

The Wrath of the Academics: Criticisms, Applications, and Extensions of the Supernatural Punishment Hypothesis

Dominic D. P. Johnson

University of Oxford

“The oldest and strongest emotion of mankind is fear, and the oldest
and strongest kind of fear is fear of the unknown.”

—H. P. Lovecraft

If you thought it was a tough life for religious believers living beneath the constant gaze of an omniscient and omnipresent God, spare a thought for the life of an academic. We spend years painstakingly putting together pieces of work that advance a niche field in tiny incremental steps, only to expose ourselves to the intense scrutiny of eagle-eyed peers whose job it is to pick through everything we say with a fine tooth-comb, ready to strike us down on any and all points.

It is thus with some trepidation that I read through the eight commentaries on *God is Watching You*, fearful of thunderbolts that might strike it down. Of course, along with fear there is always awe. It is a great honour when anyone reads one’s work, and a much greater one when they sit down to carefully critique it. In an age when we are bombarded from dawn to dusk with a blizzard of news, social media, blogs, journals, and new books it is a joy to reach anyone and a pleasure to provoke a response. I am therefore extremely grateful to all of the authors for their commentaries. They are remarkable for their diversity and brilliant ideas.

Happily, there were no thunderbolts, and instead we are lucky to find a series of thoughtful criticisms, applications, and extensions of the supernatural punishment hypothesis (hereafter, “SPH”; and supernatural punishment, “SP”). I am greatly reassured by the fact that the core idea of the book—that beliefs in supernatural punishment deter

selfishness and promote cooperation—seems to be a successful meme. While complexities remain, the empirical evidence is building and the idea is sticking. For now, the academic gods have made for fairly calm waters, but clearly there is an expanse of ocean surrounding the supernatural punishment hypothesis that needs exploration, mapping, and in particular an account of where, why, and how it varies from one place to the next. I hope the commentaries and this response will be the beginning of that journey rather than its end.

The Varieties of Supernatural Punishment: Universality and Variation in Afterlife Beliefs

Lloyd Black picks up on the book's oft-cited image of supernatural punishment beliefs as “genetic” and “universal”, and uses this as a foil to argue that, in fact, we should pay more attention to the wide cross-cultural *variation* in SP beliefs. Indeed, he suggests that the adaptive explanation for some *baseline* SP belief (one that all humans shared) might differ from the adaptive explanation for its *variation* over and above that baseline (for example, cognitive byproducts could be responsible for the former, but cultural evolution for the latter). This is an important point and one that currently lacks a good empirical test—where, when, and why do supernatural punishment beliefs *vary*? But I would argue that, at the theoretical level, we are in complete agreement. Black's commentary zeroes in on afterlife beliefs, in particular, as a specific trait on which to focus attention.

Variations on a Theme: The Supernatural Punishment Hypothesis *Predicts* Variation

I agree entirely with Black's core point, but would reframe it as a prediction, not a criticism, of the theory. The SPH *predicts* variation in the extent and form of beliefs across human populations. One of my obsessions—in the evolutionary study of human behavior in general—has been to stress the role of “behavioral ecology”, which is the recognition (and subfield of study) of how behavioral as well as physical traits *vary* across different environments in order to solve local adaptive problems. For example, northern peregrine falcons migrate (from freezing winters), but southern ones do not (where it is mild all year); waders forage together in the winter (to help find scarce food

sources) but alone in the summer (when defending breeding territories). Blanket strategies that do the same thing everywhere despite different prevailing conditions would be suboptimal in most places and disfavored by natural selection. This has been the core focus of the study of animal behavior over the last few decades (the discipline I trained in) and it applies just as importantly to humans (Davies et al., 2012; Rubenstein & Wrangham, 1986; Standen & Foley, 1989).

Unfortunately, for whatever reason, many people tend to think that if traits in humans are *evolved* traits, then they must be “fixed”, unchanging, and universal, which is of course nonsense. We all accept that physical and behavioral traits can vary significantly within populations of other animal species (“intraspecific variation”, such as the differential migration or social grouping examples above). But somehow applying evolution to humans tends to be perceived as implying that behavior is deterministic. This misunderstanding, common in the social sciences and often spilling over into anthropology, may partly explain the current (over?) excitement about cultural evolution. For many the concept of cultural evolution resolved a tension—how could evolution give rise to human traits *and* explain their diversity? If one tends to see evolved behavior as leading to fixed traits, biology on its own seems too limited. But good old-fashioned behavioral ecology does a great job of explaining much of the variation across cultures. Traits are manifested in different forms as adaptations to different socio-ecological settings, and natural selection is not blind to “cultural” practices that have good or bad outcomes for those that practice them. Where that is the case, the unit of selection may be a “cultural” trait, but the mechanism of selection is biological—natural selection favors those who express traits that improve Darwinian fitness.

The above is meant to make two points. First, complex beliefs in supernatural punishment *cannot* be genetic, it is obviously cultural in its narrative manifestation. Only its underlying cognitive mechanisms can be genetic. To me, this is obvious, but perhaps I thought it so self-evident that I did not emphasize it enough in the book. Second, *variation* in these cultural manifestations, however, can be understood within the framework of behavioral ecology, in biological selection for variants of traits that promote reproductive success within a given environment (and selection against variants which depress reproductive success).

This is not an argument against cultural evolution. Cultural evolution is just an example of behavioral ecology (since differential selection across populations can operate on either genetic or cultural traits). Moreover, where this variation occurs, it is likely to help, not hurt, the theory. In the case of supernatural punishment, cultural evolution can help to reinforce the deterrent, increase its effectiveness, and propel its spread. As Black notes, “culture may play an amplifying role in the propagation” of supernatural beliefs. I agree.

A behavioral ecological approach to the SPH suggests the following specific predictions (Johnson, 2016, pp.91-93): (1) SP beliefs are stronger and/or more widespread in environments or times when cooperation is more difficult or more in demand; and (2) SP beliefs are weaker and/or less widespread in environments or times when cooperation is more difficult or more in demand. For example, SP beliefs may be expected to be stronger in environments in which cooperation over resources is hard to achieve because of collective action problems and vulnerability to free-riding, and/or where cooperation is critical to wellbeing or survival. Several cross-cultural studies suggest that relevant beliefs are associated with situations where cooperation is indeed difficult or in demand (Dickson et al., 2005; Johnson, 2005; Peoples & Marlowe, 2012; Snarey, 1996; Sosis et al., 2007; Watts et al., 2015).

Afterlife or Afterthought: Does Behavior After Reproduction Matter for Natural Selection?

In the book I suggested an insight that might be a unique prediction of an *evolutionary* theory of religion: *afterlife* beliefs might be expected to be somewhat *less* powerful than this-worldly SP beliefs, because behavioral change due to beliefs about the afterlife cannot affect reproductive success (or, at least, affect it less, Johnson, 2016, p.84). Since the afterlife necessarily comes after reproduction, one’s genetic contribution to the next generation will occur prior to (and irrespective of) paying any consequences of one’s actions in the afterlife, and leaves time for atonement later. So from natural selection’s point of view, whether or not people are worried about what happens after death need not impact on their decisions until after reproductive age. This may be a noisy relationship

for many reasons, but the prediction stands that, from a purely biological point of view, if SP beliefs are to have an important impact on biological fitness, we might expect natural selection to have focused on beliefs about SP during one's lifetime (the perils of misfortune, droughts, crop failures, pests, disease, death, etc.). This leads to a nice prediction that someone should test: people before reproductive senescence should worry about this-life SP, while post-reproductive people should worry more about afterlife SP.

From my research for the book, I think the cross-cultural data support this greater role for this-worldly beliefs. For example, while SP in some form or other is pretty universal, it was much harder to find clear afterlife beliefs among indigenous and ancient civilizations. It was always easier to find examples of this-world SP, and there seemed to be more prevalent and to matter more in people's lives. It was also noticeable that in the growth of human civilizations, as hierarchies, elites, and cultural evolution became more important, afterlife beliefs become more prominent. Indeed, it seems to have emerged or become more salient later, especially versions that had an explicit divergence between good and bad destinations in the afterlife (Hultkrantz, 1967; Peoples & Marlowe, 2012). Going back to our previous discussion, this may have been no accident. As societies got bigger, the collective action problem became harder, and afterlife beliefs may have risen to solve them.

To summarize, it is important to push back a bit on the idea that the SPH implies a fixed, genetic adaptation. A lot of people seem to take this as a message of the book, and thus, as Black does, to conclude that such "a gene-centric theory would predict a universal application of supernatural punishment sensitivity" (he cites Ara Norenzayan on this point too). This interpretation is badly flawed. First, all that I claim is genetic is the cognitive mechanisms underlying such beliefs (Bering & Johnson, 2005). How they are manifested culturally is another issue altogether. But second, the SPH does not predict universality. Even if it was a purely genetic trait, that does *not* predict universality. Consider eye color, which is genetic. It does not lead to a universal eye color. Skin color, height, and a range of other features are examples of polymorphic genetic traits giving rise to variation from population to population and person to person in the production of enzymes, hormones, and proteins. Beyond this, we also have epigenetic processes by which genes can actually become switched on or off depending on the physical or social

environment and context (Agrawal, 2001). The brain and body are just massively variable and this is why we end up empirically with large individual variations in most traits. Just because something has a genetic basis, does not mean it is universal, or cannot account for variation. As Darwin realized, variation is critical to evolution in both its inputs (the process of selection among trait variants) and in its outputs (biological variation and diversity—among *and within* species).

We should *expect* (#1) SP to be widespread at some level (given underlying cognitive mechanisms that all humans share), but above this baseline we also should *expect* (#2) variation among individuals and populations. And critically, we should expect SP beliefs to *vary functionally* across contexts such that (#3) it is stronger and/or more prevalent where the demands or difficulty of cooperation are greater. These remain hypotheses to be tested in greater detail, across space, groups, or over time within a group (particularly with explicit measures of beliefs about supernatural *punishment* as the dependent variable). The book focused very much on collating evidence for predictions #1 and #2. Going forward, new research should now look at evidence for #3, to evaluate the extent to which SP beliefs are associated with cooperation among individuals and among cultures.

To the Fields: Supernatural Punishment and Risk-Pooling Among the Maasai

Cronk and Aktipis call our attention to a crucial role for supernatural punishment beliefs in stabilizing systems of “risk-pooling” in small-scale human societies, such as the Maasai of East Africa. In risk-pooling, individual A helps individual B with no expectation of return, as long as the latter is in need and the former is able to provide help. This means that whenever any individual in the community finds themselves in dire straits, they may rely on others for help, and thus the risk of starvation or malnutrition is spread (or “pooled”) among the community so that all can enjoy some baseline level of food security even in hard times. This form of cooperation is called “need-based transfer” (NBT). While it makes sense for the community as a whole, risk-pooling is a puzzle because the helper may never be repaid. So why do they do it? From a game-theoretical or evolutionary perspective, cooperation should be selected against. NBT differs markedly from the much more widely studied phenomenon of reciprocity (Trivers, 1971),

because there is no expectation of a return favor (of any kind at any point in the future). A gives to B but no debt is accumulated and no credit is earned.

The Bonds of Society: Supernatural Punishment and the Free-Rider Problem

Risk-pooling systems seem a great idea, but are fragile. This is because they suffer from high levels of instability resulting from the problem of free-riders. As Cronk and Aktipis explain, in need-based transfer systems it would be easy to feign need or deny help. Thus, free-riding ought to be rife on both sides of the exchange. In reciprocity, such free-riding is avoided because the repeated give and take over time means that cheats cannot prosper for long—they are simply denied further interactions. So while risk-pooling offers a significant potential service for the community, how are systems of need-based transfer underwritten?

One solution, Cronk and Aktipis propose, is the implied “threat of ... supernatural punishment to those who break the rules of the system”. Indeed, they found in their field studies that supernatural beliefs may be instrumental for need-based transfer. The Maasai themselves call the transfer system *Oсотua*, meaning “umbilical cord”, and thus carrying a powerful symbolism representing close interdependence and a concept imbued with a high degree of sacredness. When Maasai are asked to provide help in an *Oсотua* relationship, they are “obligated to do so” and report that it is “unthinkable for anyone to break the rules of *osotua*”.¹

This is a remarkable application of the supernatural punishment hypothesis to a specific cooperation problem, and one that is tightly linked to wellbeing and survival in small-scale communities like those in which humans evolved. Risk-pooling might have been a critical part of the success, spread, and growth of human societies, but a system so vulnerable to free-riding would have needed some additional mechanism, such as

¹ Interestingly, the sacred beliefs surrounding the *Oсотua* system are precisely what we might expect given the “uncertainty hypothesis”, in which superstition and rituals are especially prevalent surrounding events or activities that involve risk and uncertainty, as is risk-pooling against uncertain episodes of dangerous scarcity (Sosis & Handwerker, 2011). More predictable events can be managed instead by more secular means (e.g. by simply agreeing to divide up regular pies).

supernatural deterrence, to stabilize it. There may have been other ways to do so, but at least among the Maasai there is good evidence that the system is stabilized by the sacralization of the cooperative relationship and beliefs in the supernatural sanctions that would come with violating the system.

Cronk and Aktipis' commentary also highlights that we should pay attention to the *kind* of cooperation we are looking at (cooperation can emerge in a variety of forms: dyadic or multiplayer, mutual benefit or social dilemma, cooperation or coercion, reciprocity or risk-pooling, and so on). Supernatural punishment beliefs may be more likely (or more effective) in certain types of cooperation problems, such as those involving uncertainty or high levels of free-riding, and less likely in other forms of cooperation or social exchange, perhaps where more predictable problems can be solved by secular management and we don't need to knock on a supernatural agent's door for help (see Table 1 for some initial thoughts about how SP may apply differently among different types of cooperation problem). This suggests another important source of variation (among contexts or social activities, rather than among populations as a whole), and the prediction that supernatural punishment should not be observed everywhere (e.g. not in all cooperation problems, all types of environment, all times of the year, all communities, or even all religious sects). Rather, we should predict its presence only where it helps—bringing secular utility in solving problems, especially where other means are unavailable or have failed. The Maasai *Osotua* system appears to be precisely one such example.

Table 1. Different types of cooperation problems (or “games”) and how supernatural punishment may differentially affect and resolve them.

| Type of cooperation problem | Number of players | Free-rider problem | Utility of supernatural punishment | Example |
|-----------------------------|-------------------|--------------------|------------------------------------|---------------------------------|
| Dictator game | 2 | Absent | Low | Shariff & Norenzayan (2007) |
| Prisoners Dilemma | 2 | High | Moderate | Lane (this volume) |
| Public Goods Game | Several | High | High | O’Gorman <i>et al.</i> (2009) |
| Tragedy of the Commons | Many | High | High | Hartberg <i>et al.</i> (2016) |
| Risk-Pooling | Many | Severe | Critical | Cronk and Aktipis (this volume) |

Sheep in Wolf’s Clothing: How Fragile is the Maasai *Osotua* System?

I am in complete agreement with Cronk and Aktipis’s proposal that supernatural punishment may be an important stabilizing force in this special form of cooperation. Their insight does, however, raise some interesting questions that might warrant further thought.

First, if it is a *norm* to give to those in need (the Maasai say they are “obligated” to do so under the *Osotua* system), then individual A’s membership of the group means not only that: (a) they may sometimes have to give; but also that, (b) they *can* reasonably *expect* help from someone or other in the future if they are ever in need themselves. So, even from a purely self-interested point of view, I’d rather live in that group than live in the one down the road that has no such insurance scheme. If so, the puzzle seems to fade somewhat and one could consider it a form of “unrequited reciprocity”. One may not need to invoke *Osotua* charity for many years, or even for a whole lifetime, but it would still seem to be individually adaptive to live in a group with risk-pooling than one

without. For the average individual, the expected utility of paying a bit of *Osotua* occasionally (and one is only expected to do so if materially able to afford it) but insodoing avoiding starvation—just once—means that it seems to be an insurance premium worth paying. Needs-based transfers are defined as “systems of risk-pooling based on generosity toward those in need”, but how generous is it? (Of course, the question of how the *norm* of being obligated to help arose in the first place remains, and I suppose we would return to invoking the sacredness of *Osotua* and supernatural punishment in its origins, if not in its persistence).

Second, it may be that there *are* return benefits, perhaps invisible because they are delayed in time or returned in different forms (e.g. via deference, status, favored exchange partners, marriage opportunities, etc.). Research on *indirect* reciprocity and *image scoring* (in both models and experiments) suggests that people who are generous to others tend to accrue advantages later, even if much later, by third parties, and by other means (Milinski et al., 2002; Nowak & Sigmund, 1998). By giving, one not only gains a reputation as a good citizen that helps others, but also sends a signal that one is able to give, and thus wealthy, resourceful, or successful. Such signals may bring important returns in attracting coalition partners, status, and reproductive opportunities. If one is relatively well off, it may also make sense to spread the wealth in order to avoid losing fellow group members, or to help swell the group’s numbers. Being part of a larger group (in sheer numbers of individuals) can be an advantage to my own self-interest for at least three reasons: (1) to increase future risk-pooling capacity (the more people there are to help, the less likely any one individual is to go hungry); (2) to increase the power of collective activities within the community (e.g. hunting, foraging, farming, construction, divisions of labor); and (3) to improve deterrence, defense, and offense in inter-group competition (safety and power in numbers).

Finally, Cronk and Aktipis’s paper suggests some extremely interesting predictions that could be tested:

- Is it true in the Maasai that *other* forms of sharing and cooperation (i.e., those not prone to uncertainty and instability) are *not* surrounded by superstitious or supernatural beliefs?

- Among the 8 other societies mentioned that form their larger project, is it the case that societies facing problems in which cooperation is more difficult or more in demand are also more likely to have supernatural beliefs associated with them?
- Does Christian charity and Islamic alms-giving represent risk-pooling, and is this why they are imbued with supernatural sanctions? (the Koran specifies giving alms as one of the most important, required obligations for Muslims and failing to do so risks divine consequences, e.g. see Homerin, 2006).

Cronk and Aktipis have shown characteristic rigor in melding theory and fieldwork, and have identified an important form of cooperation, not discussed in the book, and yet for which supernatural punishment may be a vital stabilizing mechanism. For systems of mutual cooperation that provide insurance against the vagaries of life, yet are vulnerable to the dangers of free-riding, God may be the great underwriter.

Steadying the Ship: Robustness of the Supernatural Punishment Hypothesis in the Face of Skeptics

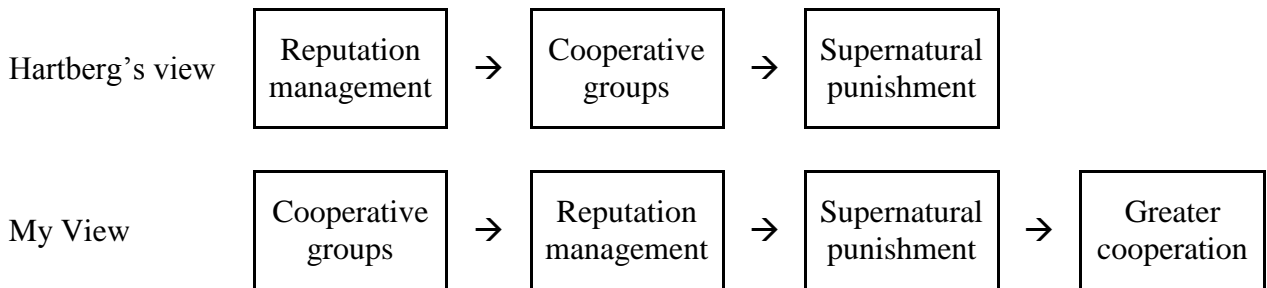
Yasha Hartberg raises the “robustness” of cultural systems, and how this may apply to the SPH. He is broadly supportive of the SPH, and indeed has provided some of the best empirical evidence supporting it in his study of 48 small-scale societies in which supernatural punishment served to promote cooperation over valuable common pool resources (a remarkable religious twist on Elinor Ostrom’s work on the effectiveness of indigenous solutions to collective action problems (Hartberg et al., 2016; Ostrom, 1990). Still, Hartberg is curious about *why* it works. Since belief in SP constrains self-interest, why don’t skeptics spread and undermine the system?

Which Came First, the Chicken or the God?

Hartberg’s broader point is that SP only works among already “highly cooperative groups with established norms” (since it is the sanctions for violating these norms that SP helps to avoid). Thus, “reputation management only becomes a problem once norms have been established”. In short, he worries that SPH is circular: “supernatural punishment increases cooperation by protecting us from consequences that accrue from living in the highly

cooperative groups that supernatural punishment is proposed to make possible in the first place.” This, he argues, represents a conundrum. Hartberg himself suggests that this is not necessarily a problem, since both cooperation and religion can co-evolve together, rather than one having to precede the other in toto, so he notes it and moves on. But I think there are at least 3 reasons to reject the charge of circularity (see Table 2 for the contrasting views).

Table 2. Hartberg’s view and my view of which features of the SPH came first.



First, humans already lived in groups *before* there was any religion. Group-living is typical of primates in general, especially among our closest relatives—chimpanzees, bonobos and gorillas. Ethnographic and archeological evidence suggest that in our own lineage we have lived in groups for millions of years, as far back as we can tell to the split with our common ancestor. The evidence also suggests these groups must have been highly cooperative for a long time (e.g. in communal living, foraging, child-rearing, hunting, and warfare). Thus, humans were already in cooperative groups long *before* religion came along.

Second, neither reputation management nor punishment are unique to humans. Both are common among non-human animals (Clutton-Brock & Parker, 1995). In many animal societies, “punishment”, in the form of costly aggression imposed on one animal by another, is widespread, and certainly does not need sophisticated human-like “norms” to emerge as a prerequisite (certainly not ones relying on complex language). Indeed, some of these species do not even cooperate with each other at all (e.g., they coalesce into groups due to ecological factors such as clumped resources or predation pressure).

Therefore, even the problem of real-world punishment from other group members predated the evolution of religion (and indeed humans!). There is, therefore, no conundrum: first came groups, then came real-world punishment, then came religion.

As I argue, what did happen is that the problem of reputation management *became much harder* as humans became cognitively and linguistically sophisticated, and that is when belief in supernatural punishment served to offer an adaptive solution. Worrying about supernatural consequences of one's actions reduced the probability of violating existing social norms and their detection. But in so doing, SP only *enhances* cooperation, it did not ever "make [it] possible in the first place", as ants or bacteria could attest.

Problems: The Robustness of Cultural Systems

While a SP system should work well once up and running (everyone believes, and everyone enjoys the fruits of the resulting higher levels of cooperation), Hartberg argues that we need to consider how cooperative groups emerge in the first place, and how they survive the threat of non-believers. Hartberg suggests that this is SPH's "Achilles heel", since "the institution would seem to be extremely fragile to skeptics".

In theory, this is a concern. If skeptics have greater Darwinian fitness than believers, then the former should spread at the expense of the latter. Yet, as he notes, this has clearly not happened—supernatural beliefs of some form or another are more or less universal across all cultures around the globe. So how has the system been so robust to skeptics? One answer is that skeptics actually fare poorly at the hands of natural selection, so there is no puzzle after all. In the book, I explicitly argue that skeptics may *not* have greater Darwinian fitness than believers. I evaluated the comparative advantage of non-believers ("Machiavellians") versus believers ("God-Fearers") and deduced that, as long as the costs of real-world sanctions for self-interested behavior are greater than the opportunity costs of refraining from them, God-Fearers will outcompete non-believers and spread. This is because the threat of SP makes believers forego some opportunities for personal gain but in so doing they avoid incurring the (greater) sanctions of fellow group members.

So theoretically, we have a potential solution. Of course, *empirically*, it is very hard to evaluate the relative costs and benefits of belief and non-belief. Moreover, these may vary across time, individuals, and environments. Yet, the point is that while skeptics are interesting and remain under-theorized (Johnson, 2012), there are good reasons to believe that believers are robust against them. This is further—and perhaps most powerfully of all—supported by the phenomenon of “positive assortment”. People can *choose* who they interact with, so believers can of course choose to preferentially interact with co-believers and avoid non-believers. This may be a central way in which religion works, and is in fact central to many theories of religion (e.g. costly signaling, Sosis & Alcorta, 2003).

Solutions: Proximate Mechanisms of Robustness

Having posed the problem of how SP stays robust in the face of skeptics. Hartberg does us a favor by laying out some compelling solutions (“proximate mechanisms” based on empirical phenomena observed in small-scale societies). He suggests that skeptics may be deterred or dissuaded because: (1) SP has “nonspecific penalties” (and thus any common event—storms, illness, bad luck—can be attributed to the gods); (2) SP has “nonspecific targets” (and thus many peoples’ misfortune can be traced back to a violation by someone else); (3) SP enjoys “communal exegesis” in that groups tend to gossip and publicly attribute natural events to violations of norms or taboos;² and (4) “evolved psychology” makes the conceptualization of SP easy and skepticism hard. These offer a variety of additional mechanisms that may shore up the robustness of SP, and I agree.

Let me take the opportunity to propose some more: (5) skeptics have often historically been shunned, ostracized, or executed (so their numbers and influence can be contained); (6) in isolated groups in the pre-scientific past, skeptics lacked any alternative belief systems or explanations for events to adopt instead (so perhaps they are common today but rare in our past); (7) Pascal’s wager might have been prominent in our evolutionary history, since while one might be skeptical about the tribe’s beliefs in

² And he brilliantly notes that this mechanism not only builds the cause-effect case for SP, but also alerts past or would-be violators to the fact that the *community* is watching, as well as the gods.

supernatural causation in trivial matters of everyday life, without a better explanation for misfortunes, illness or death, would you risk it when the stakes were high?; (8) as long as there is a majority of believers in the group, the system as a whole can be robust (i.e., a few skeptics need not destabilize it); (9) a few skeptics might, counter-intuitively, serve to *reinforce* the stability of the system, by demanding clearer and better justifications and arguments—as Richard Dawkins appears to do for many believers today (Johnson, 2012); (10) skeptics still have the fundamental human cognitive mechanisms *underlying* religious beliefs, which are common to us all, so even if their beliefs are weaker and limited, skeptics can still be deterred from selfishness and enticed to cooperate. None of these 10 reasons may guarantee the robustness of the SPH on their own, but in combination they suggest an explanation for why skeptics have generally failed to take over any society ever. They are out there, but they do not unravel the community of believers. There's only 500 million of them today, compared with 6.5 billion believers (Zuckerman, 2008), and that minority is in decline.

Interestingly, Hartberg's (and my) solutions focus on *religious* or *psychological* traits that may offer robustness against skeptics. But robustness can also derive from *organizational structure* and the properties of *complex adaptive systems* themselves. General insights about the conditions for robustness in biological or physical systems may be paralleled in religious systems, and these would be well worth exploring too (Sosis, In Press, Miller & Page, 2007).

Raindrops Keep Falling on My Head: Remaining Skeptical About Skeptics

People often raise the question of skeptics in evolutionary theories of religion, but it remains unclear to what extent they were present in human history and, to whatever extent they were present, what damage that did to the robustness of religion. Hartberg gives the example of 5th Century Diagoros of Milos, who abandoned his faith when the gods failed to punish a false accuser. However, the problem is that the lessons of such life events can work in the opposite direction. When things don't go according to plan, people may: (1) renounce their belief, like Diagoros (since justice seemed not to be done); (2) reinforce their belief, like Job (since they had not deserved the outcome, they must redouble their efforts to understand or beseech the gods); or (3) adopt a theology that god

cannot or will not always intervene, so injustices or “evil” in the world need not challenge anyone’s faith. A new book suggests that skepticism and “atheism” (which used to mean rejecting a particular version of religion as well as any god) were actually quite common in the classical world—at least among prominent philosophers and writers (Whitmarsh, 2016). There are no-doubt many such cases, but also reverse cases where “atheists” became believers—Emperor Constantine’s conversion to Christianity for one. If anything, skeptics seem to stand out precisely because they are islands in a sea of believers. Skeptics are only “skeptics” because they are skeptical of some wider view held by the majority. So almost by definition, skeptics are only exceptions that prove a more general rule. Perhaps in the future, a few die-hard believers will be the new “skeptics”. But if that ever happens, it is a long way off.

Of course, from an adaptationist perspective we might be worried if skepticism was widespread in the small-scale societies that offer a window onto our evolutionary past. Perhaps they reported religious ideas and myths but often didn’t really believe them. I looked for evidence of exactly this while writing the book, but such examples seemed to be rare rather than common, and were certainly drowned out by the great significance of religion for most societies and the seriousness with which supernatural beliefs were considered and acted upon by individuals.

Hartberg gives one example from a small-scale society in Indonesia, where a man named Muda Duni was skeptical about whether the ancestors really punish people for taking bamboo from sacred forests by supposedly unleashing storms and floods. He takes a great load of bamboo home from one such forest and is relieved to find only a few raindrops falling on his head. The gods, it seems, are weak. For me, however, this is evidence of belief, not doubt! First, it is one man doubting a majority belief in his society at large (the latter of which is the crucial sample, especially in multi-player collective action problems like the conservation of forests). Second, the man himself is quoted as saying that the fact that some rain did fall “confirmed the continuing presence of supernatural spirits” and “proved that the ancestors are still protecting their sacred resources and still have the ability to punish people for desecrating them”. The only aspect suggesting doubt was that the spirits appeared to have a “weakening power” because there were only a few drops of rain rather than a flood. Well OK, but Hartberg

offers this as “one of the most vivid” examples of skepticism in their studies of 48 societies in diverse societies and environments around the world. If that is the most vivid example of skepticism—one person and one instance, who anyway reiterated the norm and attributed the rain to supernatural agents—then the SPH seems firmly vindicated!

I don’t doubt the presence of skeptics in any society, but I do doubt that this few would present a significant counter-current to the beliefs of the group as a whole. If anything, religions just *adapt* to skepticism about particular features of belief or practice, but are not undermined by them. Indeed, they are often strengthened by them instead. In the middle ages Galileo doubted that the sun orbited around the Earth, which the Roman Inquisition of 1615 determined was “foolish and absurd in philosophy, and formally heretical since it explicitly contradicts in many places the sense of Holy Scripture”, but few Christians would defend that view any more. Skepticism emerged and Christianity adapted. Skeptics rarely threaten religion, they just challenge it—and that’s usually good for the vitality and persistence of the religion as a whole. So rather than skeptics undermining the robustness of religious systems, as Hartberg began, it may be that skeptics are part of the solution.

A Model God: Supernatural Punishment Survives and Thrives in an Agent Based Model

Justin Lane does something I’ve been hoping someone would do for a while—explore the SPH using computer simulations. Using Axelrod and Hamilton’s (1981) iterated Prisoner’s dilemma as a baseline model, Lane adds the possibility for agents to anticipate supernatural punishment for defecting. This belief deters them from defecting by some probability, allowing us to explore the effects of this novel trait on which strategies tend to spread over time. Lane finds that SP beliefs not only survive but also thrive across simulated conditions, with a variety of effects on the dominant strategies in play. So, overall, a belief in supernatural punishment evolves and is somehow adaptive in the model. The key question is why that is the case.

The Big Picture: The Success of Supernatural Punishment Beliefs

In the model, agents play an iterated Prisoner's Dilemma game, in which they are randomly paired each round to another agent. Agents start with (and can inherit) one of 5 strategies: (1) Always Cooperate (ALL C); (2) Tit-for-tat (TFT; start by cooperating, then do whatever the other player did in the previous round); (3) Sometimes Cooperate (C with $p = 0.3$); (4) Mostly Defect (C with $p = 0.6$); and (5) Always Defect (ALL D). In addition, agents can either believe or not believe in *supernatural* punishment for defecting ("defecting" means choosing *not* to cooperate in the game), on top of any material consequences of defecting. Believers reduce their probability of defection accordingly, since they expect greater net costs for doing so. At intervals, the best performing strategy replaces poorly performing strategies in proportion to its success.

The basic result is that the trait—"belief in supernatural punishment"—not only survives in the face of competitors that do not anticipate non-material consequences of their actions, but actually thrives across all of the models' simulated conditions. This in itself is surprising. Why should a strategy that imagines additional costs of self-interested behavior that are not really there do better than strategies that see the world as it is? Let us work through the logic to understand why this makes sense, and here I offer my own interpretation.

Axelrod and Hamilton's original model found that Tit-for-Tat (TFT, which notably *start* by cooperating and then do whatever the other player did on the last round), emerged as the evolutionarily stable strategy (ESS; an ESS is one that is robust and cannot be overturned by some alternative strategy emerging. However, Axelrod and Hamilton also noted that there was another possible ESS strategy that could emerge (and once in place, again could not be invaded by any other strategy), which was ALL D. In ALL D, every player always defects no matter what the other player does. This is where Lane's finding becomes especially interesting.

Axelrod and Hamilton puzzled over the fact that their model could result in either TFT or ALL D, and therefore wondered "how an evolutionary trend to cooperative behavior could ever have started in the first place". They proposed a solution based on one of the most important concepts in evolutionary biology: kin selection. They noted that, if players are genetically *related* to some extent, then there would be a slight edge to

the TFT strategy over ALL D, since by helping others you are helping copies of your own genes in other individuals, and this could get cooperation of the ground (cooperative genes would be selected for). I don't doubt that explanation. But Lane's model suggests a novel one: cooperation can evolve *even without kinship*, if instead a belief in supernatural punishment fills the role of steering people towards a slightly cooperative disposition to begin with. Intriguingly, this suggests that, while kin selection may have been one lever by which cooperation got started, beliefs in supernatural punishment affecting *individual* behavior could be another.

Given the game—an iterated Prisoner's Dilemma in which the reciprocal strategy of Tit-for-tat does well—the problem to solve in real life is how to get players to cooperate initially so that pairs of TFT players don't descend into an endless cycle of reciprocal defection. Agents that anticipate supernatural consequences for self-interested behaviors (that is, believers in SP) are more likely to cooperate, and in a world of competing strategies where agents retaliate against others' self-interested acts (as in TFT and as theorized in the SPH), being disposed towards generosity is an advantageous trait. Perhaps counter-intuitively, in a world of blind retaliators, the altruist is king.

Variations: Supernatural Punishment Under Different Model Conditions

Beyond the finding that supernatural punishment beliefs persist and thrive in the model, Lane reports specific model conditions which generate variation in: (1) the prevalence of the SP trait; and (2) the resulting strategies that emerge.

First, *belief* in supernatural punishment spreads more widely when individual transgressors are deemed to be punished, rather than when the group as a whole (or no one) is punished.

Second, the *strategy* that becomes dominant depends on the purported *target* of supernatural punishment. TFT spreads more widely when it is *individual transgressors* that perceive themselves to be the targets of supernatural punishment. If punishment is absent or perceived to effect the entire *group*, then ALL D becomes the dominant strategy (no one cooperates). As Lane notes, this result suggests that mechanisms of individual selection acting on religious belief may be especially important to the evolution of cooperation, while group-level punishment leads to widespread defection.

Third, Lane anticipated that, since believers are more likely to cooperate, as belief spreads, TFT should decrease in the population relative to cooperative strategies. In fact, however, ALL D remained the most common (just, at 34%), followed by TFT (29%), and then Mostly C (23%). However, I would be more interested to know what the correlation is at the *individual-agent level*. All we know is that, at the population level, there were X believers and Y proportion of each strategy. But which strategies were *believers* tending to play (compared to non-believers)? This would be more illuminating than the raw proportions of strategies across populations, with believers somehow scattered among them. While we don't know this level of detail, Figure 5 shows that in by far the majority of simulations, belief in supernatural punishment has spread to fixation (accounting for over half of all simulations—about 12,000 of the total 22,500 simulations). Belief in SP clearly did well, even if it remains unclear which strategies believers are adopting to achieve this.

Fourth, Lane notes that there is no correlation between the proportion of the simulated populations that cooperate and the proportion believing in SP. However, these are again raw comparisons across whole populations. We do not have the vital statistics on *within-population correlations*. That is, within any single simulated population (or over the entire population of all agents simulated), is there a correlation between individual X's SP belief and the strategy individual X adopted? Unfortunately this critical information is not known, or not presented.

Finally, Lane notes that there is a relatively low frequency of ALL C across conditions, counter to his expectation. Given that the *trait* of supernatural punishment is prevalent across conditions, he expected the prevalence of cooperation to increase just as much. I suggest two reasons why we should not expect this. The first is the problem raised in the above paragraphs—looking across population averages as a whole may not be telling us much, since we need to know whether an *individual's belief* in SP correlates with *that individual's strategy*. Only then can we judge whether SP beliefs are associated with more or less cooperation overall. Second, while Lane expects ALL C to dominate if SP is prevalent, I suggest we should in fact expect TFT to persist as well. This is because the SPH *requires* secular punishment by other real-world actors—exactly as happens in TFT—and it is this threat of material retaliation that imposes a cost on social

transgression. It is *because* of these costs that supernatural punishment beliefs becomes useful (reducing the probability of committing such social transgressions, and the material consequences of retaliation by another player). Therefore, some TFT in the population may in fact be a pre-requisite for SP beliefs to persist—without real-world dangers, there's nothing for the gods to help you avoid. I therefore *predict* a mix of TFT and cooperative strategies in the population. Interestingly, and in line with empirical evidence, not everyone has to be a punisher of social transgressions (O'Gorman et al., 2009), but as long as the possibility exists in the form of *some* TFT players lurking around, it pays to avoid their wrath.

Fair Game? Why Would Believers Cooperate Randomly with Non-Believers?

One of Lane's puzzles is why, if supernatural punishment beliefs spread so pervasively, we do not observe more cooperation. I suggest that one important reason is that the model picks *random* interaction partners from the population at each iteration of the game—pitting believers and non-believers together, as well as different strategies. In real life, interactions tend to be directed towards fellow in-group members, and away from outsiders, a phenomenon called “positive assortment” which is well established theoretically and empirically (Aktipis, 2011). This is true of groups in general, but is argued to be particularly important for *religious* groups (Wilson, 2002). Believers will tend to seek out and cooperate with other believers and avoid non-believers. Once factored into the model, I suspect that this would dramatically increase the level of cooperation among believers (and cooperation may drop significantly among non-believers, who no longer benefit from having cooperatively disposed believers around to exploit). This would bring the results back into line with Lane's expectation.

The Only Game in Town? Beyond the Prisoner's Dilemma

An outstanding broader question is whether the PD is the right game to model the evolution of religion. The PD is arguably one of the most important social dilemmas, approximating many real life scenarios, and in its multiple player forms also represents widespread collective action problems. However, PD is not the only game in town. Across the range of human social interactions, we play PDs, but also a variety of other

cooperation games including Chicken, Stag Hunt, Coordination, Hawk-Dove, Mutualism, Public Goods Games, Tragedies of the Commons, and so on. We should therefore consider: (1) which of these games is the best “model” for cooperation among early human societies; and (2) whether SP might have different effects in each such game (or would even be needed in some of them—as Cronk and Aktipis note, reciprocal altruism, for example, has its own built in punishment system, since defection is automatically reciprocated). In future modeling, it would be interesting to explore alternative games, or even what happens when individuals play a mixture of games (see e.g. Panchanathan & Boyd, 2004).

An Alternative Setup: Modeling from the Ground Up

A final question is whether this or future models could use a different formulation, as proposed in the book. Lane says that the book “never goes beyond a narrative formulation to a form that could serve as abstract rules for human interactions”. However, I did explicitly lay out a basic game theoretical formulation for the conditions under which “god-fearing” could evolve (see Johnson, 2016, Table 6.1 on p.167). This showed that, all else equal, belief in supernatural punishment can evolve when “the total expected costs of selfishness—namely, the probability of detection (p) multiplied by the cost of exposure (c)—is greater than the cost of missed opportunities for selfish rewards (m)”. Lane’s approach is interesting, especially for its comparison with the established Axelrod and Hamilton model, but to me it diverges somewhat from the rules and the “spirit” of the game I envisaged in the book, not least in its anchoring to the two-player iterated Prisoner’s Dilemma.

In the book I envisaged more of a collective action problem among groups of individuals, wherein selfish behavior would incur real-world punishment costs from the community as a whole. These real-world punishment costs, which are *the* critical drivers of selection for SP beliefs (as argued in Ch 6 of the book), are omitted from Lane’s model (see his p7). This is puzzling, especially as Lane claims that “the model here does not include mechanisms of non-supernatural second order punishment; this is deliberate because it is proposed to be unnecessary given the SPH.” Lane seems to infer that, once SP beliefs are established as a norm in a society, then supernatural agents conduct the

punishing and real people do not need to. But this is not the case (or at least not what the SPH argues). Instead, I argue that actual or threatened secular punishment by other human beings *provides* the selection pressure for a belief in SP. That belief reduces the probability of committing violations of social norms (especially ones that are likely to be detected), and thus believers may enjoy higher Darwinian fitness as a result. This critical aspect—the heart of the SPH which makes supernatural beliefs materially important—is not captured by the model. Yet Lane notes himself at the end that “an amendment addressing this sort of punishment may be required.” By the end of my scrutiny of Lane’s model, I was less concerned about this than I might have been, because the retaliation inherent to the TFT strategy does provide at least some “real-world” punishment costs for transgression (and, as noted above, this may explain why SP and TFT coexist—even coevolve—in the model). The problem is whether real world punishment of defectors should be built into the model itself, rather than relying on it being an (inadvertent) aspect of just one of the five possible strategies.

Modeling the evolution of supernatural punishment beliefs brings some striking new insights as well as raising new questions. Lane has done a great service in demonstrating that supernatural punishment beliefs thrive and spread within a famous existing model of the evolution of cooperation, and opening up a new avenue in the exploration of the supernatural punishment hypothesis.

Fear Itself: Is There a Problem with Terror?

Hillary Lenfesty and Thomas Fikes are concerned that punishment creates “cultures of anxiety”, which is not only counter-productive to social harmony, but draws on only one side of human physiological mechanisms. As well as being deterred or coerced by punishment, we can also be encouraged and incentivized by positive social mechanisms. So isn’t punishment an unnecessarily destructive way to nudge people’s behavior?

Fear can certainly take its toll. Earlier this year an Israeli man from Haifa requested a restraining order against God. He claimed that God had been excessively vindictive towards him. Over a three-year period, the police had been called to his home ten times, and repeated requests for a hearing to consider a restraining order was finally granted. The man represented himself in court while, as the *Times of Israel* reported, “A

protocol of the hearing noted that God did not turn up for the session, although it did not specify how the court determined the Omnipresent was not in fact there, as opposed to merely exercising the right to remain silent.” The judge called the case “ludicrous”, but acknowledged that the claimant may need help, but not from the law.³ The man in question, psychiatrically ill or not, may well be suffering from precisely the kind of chronic levels of fear and anxiety that Lenfesty and Fikes focus on. The case is obviously unusual, but one wonders how many other people experience mental distress about the idea of an omniscient, omnipresent being watching over their shoulder the whole time. Many people recount such fears from childhood religious education—Richard Dawkins among them.

In fact, it may be quite common—especially among certain types of people. Nava Siltan et al. (2013) found that people who believe in an “angry god” are statistically more likely to show signs of suffering from mental disorders. Data on 1426 US adults revealed that, controlling for demographic characteristics, religiousness, and strength of belief in God, belief in a “punitive God” was associated with psychiatric anxiety disorders (such as general anxiety, social anxiety, paranoia, obsession, and compulsion). Belief in a benevolent God, by contrast, was associated with fewer such disorders and, in the authors’ words, “almost protecting against psychopathology”.

Conditions for Success: Is Fear Bad for Business?

In short, I agree that SP might well be a serious source of fear and anxiety, and there is clearly some empirical evidence to support it. The (separate) question is whether that undermines the SPH. I will argue that, if anything, it reinforces it.

First, in evolutionary perspective, we may not *enjoy* or *want* to be fearful and anxious, but natural selection doesn’t care whether we feel happy or not. If a trait increases reproductive success, it will prosper and spread. A mouse lives in fear of predation all the time, but that’s a good disposition for survival. So there is no theoretical reason to think that the psychological discomfort of anxiety is a flaw in the hypothesis.

³ Stewart Winer, “Man seeks restraining order against God”, *Times of Israel*, 4 May 2016.

Second, and more importantly, if supernatural punishment induces fear or anxiety, maybe that *increases* its effectiveness (rather than compromising it). Indeed, this would help to explain why it has spread and stuck among so many cultures across the globe, and why it became so accentuated in the major world religions.

Third, in the case of SP, we don't *all* have to live in fear—only those who have transgressed (or might transgress) need to be fearful. So it may be worth scaring the bejesus out of the few for the good of the many.

Fourth, anxiety is experienced (or tolerated) differently by different people, as Siltan et. al.'s study suggests. Detrimental effects may be limited to, or only severe in, a subset of people, which need not undermine the stability or effectiveness of SP as a whole. Indeed, Hartberg suggested above that many people don't take the threat of supernatural punishment too seriously in everyday life anyway (although they might do when the stakes are high).

Fifth, leader and follower relationships may complicate the role of fear and punishment. It is in leaders' interests for their followers to fear god (if god and the leader have similar policies), or to fear leaders themselves *as* a god. If people are anxious and fearful as a result, all the better. Compared to relatively egalitarian hunter-gatherer societies, the development of large societies and authoritarian control may have elevated the dread of omniscient gods out of all proportion for the poor old trembling human mammal. Fear may be unfortunate, but despite the crippling anxiety it may induce (or precisely because of it), is deliberately emphasized by those whose purposes it serves.

Which Game Are We Playing? Punishment is a Good Tool for the Right Job

Lenfesty and Fikes broader point is that, while supernatural punishment may be effective, it only describes half of human beings. The other half, they suggest, is the positive, prosocial disposition and “neighborliness” that also “makes us human”. They point to empirical evidence of prosocial dispositions as well as underlying cognitive mechanisms that humans have to promote it.

I do not disagree with the general point. My book was not saying: punishment is the only reason humans cooperate and without it they never do. For example, if Donald Trump and Hillary Clinton were stuck in a cave by a rock that only two people could move, they would have no trouble cooperating to escape. That is cooperation for mutual benefit and there is no puzzle. I would therefore reiterate the point that we have to pay attention to the particular game or social problem we are facing. Prosociality may be adaptive for many kinds of kin and neighbor interactions (where there are direct or indirect return benefits). But prosociality *per se* cannot solve the prisoner's dilemma or the collective action problem, because indiscriminate generosity is a self-defeating strategy that would be exploited by defectors and annihilated. Some other mechanism must be in play to protect cooperators (such as partner choice, reputation management, or punishment).

Cronk and Aktipis gave us a nice example of this. They point out that systems of reciprocity have their own internal checks and balances against defectors (shunning of past defectors), where as risk-pooling is so wide open to cheating that some other mechanism such as supernatural punishment is needed to deter cheats. In short, the relative importance and effectiveness of punishment (deterrence) and prosociality (encouragement) depends on the game or domain. We should not expect one tool to fix all problems of cooperation. My book focused very much on deterring self-interested behavior and promoting cooperation in collective action problems. It is in these domains that punishment is particularly important and effective. Rewards *are* present, but—as we know—fail to produce or sustain cooperation on their own.

No Foxholes in an Atheist: Does God Beget Fear, or Fear Beget God?

Lenfesty and Fikes argue that “reinforcement” (both positive and negative) can be superior to punishment from a *behavioral* perspective, although agreeing that punishment has an edge over reward from a *cognitive* perspective (as per Baumeister et al., 2001). But they worry that the literature is already running away too quickly with the supernatural punishment paradigm, while overlooking its shortcomings, namely “the fragility of any cooperative truces that may be established, and the fact that threat

activates a fight-or-flight mechanism that encourages either escape and avoidance, or – if evasion is not possible – produces chronic states of anxiety and fear.” We have addressed the fragility/robustness problem above, with Cronk and Aktipis and Hartberg. Let us focus on the problem of inducing anxiety and fear.

I would argue that, if the threat of supernatural punishment induces escape, avoidance, anxiety and fear, that is all the more reason to think it is effective. It may be uncomfortable, undesirable, or draining on mental resources, but all those things may be ingredients of its success. If Lenfesty and Fikes are right and SP beliefs lead to a disabling level of chronic anxiety, then one would predict that the SPH is not an effective solution to solving cooperation problems (“living under the constant threat of a wrathful god would undermine humans’ evolved physiology for social engagement” and social enterprises would suffer, not succeed). First, it is not entirely clear whether such anxiety represents damage to Darwinian fitness or merely damage to psychological wellbeing. I would argue that only the former matters for an evolutionary theory. Second, in any case empirical evidence suggests the opposite effect, and indeed causal arrows running in the opposite direction: in real world situations that generate anxiety, we observe an *increase* in the strength of religious beliefs and religious rituals. The caricature is “no atheists in foxholes”, but the effect seems widespread. Lanman and others have shown how religiosity increases under situations of insecurity (Whitehouse & Lanman, 2014). In general, superstition has been shown to become elevated in situations of uncertainty, high stakes, lack of control, stress, and anxiety (viz, the “uncertainty hypothesis”). The argument is that this is in fact adaptive, and supernatural beliefs help people to cope with the situation and perform better within it. Sosis and Handwerker (2011), for example, found not only that prayer recitation was more prevalent among people living in areas of Israel at high risk of missile attacks, but that prayer recitation was explicitly associated with a subsequent *decrease in measured levels of stress* among individuals. Finally, one might also point to “Terror Management Theory” in psychology (TMT) which, though perhaps an overstated phenomenon, represents a vast literature identifying the fear of death as a foundational yet subconscious motivation underlying numerous aspects of human cognition and behavior. Anxiety may be undesirable, but it seems to have

powerful effects on people's behavior and more importantly, it appears to motivate adaptive responses.

Mechanisms: Physiological, Psychological, and Cultural Promoters of Prosociality

Thus far, Lenfesty and Fikes probably remain unsatisfied because, while I've argued that anxiety works, and may motivate behavior, they'd argue that—even if so—we still downplay powerful mechanisms promoting positive behavior. So let us explore the role these positive mechanisms play.

Lenfesty and Fikes outline both *physiological and cultural mechanisms* that underlie prosociality. Traditionally, the autonomic nervous system, which operates unconsciously in humans to regulate basic physiological functions, is divided into the sympathetic nervous system, or “fight or flight” response, and the parasympathetic nervous system, or “rest and digest” response. Following Stephen Porges’ “Polyvagal Theory”, Lenfesty and Fikes suggest that humans have a third system—the “social engagement system”—that serves to calm and sooth to promote social interactions at a very basic level in human physiology. In short, they argue, “our bodies are adapted for social engagement”. I do not disagree with this. However, I would highlight the point that these systems are not blanket strategies, nor do they come into play as a kind of average sum of their parts. Rather, they are different strategic responses to different adaptive problems. Therefore, any of the three may come to the fore depending on the context, the interaction partner, the group, and the game being played. Social engagement may work well to promote cooperation among kin and long-term allies, but it would be a dangerous strategy to employ in the face of potential cheats. I thus reiterate my main response above, which is that social engagement is good in some contexts, but not in all. Prosocial agents are always vulnerable to exploitation. As we know, in collective action problems or Prisoner's Dilemma, straight prosociality is not enough. Indeed, it would be a suicidal strategy.

Lenfesty and Fikes then move on to other proximate mechanisms in *social learning and cultural norms*, which they suggest are as much about fairness, reciprocity, and trustworthiness, of compassion and kindness, commitment and generosity than they

are about punishment and fear. But it is important not to mistake “norms” for mechanisms. Norms are prescriptions for how people “should” behave, often precisely because left to their own devices this is *not* how people will behave. Secular and religious doctrines commonly advocate such prosocial traits precisely because they are good for the wider community, but rely on *individuals* setting aside their own self-interest. While many may be motivated to follow the norm, the problem is that we need mechanisms to shield those cooperators from exploitation—even if there are only a few exploiters around. As long as there is the threat of free-riders, punishment is one way of providing stability for norms of cooperation. One might question whether such negative mechanisms are really or always necessary. But every society in the world has discovered for themselves that they need laws, police, courts, and jails. Whatever prosocial inclinations human beings have, they are clearly not strong enough for society to run on its own. As James Madison put it, “If men were angels, no government would be necessary.”

This begs a larger question: If we are to accept that humans are prosocial and neighborly, in both physiology and behavior, then why is cooperation so hard to achieve? Why do we recurrently face the free-rider problem and the collective action problem? Why do we have such detailed social sanctions and institutions to enforce them? Cheats should be rare and cooperation and generosity should be effervescent. Evidently, cooperation is hard and cheats and free-riders are ubiquitous. Even in indigenous societies, Elinor Ostrom found that the sharing of common pool resources among small-scale kith and kin groups relied on strong institutions to deter self-interest and free-riding. As psychologist David Barash observed, “people are widely urged to be kind, moral, altruistic, and so forth, which suggests that they are basically less kind, moral, altruistic, etc., than is desired.” This seems especially salient to religion, where doctrine continuously urges moral behavior which would be lost on us if we were angels already.

To sum up this section, Lenfesty and Fikes stress that “‘doing good’ engages different physiological and psychological systems and draws on different cultural norms than refraining from ‘doing bad’ ” (p.7). I do not disagree. I stressed in the book that punishment on its own is meaningless, since it is only a *means* to achieve some greater cooperative *end*. Humans benefit immensely from the rewards of cooperation, and it is

this higher level of social interaction that we strive for. Punishment, however, is a ladder that gets us up there. Most of us may be motivated to cooperate by the potential benefits that await, but for the system to work the fundamental problem is to deter the (perhaps few or infrequent) cheats and free-riders that block the path to cooperation.

Lenfesty and Fikes quote Porges himself as recognizing that, “only in a safe environment is it adaptive and appropriate to simultaneously inhibit defense systems and exhibit positive social engagement behavior”, which leads to the prediction that insecurity *activates* defense systems. This is precisely what Lanman, Sosis and others found among religious groups and societies. In our evolutionary past, it may be that religion helped to solve precisely the cooperation problems that arose in domains of insecurity—food, weather, reproduction, illness, death, war and so on. On issues that made people rightly anxious, religion, and in particular supernatural punishment, offered a tool close to hand to leverage cooperation from the jaws of self-interest. Lenfesty and Fikes are right to explore the other side of the coin from fear and punishment—the more positive aspects of human psychology and physiology that can promote cooperation and generosity. But we should be careful not to have too idealized a view of human nature. It is not clear what Lenfesty and Fikes’ own views on religion may be, but they certainly seem to believe in angels.

Supernatural Punishment on Mars: Human Beliefs and Beliefs About Humans

McKay, Ross, O’Lone and Efferson accept my core argument that selfish behavior and social transgressions became especially costly over human evolutionary history, and beliefs in supernatural punishment may have helped to inhibit them. However, they “play devil’s advocate” in challenging the idea that a belief in supernatural punishment would be necessary, or better at avoiding such costly mistakes than an unbiased adjustment to prevailing real-world risks. Their core argument is that “Bayesian decision-makers” (agents who update their beliefs over time to reflect changes in actual costs, benefits, and probabilities) would do better than God-Fearers who harbored imaginary beliefs that supernatural agents are watching them.

My response is five-fold: (1) human beings are *not* Bayesian decision-makers (far from it), so McKay et al.'s entire logic is based on a model of intelligent beings that we might find on Mars but not on Earth; (2) even a Martian decision-maker or robot would, I argue, still make more behavioral *mistakes* (as opposed to cognitive *errors*) that would make supernatural beliefs adaptive even for them; (3) the relevant information in the environment on the costs, benefits, and probabilities (of both detection and punishment), are highly uncertain, variable, and unavailable to decision-makers, and under such conditions of poor information, heuristic biases help to steer a safe course through uncharted waters; (4) even if accurate information were available, natural selection has had to work with a human brain that does not function like a rational computer, and heuristic biases offered effective solutions because they are faster, demand less processing power, and are readily available; (5) empirically, belief in supernatural punishment *is* an exaggerated overestimate of being watched (that was the starting point, rather than the conclusion of the book), so the challenge is not to explain why it should *not* exist, but to explain why it does.

First of all, let me stress that McKay et al. focus on the argument that people believe in supernatural punishment *as a consequence of Error Management Theory* (EMT, Haselton & Buss, 2009; Johnson et al., 2013). However, the SPH does not rely on EMT—it is only one of a set of possible mechanisms. A belief in supernatural punishment may spread because of: (1) conventional competition among strategies (the costs and benefits of belief outperforming the costs and benefits of non-belief, as per my table comparing alternative strategies in the book on p.167); and/or (2) EMT (an *exaggerated* perception of surveillance helping to avoid the real-world costs of social transgressions); and/or (3) some other reason (such as elite coercion, or cultural selection for stabilizing norms). EMT is interesting because it offers an explicit mechanism for why the exaggerated perception of supernatural agency, as manifested in our so-called Hyper-Active Agency Detection Device (HADD, Barrett, 2000), might in itself be adaptive. But belief in supernatural punishment can evolve even if there is no error management. The SPH, therefore, does not stand or fall on arguments about EMT. Having said that, let me defend the EMT approach since McKay *et al.* chose to attack it.

Is Supernatural Punishment the *Only* Way to Avoid Errors?

The first aspect of their critique is easily cleared up. McKay et al., present my thinking as that “*only* exaggerated estimates—erroneous beliefs—help to avoid costly errors”. This is a false reading of the argument. Of course rational decision-makers are perfectly able to avoid costly errors by erring on the side of caution, without any exaggerated beliefs (e.g. people drive on the right side of the road not because they exaggerate the chance of the police spotting the infringement, but because if they drove on the wrong side they would certainly crash!). So we agree on this basic point. The passage McKay et al. quote from book (p.169) was intended to explain that, given a situation in which the cost of transgressing (if caught) exceeds its benefit (if not caught), a bias will only help you if it acts in the *direction* of avoiding the former outcome rather than the latter (that is, *exaggerating* surveillance, not underestimating it). Hence, “*only exaggerated* estimates ... will help you to avoid the worst of the two errors” (Johnson, 2016, p.169, original italics). The key emphasis is on “exaggerated”, not “only”. McKay et al. read this as “*only* exaggerated estimates ... will help you to avoid errors” *in general*, which is incorrect—there are of course other means to this end. (McKay et al. altered the italics of the original [“*only exaggerated*”] into their own italics [“*only exaggerated*”], changing the meaning of the argument).

On p.167-168 of the book I gave a more fundamental, *non*-EMT explanation of why belief in supernatural punishment may outperform non-believers, which is equivalent to McKay et al.’s Russian roulette example. Essentially, natural selection adds up the material costs and benefits over time and God-Fearers will win “*as long as* the total expected costs of selfishness—namely, the probability of detection (p) multiplied by the cost of exposure (c)—is greater than the cost of missed opportunities for selfish rewards (m)” (that is, when $pc > m$; p.167, emphasis in original). In short, material probabilities decide the day, and “God-Fearers” are just rational beings who avoid shooting themselves in the head (in my formulation on p.167 even God-Fearers do not *exaggerate* the probability of being watched, they just weigh up the real-world, empirical costs, benefits, and probabilities of the possible outcomes, and rightly believe it is a dangerous game and avoid risking the consequences—like McKay playing Russian roulette). To be clear, God-Fearers’ calculations are based purely on material

consequences. That only leaves the question of why I call them “God-Fearers” rather than, say, rational automatons. The answer is that these are the conditions under which God-Fearing could evolve (and in the book I argued that if deeply entrenched self-interest meant humans were in danger of taking too many social risks, then God-Fearing provided a useful corrective to balance the books). Over a beer, I believe I could convince McKay et al., that we are on exactly the same page here (in their critique they are stuck on page 169, and do not see the bigger picture). Only after laying that groundwork did I go on to consider the added explanatory value of EMT.

The EMT explanation pops out as important because, if transgressing is costly but humans are impulsive, we need help to avoid transgressing even when in the calm light of day we should be able to calculate that doing so would be costly (evidence the many people in jail who regret what they did). We also face the problem of being able to evaluate p , c , and m correctly (evidence the many other people in jail who might not regret what they did but underestimated the chances of getting caught). One can make an argument that even McKay’s rational beings playing Russian roulette would, in fact, benefit from a cognitive bias, because a better analogy to real life decisions may be a gun with 1000 chambers (instead of 6) and a prize of survival or reproduction (instead of 1 dollar). Their stylized case is useful for illustrative purposes but rarely do we know such probabilities in real-life, especially with such precision (in their example, an exact $4/6$ chance of winning 1 dollar, and an exact $2/6$ chance of death). In a case of 1000 chambers, it is much harder to estimate the precise expected utilities of the different outcomes (who knows precisely how many chambers there are in a given “gun”, or precisely what the positive and negative consequences will turn out to be in a given scenario?).

The two conditions recognized as prerequisites for EMT to operate (Johnson et al., 2013)—(a) asymmetric costs (i.e., missing out on a small prize versus death), and (b) uncertainty (i.e., will the gun fire a bullet or a blank?)—are present in both cases above. However, in McKay’s analogy the probabilities are precisely known and the cost asymmetry extreme, so any idiot should make the right decision. By contrast, in my real-life example the probabilities are not known and the cost asymmetries are far less clear. Natural selection may “know” them in terms of average success rates over time, and thus

select for an optimal decision threshold. But the humans it gives rise to must still have some cognitive mechanism to put those statistical odds into action. So, we can either postulate a Martian decision-maker with perfect information and an unrestricted ability to integrate complex statistical odds, or we can postulate an imperfect human being with cognitive heuristics to guide us in making reasonable choices under uncertainty. Many such examples of the latter have been documented in the literature (Johnson et al., 2013), but I know of no examples of the former.

Given the reality of cognitive constraints (limits on information and processing power) and evolutionary constraints (limits on efficiency, speed, evolvability, and the “adaptive landscape” allowing some kinds of adaptive change more easily than others), humans are far from Bayesian decision-makers. Indeed, Nobel Laureates Daniel Kahneman and Amos Tversky concluded from their vast empirical observations that humans are “not Bayesian at all” (Kahneman & Tversky 1982, p. 46).

Cognitive biases seem to have been favored in our evolution over precise, case-by-case cost-benefit calculations because they have lower operating costs, were more easily available to natural selection, and are capable of reaching decisions faster, especially in novel, complex, uncertain, or fast moving situations. Empirically, there are many examples of biases in human judgment and decision-making that seem to be adaptive precisely because they offer simple heuristics that guide us (or indeed often deceive us) into fitness-maximizing behavior. McKay et al. admit as much on their final page when they qualify their conclusions with: “when considering a single decision-making domain *without constraints on the cognitive processes available*” (my italics). This may be the end point for economics (the wall they hit at the end of their road, which is laid on foundational assumptions about rational choice), but it is the starting point for cognitive psychology. Given the many evolutionary and cognitive constraints that the human brain has been under, cognitive biases are effective decision-making strategies that natural selection has utilized to maximize Darwinian fitness under the imperfect conditions of life on Earth.

Is Supernatural Punishment the *Best* Way to Avoid Errors?

Having argued that a Martian could avoid decision-making errors just as well as a god-fearing human being, McKay et al., go on to consider whether god-fearing is a *better* way to do so. Here, they slightly but crucially misunderstand the original argument again. They quote, “The best solution to avoiding detection... is a mechanism that overestimates the true probability that detection will occur—*exaggerated* estimates outperform *accurate* estimates, because the latter will engender more [costly] mistakes” (p.169; with original italics). They then launch into an argument that what matters is not minimizing the number of errors, but minimizing the overall costs of the errors made. Absolutely correct. However, their footnote 3 betrays their misunderstanding in what this means for the SPH. It was *they* who added the qualifier of “[costly]” mistakes in square brackets in the quote above, but this is incorrect—the original was intended. It is vital to distinguish between “errors” and “mistakes”: (1) *errors* are incorrect estimates of the true probability of some event (e.g. of getting caught cheating); while (2) *mistakes* are choosing the more costly behavior as a result of the estimate in #1 (e.g. deciding to steal when you will tend to be caught). Counter-intuitively, perhaps, *errors* can lead to making either the right decision or the wrong decision. Here’s an example. Let’s say there is a 75% chance of rain. If you estimate there is a 100% chance of rain, when the real probability is 75%, then you have made an error, but you nevertheless made the right decision to take your umbrella with you to work. By contrast, if you estimate there is a 0% chance of rain, when the real probability is 75%, then you have again made an error but this time you also made a mistake—leaving the umbrella behind and likely getting drenched).⁴ With this crucial distinction now in mind, here’s my original argument again: exaggerated estimates (e.g. God-Fearers) and accurate estimates (e.g. Martians) make *the same number of errors*, but they make *a different number of mistakes*. Where $pc > m$, the Martians lose.

This discrepancy also explains why McKay et al. are enchanted by their Figure 1, which describes *behaviors* (mistakes), not *beliefs* (errors). Their claim that “Unbiased beliefs ensure that one never chooses the option with the lower expected payoff ... [and] Biased beliefs ... ensure that one will, with some positive probability, choose

⁴ Note that, even if it did not rain, you still made a mistake not to take your umbrella because the chance of rain was better than even. On overage, this strategy would not pay off.

suboptimally” is precisely incorrect. The opposite is true. Unbiased beliefs make X errors but 50% of them lead to mistakes (the behavior that incurs the greater cost). Biased beliefs also make X errors *but* only a much lower proportion (0% – 49%) of these errors lead to mistakes (see e.g., Haselton & Nettle, 2006; Nettle, 2004). Hence, a biased belief is useful precisely because it shifts the scatter of errors “above the line” into the domain where the right *decision* is usually made, even if the *probability estimate* is wrong (and hence making fewer mistakes). Unbiased beliefs ensure that the scatter of errors fall both “above and below the line” and thus lead to the wrong decision (mistakes) half of the time. Thus, supernatural punishment may not be the “best” of all possible ways to avoid social transgressions, but in conditions of poor information and uncertainty, such biases can be effective heuristics for adaptive behavior (Cosmides & Tooby, 1994).

And Never the Twain Shall Meet: Human and Alien Natures

One can forget all of the above and consider something more fundamental. What is our model of human nature? McKay et al. explicitly state that, “because someone with Bayesian beliefs has integrated all available evidence in the optimal way, we are explicitly considering a decision maker *without* a cognitive bias.” Herein lies the central problem: humans are not Bayesian decision-makers (so the model is wrong theoretically), and humans do in fact demonstrate a range of cognitive biases, including specifically a bias in overestimating the presence of supernatural agency (so the model is also wrong empirically). McKay et al.’s examination of whether Bayesian decision-makers would benefit from a belief in God is an interesting exercise, but a theoretically and empirically flawed model with which to understand *Homo sapiens*. McKay et al. assume that humans are (a) rational actors, (b) have perfect information, and (c) update accurately based on this information. I would argue that all of these assumptions are false. This is why the people they describe seem more like Martians to me, and are certainly far removed from the imperfect, emotional, cognitively biased, believing primates we know and love on Earth. Our challenge is to explain why the perception of supernatural agency exists, not to explain why something we observe empirically should be absent. But I thank the devil for sending his advocates to test the argument.

Supernatural Punishment Works, But Why? Adaptive and Superior or Accidental and Inferior?

Azim Shariff notes that, perhaps more than anyone else in the field, we share a lot of the same views on the role and importance of supernatural punishment. He has been at the forefront of going out and actually testing whether supernatural punishment does in fact promote cooperative behavior, and has provided some remarkable experimental and cross-cultural evidence. He remains skeptical about two aspects of the theory, however: (1) whether supernatural punishment beliefs have been subject to natural selection rather than cultural selection; and (2) whether supernatural punishment can match secular punishment in its effectiveness. As he points out, both of these debates have implications beyond mere academic interest—they make different predictions for if and how religious beliefs may spread and remain useful in the future. Let me respond to each in turn.

Naturally Selected Adaptation or Culturally Selected Byproduct? Actually, Both

Shariff suggests that the SPH contains the assumption that supernatural punishment beliefs is a “modular genetically-selected adaptation” that occurs spontaneously in human beings. I’m slightly puzzled by this characterization because, as I noted in response to Lloyd Black above, I disagree. I don’t think we are as far apart on this issue as people seem to believe. We can break the problem down into two parts: (1) are supernatural punishment beliefs a genetic or cultural?; and (2) are supernatural punishment beliefs an adaptation or a byproduct?

The first is easy to answer. Specific beliefs in supernatural punishment (such as belief in Hell or Zeus) *cannot* be a genetic trait. No detailed cultural idea can be genetic. Rather: (1) the underlying cognitive dispositions that *enable* religious beliefs and behavior—e.g., sensitivity to threats, theory of mind, cause and effect reasoning—are -genetic adaptations, while (2) the specific beliefs that supernatural agency X carries out supernatural punishment Y for doing activity Z are *cultural* phenomena. As Shariff notes, it is important to qualify this distinction with the point that #2 (cultural phenomena) can in part be explained by #1 (genetic predispositions), because, for example, what we tend to pay attention to, learn, and remember, are themselves influenced by underlying genetic

cognitive adaptations. Thus, whatever supernatural agents we believe in, many of them tend to have common characteristics. Still, that only goes so far, we absolutely need cultural evolution to explain how explicit religious beliefs have emerged and spread over time.

I would argue that the above is more or less a consensus in the field (though often misunderstood, Sosis, 2009), so why does Shariff (and others above) continue to believe that the SPH implies a genetic adaptation? I think the answer is because, in the book, I argue that there are individual Darwinian fitness benefits from belief in SP (that is, SP beliefs increase reproductive success). If so, that might seem to imply that individual A's beliefs in God X is subject to natural selection at the genetic level. However, I would put it differently. The SPH doesn't care *which* God or *which* form of punishment individual A believes. That doesn't matter. What matters is that individual A anticipates supernatural consequences—of some kind or another—for his actions. Thus, what I and the book intended to argue is that belief in supernatural consequences (in general, and the underlying cognitive mechanisms giving rise to them) may have been subject to natural selection. But not the content of those beliefs. If supernatural punishment beliefs vary among people, are heritable, and have consequences for reproductive fitness, then they could not fail to be subject to natural selection (and, if they promote fitness, they will spread). As Darwin noted, we only need some degree of variation, some level of heritability, and some mechanism of replication for evolution to work. We know religious beliefs vary empirically, we know religious beliefs tend to have high heritability, and we know that people reproduce. Hence, we don't need to invoke cultural evolution at all to explain the origin and spread of the *underlying cognitive mechanisms* that give rise to religious beliefs. Foster and Kokko (2009) even argue that such mechanisms have been widely selected for among *animals*, let alone humans. Overestimating cause and effect, especially for negative outcomes, is generally thought to be adaptive (Baumeister et al., 2001; Rozin & Royzman, 2001).

To some extent I would argue that this logic extends to cross-cultural variation as well. For example, many non-human animal species have very different physiological and behavioral traits in different parts of their range. We don't need culture to explain that. There is thus no reason to assume, in principle, that trait variation across human

populations needs a cultural explanation. For example, Central African pygmy houses have doors that are quite small compared to the doors of American houses, but this does not have (or need) a cultural explanation. It's because pygmies are smaller. How might this apply to religion? Snarey (1996), for example, found that moralizing gods were more likely in environments where water was scarce (and thus people had to share more to survive). The idea that gods are "moralizing" is clearly cultural, but the variation itself can be explained by ecological circumstances (behavioral ecology). I therefore suspect that cultural explanations are being given too easily for behavioral variation that can be chalked up more simply to local adaptations to prevailing ecological conditions. Where these adaptations promote survival or reproductive success, they will be favored by natural selection (not cultural selection). One may argue that such variation itself represents "culture". This is just like variation in behavioral traits within the populations of numerous other species and natural selection can explain it perfectly well.

Universal and Born Believers: Beware the False Prophets

Shariff attacks two lines of evidence that are supposed to support the adaptationist approach: (1) universality (everyone believes in supernatural punishment); and (2) developmental psychology (all children develop beliefs in supernatural punishment). He argues the first is challenged by variation, since different cultures vary in the form and extent to which they believe in supernatural punishment. However, this is in fact precisely what I would predict. Evolutionary theory does not predict a trait to occur at the same level everywhere, because of (1) naturally occurring variation in traits (no two individuals are identical); and (2) the above argument that traits are adaptations to local (not global) adaptive problems, so they will only reach high levels or only emerge where they contribute to fitness (and will fade out or be selected against where they do not). In the case of supernatural punishment, as I stressed in the book (Johnson, 2016, p.92):

"Darwin's theory of evolution was as much about variation in characteristics, selection among those variants, and the generation of new variants, as it was about any universal features of life. My theory is that beliefs in supernatural punishment were favored by selection because they helped to avoid the costs of selfish behavior and promote cooperation. If so, then in places or times when cooperation was difficult or in demand,

we may expect supernatural punishment to come to the fore. And in places or times when the fruits of cooperation could be harvested just as well or better by other means, supernatural punishment may be superfluous or even counterproductive. An adaptive theory would therefore *expect* it to be reduced or absent in some circumstances, and accentuated in others (whether the adaptation arises from biological or cultural selection). In short, if belief in supernatural punishment has evolutionary causes and consequences, then it should vary depending on our prevailing social and physical “ecology”—just as other functional biological and cultural traits do.”

Shariff specifically notes that “belief in monitoring, moralizing, and punishing gods is not universal, but highly culturally variable ...[and] that variation appears to covary with prosocial behavior”. This is precisely what the SPH predicts, yet he suggests this is “troubling for any theory that suggests humans have universally evolved to hold faith in these specific types of gods”. The SPH is certainly no such theory!

As for the developmental psychology argument, I would refer back to my earlier distinction between *underlying cognitive mechanisms* versus *culturally explicit content*. I do not believe children are born to believe in punishment by any particular god—they could not. However, I do believe we are born with underlying cognitive mechanisms that predispose us to believe in supernatural as well as natural explanations. Here, I am not saying anything different from the standard cognitive science of religion literature (Barrett, 2012; Bloom, 2004; McCauley, 2011). One consequence of these underlying mechanisms is that we tend to ascribe cause and effect to life events, even when there is none. Combined with the negativity bias (Baumeister et al., 2001) and negative agency bias (Morewedge, 2009), we are primed to believe that bad things, in particular, happen for a reason.

Having argued for the importance of *natural* selection in explaining the origins, spread, and variation in supernatural punishment beliefs, there are clearly aspects of human beliefs and behaviors that cannot be explained by natural selection alone, and here I have no problem invoking cultural selection. To explain the *content* of beliefs (and the *context*, in terms of social structures and role models, as Shariff points out), we also need to understand how cultural traits vary, stick, and are passed on. One crucial difference is that natural selection only favors traits that increase biological fitness, whereas cultural

selection can favor traits that do not necessarily increase fitness. So cultural evolution is especially important for explaining human beliefs and behaviors that are counter-productive.

In short, I think we agree that: (1) *general* beliefs underlying supernatural punishment are made possible by basic cognitive tendencies (which are genetic adaptations resulting from natural selection); and (2) *specific* beliefs about supernatural punishment are made possible by cultural narratives (which are byproducts resulting from cultural selection); (3) the forms that these specific supernatural punishment beliefs take is partly a product of content biases (programmed learning), partly selective pressures of the local environment (behavioral ecology), and partly a result of the spread of ideas and practices (cultural evolution).

What Works Best: Supernatural or Secular Punishment?

Shariff's second critique is of the notion that supernatural punishment is *more* powerful than secular punishment. In the book I argue that there are indeed reasons to believe that, in certain circumstances, supernatural punishment may be more effective than secular punishment. This may sound silly to an atheist—how could imaginary beliefs be more effective than the cold reality of material punishment? But in fact there are good reasons to believe this is the case.

First, if people firmly believe that they will be punished in Hell for all eternity then there is no conceivable secular punishment that could come close to matching that. Thus, for a true believer, supernatural punishment is clearly the most powerful deterrent imaginable. The question just becomes *to what extent* in real life a given individual really believes that, or to what extent they may anticipate being able to *atone for their sins* before judgment is passed. This will vary among people, and for some it may bring down the threat of SP below whatever deterrent is offered by secular means. For others, it will maintain an edge.

Second, from a game theoretical perspective, SP may remain powerful not because people believe it is so severe, but because it is more effective in catching them. The police can't be watching all the time, but God can. So even if one is somewhat skeptical that God will punish you all that much for sin X, at least sin X will have a high

probability of being detected and punished. So the overall effect of SP beliefs in deterring selfish behavior and solving the collective action problem can be high even if the extent of punishment itself is limited. I give some other reasons in the book (Johnson, 2016, pp. 71-73 and 227-229).

But let us address Shariff's critique in more detail. In the book I illustrated the point with a curious natural experiment: when the police went on strike in Montreal in 1965, crime soared. My point was that humans will pursue self-interest where they can, and lifting secular justice demonstrated that people would express their potential for free-riding and exploitation. Shariff counters that, if anything, the Montreal case just shows that supernatural punishment doesn't work either, since the crimes were committed while the police were absent but God was still there (at least for believers). However, this problem can be turned around at Shariff. While supernatural punishment clearly did not *prevent all* citizens from committing crime (of course, it rarely does), neither did the absence of the police *prompt all* citizens to transgress. So while secular punishment seems to have been an effective deterrent for some subset of the population (who changed their behavior while the police were gone), Montreal also demonstrated that secular punishment is not necessary for the majority of people—all the rest who remained good. Why did that great majority of the population stay at home and abide by the law? The threat of supernatural punishment may help explain why 99% of them stayed good (as well as, of course, fear of delayed secular punishment and other factors).

A similar inference fallacy comes into play in his discussion of experimental cooperation games. When economists allow punishment, cooperation increases. Shariff notes that, "In the conditions without this (earthly) punishment, the threat of supernatural punishment was insufficient at preventing freeriding from taking over." That is evidently true, but supernatural beliefs were not measured in these experiments (Johnson et al., 2003), so absence of evidence about the role of supernatural punishment is not evidence of absence. When such games have been replicated by others, religious beliefs *have* in fact been associated with cooperation behavior, as I explore below.

As Shariff notes, a problem for the theory has been the lack of studies looking at explicit beliefs in supernatural *punishment* as opposed to religious beliefs or religiosity in general. This lacunae is gradually being addressed, not least by Shariff's own studies, but

also by others. For example, our old friend Ryan McKay and colleagues found that priming people with religious concepts made some of them *more* likely to personally carry out sanctions against other people who were behaving unfairly, and—notably for our discussion here—supernatural primes were *more* effective than secular primes (McKay et al., 2010). This does not tell us whether people in general were more likely to cooperate themselves if they believed in punishing gods, but Shariff’s own experiments suggest that is precisely the case! (Norenzayan & Shariff, 2008; Shariff & Norenzayan, 2007; Shariff & Norenzayan, 2011; Shariff et al., 2009; Shariff & Rhemtulla, 2012). New cross-cultural studies by others also provide strong support for the hypothesis that supernatural punishment, explicitly, is associated with greater levels of cooperation (Purzycki et al., 2012; Purzycki et al., 2016; Watts et al., 2015).

Shariff also suggests that the effects of supernatural punishment beliefs may be quite fleeting, working while people are primed or attentive to them, but lapsing back into normal behavior quickly afterwards. Where this is the case, I would argue that: (1) this is precisely the reason religious doctrine instills such beliefs into followers via narratives and rituals that are repeated over and over again (we do indeed need to be reminded); and (2) people’s behavior is likely to contemplate supernatural consequences more carefully when the stakes are high. In lab experiments for a few dollars, God is perhaps unlikely to mind too much what we do (perhaps he is just as keen to thwart the economists). But in high stakes decisions with consequences for survival and reproduction, supernatural consequences may carry significant weight. Unfortunately, these kinds of momentous decisions are hard to study in the lab, but some field studies have identified an association between increased religiosity in times of threat and insecurity (Sosis, 2007; Sosis & Handwerker, 2011).

Important questions remain about the role of supernatural *rewards* versus *punishments*, and *secular* versus *supernatural* punishments. Moreover, we need to know the *conditions*—the sources of variation—under which each may become more or less effective. But I remain impressed by the power of supernatural punishment to have found its way into the beliefs of all cultures in so many varied ways, and to have found such an elegant solution to the problem of collective action. Punishment is important for society,

but no one wants to do it or pay for it. Here, the gods work hard at what we find hard to make work.

Shariff ends by arguing that, if supernatural punishment beliefs are more “cultural byproducts” rather than “hard-wired adaptations”, and secular punishment can be just as effective as supernatural punishment, then there is little reason to expect religious beliefs to persist into the future, and little reason to expect religion to help solve the needs and problems of society. Here, I would take a different view.

First, as I have argued, I think the assumptions underlying this prediction are false: SP are not purely cultural byproducts, they have engrained cognitive components which are here to stay; and secular punishment is not always more present and effective than supernatural punishment.

Second, from a behavioral ecological perspective, our environment (today’s world) is clearly changing. But it is changing in different ways in different regions. In western democracies, largely untroubled by poverty, violence, and war, supernatural punishment may be less necessary or important because an effective government and judiciary promotes and enforces cooperation. But in other parts of the world, riven by poverty, violence, and war, other mechanisms of cooperation may be desperately needed. Where solid institutions of governance are absent, religion offers a powerful means of promoting cooperation and cohesion in the face of severe threats. Maybe we don’t “need” religion any more in the West, but in many parts of the world it is all people have got. And in the future, it may return to be a vital source of help for us all.

Out of the Armchair and into the Field: Morals, Evidence, and Westerners

Montserrat Soler is in general agreement about the potential importance and power of supernatural punishment, but identifies problems with the relevant empirical data that leaves a wide scope for further research. She worries in particular that: (1) SP is not always about morals; (2) cross-cultural evidence for moralistic SP remains limited; and (3) non-western cultures are more diverse than we realize.

Doing Good and Doing Well: Is the Supernatural Punishment Hypothesis About *Moral Behavior*?

Succinctly, Soler articulates two phenomena: “One, supernatural entities punish people. Two, they punish *because* of moral failings” (my italics). Her point is that we must not conflate the two, and feels that the book does (“The premise of Johnson’s book is that the threat of *punishment* from vigilant deities fosters *prosocial* behavior”, my italics). I wanted to emphasize mainly that supernatural punishment tends to occur for some *reason*, as a consequence of certain behaviors and not others. Often, the offending behaviors are self-interested, anti-social, or non-cooperative. Still, Soler presses the point that “it does not follow that the ‘systematic reasons’ are always moral”. I agree.

I tackle this problem head on in the book, acknowledging the many examples of capricious gods and other non-moralistic supernatural agents (including the Greek and Roman gods, the Book of Job, and the trickster gods of indigenous small-scale societies). To my mind, these curiosities are interesting and important, but they do not present a challenge to the theory, for two reasons. First, they are the exceptions rather than the rule (both *across* societies and *within* most of them), and therefore stand out as striking precisely because they break an otherwise strong pattern. Typically, SP punishes violations of social norms (whatever those norms may be). Second, even when these instances may seem, at face value, to suggest SP is unrelated to what we might consider “moral” behavior, I think they are in fact related after all. What they do is serve to bind the group together, promoting social cooperation towards common goals and adherence to prevailing norms. That may be done via moralistic behavior or non-moralistic behavior. I see the supernatural punishment theory as being about promoting cooperation, not promoting altruism or generosity.

To give a stark example, let’s say Nazis believed in supernatural punishment for not performing their duties. A belief in SP would: (1) deter them from carrying out self-interested behaviors or violating Nazi norms; and (2) encourage them to help and cooperate with other Nazis. This would be an example of supernatural punishment deterring self-interest and promoting cooperation. But it hardly says anything about morals. The supernatural punishment hypothesis—strictly speaking—is independent of morality. This is partly because it is about cooperation (getting people to do X rather than

Y), not moralistic altruism (doing good instead of bad). But the other reason, perhaps more importantly, is because it has to be a culturally relativistic theory (as an evolutionary psychologist, I have reservations about unrestrained cultural relativism, and recognize many cross-cultural universals, but some balance here is needed). We can all think of examples of religious beliefs and practices that violate our own moral standards (from Aztec sacrifices to burning heretics at the stake). But no number of such examples can detract from the argument that SP beliefs may have made those individuals and societies more effective in terms of Darwinian fitness. Evolution favors what works, not what is nice.

Having said that, I do, however, argue that under a wide range of conditions, the promotion of cooperation does correlate with “moral” or “prosocial” behavior. I think it makes perfect sense that it should do so, because cooperation is often about suppressing self-interest and helping others for the greater good. But the point is that it need not always do so—precisely as we observe empirically. I agree entirely with that point.

I think part of the reason for this common criticism comes from the literature’s adoption of the word “*prosocial*” (Gintis, 2003; Norenzayan & Shariff, 2008). This implies moralistic behavior, but really we’re trying to explain cooperation (at least, I was). Were the Nazi’s prosocial? To each other, yes. *Cooperation* is the key phenomenon at issue in explaining the origins and growth of human societies in the face of severe collective action problems, and “*prosocial*” muddies the waters, imposing moral standards on our understanding of adaptive behavior.

At the end of the day, all of this is about beliefs in supernatural *consequences* for one’s actions. And rather than arbitrary, they are typically linked to things that help the group as well as the self. Things that help the group tend to align with (but do not always match) moralistic individual and group behavior (certainly, it is not moralistic *between* groups). The best example may be the book of Leviticus in the Bible. God puts great importance on following apparently arbitrary ritual practices, and severe punishments for violating them. At face value, this challenges any utilitarian explanation. But the point is that, by these norms, the people are unified around a set of common practices that signal and identify themselves as co-believers—and distinct from others. Hence, Soler’s own examples of the Brazilian resguardos and the Mexican Santa Muerte Cult in fact make

perfect sense to me. They may not promote *moral* behavior, but they certainly promote what the SPH is really about: suppression of self-interest, collective action, and cooperation within the group.

The Relevant Data: Cross-Cultural Evidence for (Moralistic) Supernatural Punishment

Given the above, I am not too concerned about Soler's second point—a lack of cross-cultural evidence for moralistic supernatural punishment. As I found in researching the book, it is easy to find examples of supernatural punishment for violations of social norms (which may or may not be “moral”). It is harder to find examples of supernatural punishment for “immoral” behavior (not least because how do we say what is “moral” or “immoral” for other societies?). We do, however, have increasing evidence for this in recent cross-cultural studies which find generosity to others (probably a universal moral notion) is greater among those who believe in supernatural agents that are more punishing (Purzycki et al., 2016).

Soler suggests that in insecure environments of widespread poverty, crime, or misfortune, there would be a diminished utility of SP, and “transactional” features of religion instead become more important—a way of bargaining for resources that are in demand. However, I think this remains open for debate. For example, as noted earlier, Sosis (2007; 2011) found insecurity *increases* superstition and religiosity, and even the cult of Santa Muerte in narcocrime-ridden Mexico includes punishments for violating religious obligations.

So in short, I would argue that finding evidence for SP of moral infractions is less important than finding evidence for SP for lapses in cooperation. But even in the former the evidence is striking and growing, including from other non-western cultures. Exceptions are clearly found, but defy a broader pattern that we are all too familiar with in both modern, past, north, and south religions: gods punish those who violate the norms of the community, and oftentimes, that promotes “prosocial” behavior rather than bad behavior.

Seeing the Forest for the Trees: Diversity and Design in Non-Western Cultures

Soler's third point is that non-western cultures are themselves very heterogeneous, and we must pay attention to that variation as well. I agree! We must be careful not to generalize. However, existing surveys remain important. Ake Hultkrantz (1967), for example, was a good source because he maintained a broad vision across two continents (North and South America), not *too* distracted by the details of any one group or people. He thus was able to draw out patterns amongst the kaleidoscope of Amerindian religions. But I agree we must also explore the massive variation within regions. I tried in fact to highlight this in the book, and emphasized it again above in my response to Black (a behavioral ecology approach, explaining variation in religious beliefs and behaviors across environments, is what the SPH predicts, not fears).

This remains work to be done. The book's job was primarily to show that SP is important and widespread. The next step is to explore how well the SPH explains variation in supernatural punishment beliefs as adaptations to local cooperation problems. All such variation is welcome, as it offers more opportunities to test whether SP co-varies as expected with the difficulty or demand for cooperation across time and space.

Finally, I agree wholeheartedly with Soler's lament about the problem of armchair ethnographies and the need for detailed fieldwork. Of course, there are some advantages of earlier accounts in that, while less systematic or methodologically rigorous, they can be better (or vital complementary) data because they describe societies from a time when there was least interference from western religions and ideas. There is a trade off here between flawed descriptions of uncontacted people, and the rigorous study of conflated ones. But certainly Soler's call for more fieldwork is a laudable one that will help us all to hold our theories closer to the fire of data.

Supernatural Punishment: Past, Present, and Future

The supernatural punishment hypothesis was born of an interest in the evolution of cooperation. While cooperation is everywhere in nature and among human societies, it is plagued by severe obstacles that privilege self-interest and free-riding. Punishment offered one possible solution, raising the costs of free-riding high enough to deter or damage free-riders. *Theoretically*, in both evolutionary theory and game theory,

punishment offered an effective solution. *Empirically*, around the globe and across much of recorded history, all human cultures appeared to have recognized the necessity of institutions of laws and sanctions for those who flout them. Without these mechanisms of punishment, large-scale human civilizations would probably not have been possible. However, the power and pervasiveness of secular punishment does not resolve the evolutionary puzzle: how did humans solve the collective action problems over the long period of time *before* the advent of secular institutions that could effectively impose and enforce punishment? One solution is the “institution” of religion, something apparently almost as old as our species, in which *supernatural* agents are believed to punish free-riders. Holding beliefs in imaginary supernatural agents that deter self-interested behavior seems counter-intuitive in Darwinian terms, unless, that is, self-interest itself brings real-world costs. In *God is Watching You*, I argued that as humans became cognitively and linguistically sophisticated, self-interest did indeed bring real world costs that impulsive human behavior found hard to avoid. Transgressions brought retaliation and reputational damage (via harm, sanctions, or isolation), which could be carried out at low cost to the community and high cost to the transgressor. In such circumstances, a belief that the gods are watching would be adaptive, promoting Darwinian fitness at the expense of those who were less prudent in their social behavior.

This Supernatural Punishment Hypothesis has been around for quite a while now, at least since its first incarnation with Oliver Krueger in 2004 (Johnson & Kruger, 2004). Over this time, evidence has grown for the breadth of beliefs in supernatural punishment across cultures and over history, as has evidence that supernatural punishment beliefs do deter self-interest and promote cooperation, in the lab, the field, and in cross-cultural analyses (most notably, Purzycki et al., 2016; Watts et al., 2015). This symposium has highlighted a variety of important debates that remain, which can be roughly summarized as follows (see Table 3): the role of variation in SP (Black), the role of SP in different cooperation games (Cronk and Aktipis), the robustness of the SPH in the face of skeptics (Hartberg), SP’s performance in competition with classic strategies in game theory (Lane), the role of the carrot versus the stick in human cooperation (Lenfesty and Fikes), why we need “false” beliefs for adaptive behavior (McKay et al.), whether SP is an

adaptation or byproduct and arose by genetic or cultural selection (Shariff), and SP's relationship to moral behavior and its variation non-western cultures (Soler).

All of these are important questions to address. Indeed, many of them apply to other theories of the evolution of cooperation, and other theories of the evolution of religion as well. With these commentaries on the book, I am simply glad that people are engaging with argument, and eager to understand why beliefs in supernatural punishment have been so cross-culturally recurrent across the globe and across human history. Both their universality, and their variation over and above this baseline, suggests to me an important *function* that may help to explain the evolution of religion, the cultural diversity of religious beliefs and practice, and ultimately the success of human societies. Whatever the diminishing role of religion today, and the rise of secular institutions in its stead, the supernatural remains important for understanding our past, and punishment remains important for the future—from whatever source it must come. Since human beings are not angels, societies adopted gods that have had to be cruel to be kind.

References

- Agrawal, A.A. (2001). Phenotypic plasticity in the interactions and evolution of species. *Science*, 294, 321-326.
- Aktipis, C.A. (2011). Is cooperation viable in mobile organisms? Simple Walk Away strategy favors the evolution of cooperation in groups. *Evolution and Human Behavior*, 32, 263-276.
- Axelrod, R. & Hamilton, W.D. (1981). The evolution of cooperation. *Science*, 211, 1390-1396.
- Barrett, J. (2012). *Born Believers: The Science of Children's Religious Beliefs*: Atria Books.
- Barrett, J.L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, 4, 29-34.
- Baumeister, R.F., Bratslavsky, E., Finkenauer, C. & Vohs, K.D. (2001). Bad is stronger than good. *Review of General Psychology*, 5, 323-370.
- Bering, J.M. & Johnson, D.D.P. (2005). 'Oh Lord, you hear my thoughts from afar': Recursiveness in the cognitive evolution of supernatural agency. *Journal of Cognition and Culture*, 5, 118-142.
- Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

- Clutton-Brock, T.H. & Parker, G.A. (1995). Punishment in animal societies. *Nature*, 373, 209-216.
- Cosmides, L. & Tooby, J. (1994). Better than rational: evolutionary psychology and the invisible hand. *American Economic Review*, 84, 327-332.
- Davies, N.B., Krebs, J.R. & West, S.A. (2012). *An Introduction to Behavioural Ecology*. Chichester: Wiley Blackwell.
- Dickson, D.B., Olsen, J., Dahm, F. & Wachtel, M. (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, 53-79.
- Foster, K.R. & Kokko, H. (2009). The evolution of superstitious and superstitious-like behaviour. *Proceedings of the Royal Society B: Biological Sciences*, 276, 31-37.
- Gintis, H. (2003). Solving the Puzzle of Prosociality. *Rationality and Society*, 15, 155-187.
- Hartberg, Y., Cox, M. & Villamayor-Tomas. (2016). Supernatural monitoring and sanctioning in community-based resource management. *Religion, Brain & Behavior*, 6, 95-111.
- Haselton, M.G. & Buss, D.M. (2009). Error management theory and the evolution of misbeliefs. *Behavioral and Brain Sciences*, 32, 522-523.
- Haselton, M.G. & Nettle, D. (2006). The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review*, 10, 47-66.
- Homerin, T.E. (2006). Altruism in Islam. In J. Neusner & B. Chilton (Ed), *Altruism in World Religions* (67-87). Washington, DC: Georgetown University Press.
- Hultkrantz, A. (1967). *The Religions of the American Indians*. Berkeley: University of California 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. (2012). What are atheists for? Hypotheses on the functions of non-belief in the evolution of religion. *Religion, Brain & Behavior*, 2, 48-70.
- Johnson, D.D.P. (2016). *God is Watching You: How the Fear of God Makes Us Human*. New York: Oxford University Press.
- Johnson, D.D.P., Blumstein, D.T., Fowler, J.H. & Haselton, M.G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, 28, 474-481.
- Johnson, D.D.P. & Kruger, O. (2004). The good of wrath: Supernatural punishment and the evolution of cooperation. *Political Theology*, 5.2, 159-176.
- Johnson, D.D.P., Stopka, P. & Knights, S. (2003). The puzzle of human cooperation. *Nature*, 421, 911-912.

- Kahneman, D. & Tversky, A. (1982). Subjective probability: a judgment of representativeness. In D. Kahneman, D. *et al.* (Eds), *Judgment under Uncertainty: Heuristics and Biases* (32–47). Cambridge: Cambridge University Press.
- McCauley, R.N. (2011). *Why Religion is Natural and Science is Not*. Oxford: Oxford University Press.
- McKay, R.T., Efferson, C., Whitehouse, H. & Fehr, E. (2010). Wrath of God: religious primes and punishment. *Proceedings of the Royal Society B: Biological Sciences*, 278, 1858-1863.
- Milinski, M., Semmann, D. & Krambeck, H. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, 415, 424-426.
- Miller, J.H. & Page, S.E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, NJ: Princeton University Press.
- Morewedge, C.K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology*, 138, 535-545.
- Nettle, D. (2004). Adaptive illusions: Optimism, control and human rationality. In D. Evans & P. Cruse (Ed), *Emotion, Evolution and Rationality* (193-208). Oxford: Oxford University Press.
- Norenzayan, A. & Shariff, A.F. (2008). The origin and evolution of religious prosociality. *Science*, 322, 58-62.
- Nowak, M.A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393, 573-577.
- O'Gorman, R., Henrich, J. & Van Vugt, M. (2009). Constraining free-riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 276, 323-329.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
- Panchanathan, K. & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432, 499-502.
- Peoples, H.C. & Marlowe, F.W. (2012). Subsistence and the evolution of religion. *Human Nature*, 23, 253-269.
- Purzycki, B., Finkel, D., Shaver, J., Wales, N., Cohen, A.B. & Sosis, R. (2012). What Does God Know? Supernatural Agents' Perceived Access to Socially Strategic and Nonstrategic Information. *Cognitive Science*, 36, 846-869.
- Purzycki, B.G., Apicella, C., Atkinson, Q.D., Cohen, E., McNamara, R.A., Willard, A.K., Xygalatas, D., Norenzayan, A. & Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. 530, 327-330.
- Rozin, P. & Royzman, E.B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, 5, 296-320.
- Rubenstein, D.I. & Wrangham, R.W. (1986). *Ecological Aspects of Social Evolution*. Princeton: Princeton University Press.

- Shariff, A.F. & Norenzayan, A. (2007). God is watching you: Supernatural agent concepts increase prosocial behavior in an anonymous economic game. *Psychological Science*, 18, 803-809.
- Shariff, A.F. & Norenzayan, A. (2011). Mean Gods Make Good People: Different Views of God Predict Cheating Behavior. *International Journal for the Psychology of Religion*, 21, 85-96.
- 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 (Ed), *Evolution, Culture and the Human Mind* (119-136). New York: Psychology Press.
- Shariff, A.F. & Rhemtulla, M. (2012). Divergent Effects of Beliefs in Heaven and Hell on National Crime Rates. *PLoS ONE*, 7, e39048.
- Silton, N.R., Flannelly, K.J., Galek, K. & Ellison, C.G. (2013). Beliefs about God and mental health among American adults. *Journal of Religion and Health*, *In Press*.
- 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.
- Sosis, R. (2007). Psalms for safety: Magico-religious responses to threats of terror. *Current Anthropology*, 48, 903-911.
- Sosis, R. (2009). The adaptationist-byproduct debate on the evolution of religion: five misunderstandings of the adaptationist program. *Journal of Cognition and Culture*, 9, 315-332.
- Sosis, R. & Alcorta, C. (2003). Signaling, solidarity, and the sacred: the evolution of religious behavior. *Evolutionary Anthropology*, 12, 264-274.
- Sosis, R. & Handwerker, W.P. (2011). Psalms and Coping with Uncertainty: Israeli Women's Responses to the 2006 Lebanon War. *American Anthropologist*, 113, 40-55.
- Sosis, R., Kress, H. & Boster, J. (2007). Scars for war: evaluating alternative signaling explanations for cross-cultural variance in ritual costs. *Evolution and Human Behavior*, 28, 234-247.
- Standen, V. & Foley, R. (1989). *Comparative Socioecology: The Behavioural Ecology of Humans and Animals*. Oxford: Blackwell Scientific Publications.
- Trivers, R.L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46, 35-57.
- Watts, J., Greenhill, S.J., Atkinson, Q.D., Currie, T.E., Bulbulia, J. & Gray, R.D. (2015). Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proceedings of the Royal Society B: Biological Sciences*, 282, 1-7.
- Whitehouse, H. & Lanman, J.A. (2014). The ties that bind us: ritual, fusion, and identification. *Current Anthropology*, 55.

- Whitmarsh, T. (2016). *Battling the Gods: Atheism in the Ancient World*: Faber & Faber.
- Wilson, D.S. (2002). *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago: University of Chicago Press.
- Zuckerman, P. (2008). *Society without God: What the Least Religious Nations Can Tell Us About Contentment*. New York: NYU Press.

Table 3. The commentaries on *God is Watching You*, their key points and my responses (SP = supernatural punishment; SPH = SP Hypothesis).

| Authors | Field | Issue | Key Points | Key Responses |
|------------------|--------------|---------------------------------------|--|--|
| Black | Anthropology | Afterlife belief variation | <ul style="list-style-type: none"> • If SP beliefs are a “genetic” <i>universal</i>, how do we explain the great cultural <i>variation</i> in SP? | <ul style="list-style-type: none"> • SP beliefs are not a genetic universal (only the underlying cognitive traits enabling them are) |
| | | | <ul style="list-style-type: none"> • <i>Universal</i> aspects and <i>variable</i> aspects of SP beliefs may have <i>different</i> explanations | <ul style="list-style-type: none"> • Agree; interesting point worthy of further exploration |
| | | | <ul style="list-style-type: none"> • Cultural evolution may better explain the latter | <ul style="list-style-type: none"> • Agree; only cultural selection can explain specific manifestations of SP beliefs (e.g. Hell or Zeus) |
| Cronk & Aktipis | Anthropology | Risk pooling | <ul style="list-style-type: none"> • Helping can occur via: <ul style="list-style-type: none"> #1 Reciprocity (which generates debt and credit) #2 “Needs-based transfers” (no debt or credit) • However, #1 has obvious and instant secular sanctions (debt and credit); #2 does not (no debt or credit). Hence, cheating is easier in the latter, so needs reinforcement. | <ul style="list-style-type: none"> • Great insight (and raises the important question of <i>which</i> game individuals are playing, and whether SP may affect or resolve such games differently) |
| | | | <ul style="list-style-type: none"> • SP may stabilize #2 | <ul style="list-style-type: none"> • I agree. How generalizable is it? #2 evident among the Maasai; is it true of the other 8 societies in their study? • A prediction: Among the Maasai, is SP greater in cooperation problems involving unpredictable payoffs? |
| Hartberg | Anthropology | Robustness of cooperative systems | <ul style="list-style-type: none"> • Is the SPH robust? | <ul style="list-style-type: none"> • Important question. |
| | | | <ul style="list-style-type: none"> • Doesn’t the SPH need cooperative norms <i>already</i> in place to work (and thus, isn’t it circular)? | <ul style="list-style-type: none"> • No: Cooperative groups, and punishment of transgressions, predate humans; SP came afterwards. |
| | | | <ul style="list-style-type: none"> • Skeptics are common and may undermine the SPH | <ul style="list-style-type: none"> • Skeptics are not that common and may <i>help</i> the SPH |
| Lane | Anthropology | Computer modeling | <ul style="list-style-type: none"> • How does SP belief fare in Axelrod and Hamilton’s (1981) classic model of the evolution of cooperation? | <ul style="list-style-type: none"> • Excellent idea (although we should consider which game best models cooperation in small-scale human societies) |
| | | | <ul style="list-style-type: none"> • The SP <i>trait</i> is successful and spreads widely, across a wide range of model conditions | <ul style="list-style-type: none"> • Remarkable finding; SP’s disposition to cooperate may take the place of Axelrod and Hamilton’s kin selection |
| | | | <ul style="list-style-type: none"> • Presence of the SP trait leads to mixed <i>strategies</i> of Tit-for-Tat, Mostly Cooperate, and Always Defect | <ul style="list-style-type: none"> • TFT may remain in the population because it provides the “secular punishment” that SP helps to avoid • Need data on whether <i>individual</i> X plays strategy Y (data presented as population summaries only) |
| Lenfesty & Fikes | Psychology | Fear is only one side of human nature | <ul style="list-style-type: none"> • Punishment problematic: leads to chronic anxiety | <ul style="list-style-type: none"> • Chronic anxiety my <i>help</i> rather than hurt the SPH • Natural selection doesn’t care about psychological comfort • Not necessary to scare everyone (only cheats and free riders) |

| | | | | |
|---------------------|------------------------|--|--|--|
| | | | | need fear punishment) |
| | | | • SP beliefs lead to fear | • Fear leads to SP beliefs! |
| | | | • Social engagement mechanisms key to prosociality | • Too rosy a view of human nature? • Depends on the game being played. In some contexts, rewards may be enough; in others, punishment essential • Complex adaptive systems approach: its not one or the other, you need both (punishment and positive social engagement mechanisms). |
| McKay <i>et al.</i> | Psychology / Economics | Benefits of false beliefs | • Beliefs in supernatural agents not the <i>only</i> way to avoid mistakes | • Agree (a rational actor can err on the side of caution); however only <i>exaggerated</i> perceptions of SP can help |
| | | | • Beliefs in supernatural agents not the <i>best</i> way to avoid mistakes | • Disagree; exaggerated perceptions of SP help to make fewer behavioral <i>mistakes</i> (even if many errors of perception) |
| | | | • Humans are Bayesian decision-makers, with perfect information and accurate computation | • Humans are not Bayesian decision-makers; perceptions and decisions suffer significant uncertainty, poor information, and cognitive constraints; thus, perceptual biases adaptive |
| Shariff | Psychology | Sources and effectiveness of SP | • SP is more cultural than genetic | • Broadly agree (genetic in underlying dispositions, cultural in how exactly it is manifested) |
| | | | • SP is less powerful than secular punishment | • Highly context dependent, but the jury is still out on this |
| | | | • If the above are true, religion will die | • Behavioral ecology suggests a different future: religion, and SP, will endure in <i>places</i> and <i>contexts</i> in which it remains effective at promoting cooperation (e.g. poor governance, high insecurity, and hard or vital cooperation problems) |
| Soler | Anthropology | Morality, field data, and western bias | • SP is not always about morals | • Agree (the SPH is about deterring self-interested behavior and promoting cooperation problems, <i>moral or not</i> ; however, cooperation and morality are often correlated) |
| | | | • Cross-cultural evidence for <i>moralistic</i> SP remains limited | • Agree, although given the above I am not surprised |
| | | | • Non-western cultures more diverse than we think | • Absolutely, and this variation is an important test for SPH |