully agree that for god-fearing (GF) to become established as a PA strategy, the cost of being detected and punished needs to exceed the cost of missed opportunities to defect. This simply constitutes an environmental condition necessary for GF to be favored by selection, and we did not state or intend to imply otherwise. The questions are: (a) are there empirical grounds for believing this condition has been met, and (b) are there theoretical grounds to believe it is likely? On the first point, we and Johnson agree that the requisite data are not yet available.

On the second point we suggest two refinements to Johnson's model, proposing that religious costs and the possibility of differential detectability between GF and Machiavellian (M) be factored in. Although Johnson countenances this, he argues that these additions would exert no net impact if religious costs were balanced by other kinds of benefits to religious commitment and if the increased likelihood of a GF being caught were equal to the reciprocal of the reduced likelihood of attempted defection. Absolutely true! Yet seemingly without justification: why make these assumptions about parameter values? We should be clear that unlike some critics of PA, our point is not to argue that the environmental scenario necessary for selection of GF is unlikely. However, we are unpersuaded by Johnson's claim that the model is "favorable" to conditions necessary for GF evolution. Our modest point is that although the model helpfully illuminates the conditions under which belief in SP will evolve, having no idea of the values for variables that inform it, we simply do not know how likely the requisite conditions are.

Third, Johnson argues that PA and CE need not be mutually exclusive accounts and posits that apparently contradictory elements can be reconciled by understanding belief in SP as a sequential adaptation to differing ecological conditions. This is plausible in principle, it may help solve problems with each independent account, and it is also complemented by aspects of De Aguiar and Cronk's analysis. In fact, we conclude our paper with a version of the same proposal. However, both ours and his face two serious questions. One is the historical issue of whether the kinds of moralizing supernatural agents PA requires, even in a coarse-grained sense, indeed typify Pleistocene societies. Advocates of CE point to ethnographic studies that suggest they do not. The other question is theoretical: the selective regimes posited by PA and CE are not just different, but at face value are countervailing. Shariff points out that the conditions favoring SP under CE should actually cause a reduction of such beliefs under PA. One response involves the possibility that belief in SP may represent a cognitive phenotype with mixed etiologies that are subject to orthogonal selective forces operating at individual and group levels, on genetic and cultural replicators. Thus, it is possible that genetically mediated cognitive dispositions to believe in SP were selected or exapted in conditions where PA conferred benefits to the individual, but as group size increased and these benefits declined, cultural group selection (Shariff) exploited receptivity to ideas of SP and facilitated large scale cooperation. However, far from resolving the issue of compatibility between PA and CE, this speculative if plausible and potentially fruitful scenario magnifies the present range of empirically under-determined theoretical options. At this point the comment that Michael Ruse made about evolutionary theories of morality 25 years ago applies to accounts of religion: "the question is not whether evolution is linked to [religion], but how," (Ruse, 1986, p. 95).

Note

1. We should comment on a point that, while not crucial to our treatment, is quite important to the general issue of religion and cooperation. Nichols claims that the paper "lacks a definition of cooperation" and that "humans are not as uniquely cooperative as many – especially proponents of 'strong reciprocity' – have claimed." Although terminological conventions are a subject of current debate (West, Griffin & Gardner, 2007), the conception of cooperation that is made explicit in the very first paragraph of the paper and that runs throughout is a widely prevailing one (Nowak, 2006): the exchange of benefits that is vulnerable to defection. This contrasts with byproduct mutualism, which

exchanges benefits without cost and therefore entails no added advantage to defecting, and altruism, in which net benefits are provided but not received. Contrary to Nichols' assertion, addressing the group vs individual selection debate does not require a more specific definition, since it can be engaged in terms of the contribution of cooperation and defection to intra- vs inter-group variance in fitness. Moreover, claims of the unique character of human cooperation to which many SP accounts are tied (and for that matter, morality and other cultural mechanisms of social control), do not hinge on strong reciprocity. Decades before this work E.O. Wilson described the unique nature of human cooperation as the "culminating mystery of all biology," (2000/1975, p. 382) and countless studies since then have identified the distinctive aspects of human cooperation most needing explanation as (a) the very existence of reciprocity and (b) the scale at which cooperative exchange occurs between non-kin (Alexander, 1987; Hauser, McAuliffe & Blake, 2009; Melix & Semmann, 2010; West et al., 2006).

References

- Alexander, R. (1987). *The biology of moral systems*. Piscataway, NJ: Aldine Transaction.
- Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon*, 29, 81–101.
- Hauser, M., McAuliffe, K., & Blake, P. (2009). Evolving the ingredients for reciprocity and spite. *Philosophical Transactions of the Royal Society*, 364, 3255–3266.
- Johnson, D.D.P. (2009). The error of God: Error management theory, religion, and the evolution of cooperation. In S.A. Levin (Ed.), *Games, groups, and the global good* (pp. 169–180). London: Springer.
- Melix, A., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B*, 365, 2663–2674.
- Nowak, M.A. (2006). Five rules for the evolution of cooperation. *Science*, 314, 1560–1563.
- Plotkin, H. (1997). *Darwin, machines and the nature of knowledge*. Cambridge, MA: Harvard University Press.
- Plotkin, H. (2000). *Evolution in mind*. Cambridge, MA: Harvard University Press.
- Ruse, M. (1986). Evolutionary ethics: A phoenix arisen. *Zygon*, 21, 95–112.
- Schloss, J.P. (2007). He who laughs best: Religious affect as a solution to recursive cooperative defection. In J. Bubulia, R. Sosis, E. Harris, R. Genet, C. Genet, & K. Wyman (Eds.), *The evolution of religion: Studies, theories, critiques* (pp. 205–215). Santa Margarita, CA: Collins Foundation Press.
- West, S., Gardner, A., Shuker, D., Reynolds, T., Burton-Chellow, M., Sykes, E., Guinee, M., & Griffin, A. (2006). Cooperation and the scale of competition in humans. *Current Biology*, 16, 1103–1106.
- West, S., Griffin, A., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity, and group selection. *Journal of Evolutionary Biology*, 20, 415–432.
- Wilson, E.O. (2000). *Sociobiology: The new synthesis*. Cambridge, MA: Harvard University Press.