

## 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

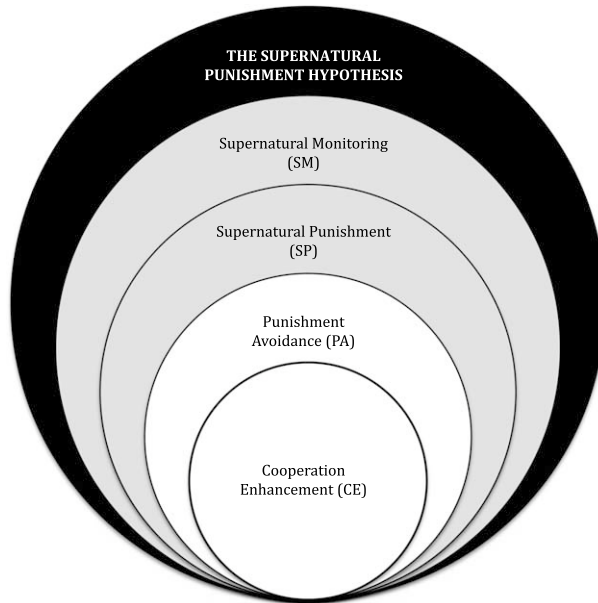


Figure 1. SPH (black circle) proposes that a fear of “supernatural” detection and punishment for selfish actions increases Darwinian fitness by reducing the “real-world” costs of norm transgressions. SPH requires two core “mechanisms” (in grey): supernatural monitoring (SM) and supernatural punishment (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

Downloaded By: [Sosis, Richard][Sosis, Richard] At: 23:02 14 June 2011

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.