Evolutionary Psychology

human-nature.com/ep – 2006. 4: 219-233

Original Article

Hand of God, Mind of Man: Punishment and Cognition in the Evolution of Cooperation*

Dominic Johnson (corresponding author), Society of Fellows, Princeton University, Joseph Henry House Princeton NJ 08544, Tel: 609-258-4835, Fax: 609-258-2783, Email: dominic@princeton.edu

Jesse Bering, Institute of Cognition & Culture, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland UK, Email: j.bering@Queens-Belfast.ac.uk

Abstract: The evolution of human cooperation remains a puzzle because cooperation persists even in conditions that rule out mainstream explanations. We present a novel solution that links two recent theories. First, Johnson & Kruger (2004) suggested that ancestral cooperation was promoted because norm violations were deterred by the threat of supernatural punishment. However, this only works if individuals attribute negative life events (or a prospective afterlife) as intentionally caused by supernatural agents. A complementary cognitive mechanism is therefore required. Recently, Bering and Shackelford (2004) suggested precisely this. The evolution of "theory of mind" and, specifically, the "intentionality system" (a cognitive system devoted to making inferences about the epistemic contents and intentions of other minds), strongly favoured: (1) the selection of human psychological traits for monitoring and controlling the flow of social information within groups; and (2) attributions of life events to supernatural agency. We argue that natural selection favoured such attributions because, in a cognitively sophisticated social environment, a fear of supernatural punishment steered individuals away from costly social transgressions resulting from unrestrained, evolutionarily ancestral, selfish interest (acts which would rapidly become known to others, and thereby incur an increased probability and severity of punishment by group members). As long as the net costs of selfish actions from real-world punishment by group members exceeded the net costs of lost opportunities from self-imposed norm abiding, then god-fearing individuals would outcompete non-believers.

Key words: Cooperation, selfishness, religion, punishment, strong reciprocity, theory of mind, intentionality system, language, cognition.

* A version of this paper is forthcoming in "*The 'Nature' of Belief: Scientific and Philosophical Perspectives on the Evolution of Religion*", edited by Jeffrey Schloss and Alvin Plantinga.

Introduction

> We're in hell … they never make mistakes and people are not damned for nothing.
>
> Jean-Paul Sartre's Inès, in *Huis Clos*

The puzzle of human cooperation

Cooperation is widespread among mammals, birds, insects, cells, microscopic organisms, and different organs of the body (Gadagkar 2001; Wilson 2000). Sometimes cooperation results in mutual payoffs to all actors involved, and can therefore be easily understood as each pursuing their own selfish interest. However, other instances of cooperation are more surprising, because individuals help others despite incurring a cost in doing so. In the last half century, a number of theories have come to understand such behaviour as the result of motives that, while they may be *apparently* altruistic at first glance, ultimately serve selfish genetic interests (they incur an immediate cost, but result in a net gain to inclusive fitness overall, Dawkins 1986). The four dominant theories are: "Kin-selection," in which cooperation is genetically rewarded by favouring kin (Hamilton 1964); "reciprocal altruism," in which altruistic acts are returned later on (Trivers 1971); "indirect reciprocity," in which one's reputation for cooperation is rewarded indirectly through the favour of third-party observers (Alexander 1987; Nowak and Sigmund 1998); and "costly signalling," in which generosity serves as an advertisement of high fitness to would be mates or allies (Gintis et al. 2001; Zahavi 1995). Formerly puzzling examples of animal cooperation have now been routinely explained in terms of these theories (for a review, see Dugatkin 1997).

By contrast, cooperation *among humans* is still not understood. Although people do increase cooperation when kin-selection, reciprocal altruism, indirect reciprocity, and costly signalling are at stake, we also continue to cooperate when they are not (Fehr and Fischbacher 2003; Gintis 2003). In the words of two leading scholars, "people frequently cooperate with genetically unrelated strangers, often in large groups, with people they will never meet again, and when reputation gains are small or absent," leaving human cooperation as an "evolutionary puzzle" (Fehr and Gächter 2002, p.137). The key evidence for such puzzling behaviour comes from controlled laboratory studies demonstrating that people cooperate even when any possible self-interested payoffs via existing theories are carefully eliminated one by one. The result is that, when asked to play simple games that represent every-day social dilemmas, people from both modern and pre-industrial societies around the globe cooperate to a greater extent than can be accounted for by traditional theory – a phenomenon dubbed "strong reciprocity" (Fehr and Fischbacher 2003; Henrich et al. 2004). So far, no one has come up with a consensus explanation for this phenomenon. A number of scholars have invoked group selection as a possible explanation (Boyd et al. 2003; Gintis 2000). Another explanation may be that our psychology simply fails to optimise behaviour in evolutionarily novel circumstances (such as laboratory experiments or big cities), and better reflects the constraints of the environment in which we evolved, where we lived in small groups of extended kin, few strangers, strong hierarchies and lasting reputations (Barkow et al. 1992; Burnham and Johnson 2005; Johnson et al. 2003). In this paper, we take an entirely new approach. We suggest that religious beliefs, specifically the moralizing and sanctioning behaviour they generate, may serve as a common origin for human cooperation.

Religion as a solution to the puzzle

It would be incredible to suggest that religion has nothing to do with cooperation – either in ancient or modern societies. Anthropologists have long noted such links, and over the years have both championed and criticized functionalist accounts of religion's apparently numerous socially beneficial features (Morris 1987; Pals 2006; Weber 1922/1978). However, scientific progress on the topic reached a "theoretical impasse" until the advent of approaches that explicitly couched the benefits of religion in terms of natural selection (simply observing possible benefits ignored the problem of how the prerequisite costly beliefs initiated, and why cheats did not thrive, Sosis and Alcorta 2003). The new evolutionary approach has given rise to a number of theories arguing that religion was a key promoter of within-group cooperation during human evolution (e.g. Cronk 1994; Irons 2001; Roes and Raymond 2003; Sosis 2003; Wilson 2002), but this work remains totally absent from the literature on "strong reciprocity" and the puzzle of cooperation (Johnson et al. 2003; Schloss 2004).

In fact, proponents of strong reciprocity have specifically denied any link between cooperation and religion (e.g. Fehr and Gachter 2003), despite mounting empirical evidence supporting such an intuitive link. For example, Richard Sosis has shown that, among a large sample of 19th century communes, religious groups with more costly rituals out-survived secular groups and religious groups with fewer rituals (Sosis and Bressler 2003). Among Israeli kibbutzim, groups with more religious rituals also demonstrated higher levels of cooperation than secular groups and religious groups with fewer rituals (Sosis and Ruffle 2003), which may explain why religious kibbutzim are economically successful while secular ones have faced bankruptcy (Fishman and Goldschmidt 1990). There is also evidence that religion tends to promote cooperation in a broad range of historical and pre-industrial societies (Johnson 2005; Wilson 2002). That religious beliefs are associated with higher levels of within-group cooperation is not in doubt. What remains intriguing is *why*.

A New Theory

We outline a precise, proximate cognitive mechanism that suggests it is the expectation and fear of supernatural punishment that serves to promote cooperation. We also argue that this mechanism evolved via individual selection (any group selection effects, though they are not necessary, would help drive the system). The theory builds on two recent and complementary ideas: (1) supernatural punishment as a positive impact on cooperation (Johnson and Kruger 2004); and (2) human cognition as an evolutionarily novel canvas for the workings of natural selection (Bering and Shackelford 2004).

Supernatural punishment and cooperation

It is increasingly accepted that punishment is key to ensuring cooperation (Andreoni et al. 2003; Clutton-Brock and Parker 1995; Fehr and Gächter 2002; Sigmund et al. 2001; Trivers 1971). However, the act of punishing cheats entails costs, so punishment itself represents a "second-order" public good (Hackathorn 1989; Yamagishi 1986). The original puzzle of cooperation therefore just re-appears at a new level: "second-order" cheats may cooperate towards the public good, but then defect from contributing to punishment. So how is cooperation enforced? Four solutions to this conundrum have emerged in the literature. Three are deemed unsatisfactory (Henrich and Boyd 2001, p. 80), and the fourth is contested: (1) punishment is administered by an external institution (however, while this may be true in western societies today, cooperation evolved long before modern institutions existed, and is evident even in remote societies that are not subject to state regulations); (2) punishment is

not costly after all (however, administering punishment must incur some cost, however small, of time and/or effort which, combined with the risk of reprisals from punished individuals or their allies, simply returns us to the original dilemma); (3) both regular defectors and those who refuse to punish are punished (however, as Henrich and Boyd put it: "Do people really punish people who fail to punish other non-punishers, and do people punish people who fail to punish people, who fail to punish non-punishers of defectors and so on, *ad infinitum*?"); (4) Some fraction of people altruistically punish defectors for the good of the group (Fehr and Gächter 2002; Fehr and Fischbacher 2003), and this trait is propagated by group selection (however, this requires that humans are genuinely altruistic, a claim that is problematic for a number of reasons, see Burnham and Johnson 2005; Johnson et al. 2003). The puzzle therefore remains: Without institutions of law and order, and without a good incentive for people to punish each other, how could early human societies establish cooperation with a credible deterrent threat against cheats?

We believe solution 1 is discounted too readily. Although most legal and law enforcement institutions are indeed modern inventions, Henrich and Boyd (2001) neglect another "external" category of norm setting and enforcement that reaches as far back as we can see into human history – religion.

Johnson and Kruger (2004) argued that, over our evolutionary history, individuals would be dissuaded from free-riding if they feared supernatural retribution as a consequence of their actions. Religious codes, taboos and mythology provided the "laws" – the rights and wrongs which defined the norms of conduct promoting, among other things, cooperation. These norms were enforced by the threat of supernatural punishment, either in the present and/or in the afterlife (commonly endorsed by folklore, explanations for other people's misfortune, and supernaturally sanctioned worldly punishment by real group members). If supernatural punishment is held as *a belief*, then this threat becomes a deterrent *in reality*, so the mechanism can work regardless of whether the threat is genuine or not (following Thomas' dictum: "If men define situations as real, they are real in their consequences" (Thomas and Thomas 1928, p. 572)).

Cooperation enforced by the threat of supernatural punishment has four major selective advantages that evade the classic public goods problems troubling current theoretical work: First, there is no second-order free rider problem (supernatural agents are envisioned as administering the punishing). Second, since other group members do not have to be vigilantes, they do not risk reprisals that could undermine future cooperation. Third, (believing) defectors can expect to be automatically caught (the idea is encapsulated in Matt. 5:28: "whosoever looketh on a woman to lust after her hath committed adultery with her already in his heart"). Fourth, (believing) defectors can expect to be automatically punished (the act itself triggers the punishment).

Considerable ethnographic evidence suggests that the threat of supernatural punishment for norm transgressions exerts a powerful effect on people's behaviour – believers literally alter their everyday decisions in order to avoid supernatural retribution (see examples in Bering and Johnson 2005; Boyer 2001). Not only is supernatural punishment commonly feared in diverse cultures around the world, both ancient and modern, it is also commonly linked to taboos concerning life or death collective action problems, such as scarce resources, sexual access, food sharing, hunting, divisions of labour, defence, or warfare (see Boyer 2001; Earhart 1993; Weber 1922/1978).

Supernatural punishment may come from any mix of gods, dead ancestors, witches or sorcerers. One or more feature prominently in hunter-gatherer societies, and all are commonly attributed to the cause of ill fortune (Boyer 2001, see p. 160; Murdock 1980). Dead ancestors are commonly offered gifts and attention specifically to avoid their retribution (Bonsu and Belk 2003). In Medieval Europe concerns for the dead were so prevalent in the

conduct of daily life that one historian treated them as a separate age group (Bering Forthcoming). In ancient Hawaii, the "souls" of the dead (*akua*), once unconstrained from bodily limitations and senses, could be in several places at one time, know the thoughts of others, and were in constant interaction with the living (Dudley 2003). There are some cultures that are apparently not particularly concerned about supernatural punishment, such as the Amazonian Yanomamo whose spirit of judgment after death can be lied to about one's wordly conduct because he is stupid (Chagnon 1997). Nevertheless, such cases appear to be exceptions to an otherwise widespread significance of supernatural punishment to cultures across the globe and across history.

The significance of supernatural punishment is common to modern religions as well. Christians who act contrary to God's will expect divine retribution either immediately by sanctions (e.g. struck down with an affliction or some other misfortune), or later, in hell. Even if they don't believe that, they commonly attribute positive and negative life events to their conduct before God. Either way, "it is plain from the bible that sin will be punished" (Harrison et al. 1960, p. 196). Supernatural punishment is also a central theme in Islam, where salvation depends on "human effort as well as God's mercy in following the Qur'an's teachings" (Coward 2003, p. 164-165). Similar concerns for the afterlife are prominent in East Asian and Indian religious traditions, as well as in ethnographic evidence on the world's far more numerous and diverse pre-industrial and historic cultures (for some examples and evidence, see Bering and Johnson 2005; Bering Forthcoming; Boyer 2001; Johnson and Kruger 2004; Johnson 2005; Wilson 2002).

Why punishment is more important than reward

It may seem odd to focus on punishment, because most religions also offer the prospect of rewards for good behaviour (in fact many people, religious or not, see positive events as felicitous signs of supernatural forces – e.g., "it was *meant* to be" (Bering 2002; Gilbert et al. 2000)). Such beliefs would, like punishment, serve to induce cooperative behaviour if one was rewarded for pro-social actions.

However, the effects of carrots and sticks on the level of cooperation are not symmetrical, even when of equivalent magnitude: punishment is inherently *more* effective at promoting cooperation than rewards. Carrots are not enough because, although they may encourage *some* people to cooperate, they do not prevent *all of them* from cheating. Even if the rewards of cooperation are large and obvious to everyone involved, they provide no credible deterrent against defectors – cheats will not be deterred if they can gain even more by shirking the costs of cooperation (Schelling 1960; Sigmund et al. 2001). This reflects the fundamental paradox behind the famous "Prisoner's Dilemma" game. Even though each player is aware of the substantial rewards if they both cooperate, rational actors defect because this is the only way to avoid exploitation and it may bring an even greater payoff – and there is no credible deterrent against doing so (Axelrod 1984; Poundstone 1992). In other contexts too, rewards turn out to be less effective than equivalent levels of punishment in promoting cooperation. Empirical experiments bear out this claim: despite its potential mutual rewards, cooperation collapses in real-life groups if there are no additional binding agreements to prosecute or punish dissenters (a single cheat can cause otherwise cooperative agents to withdraw their own contributions, Fehr and Gächter 2002; Ostrom et al. 1992; Yamagishi 1986). Such results have led to a convergence of opinion among experimental economists, game theorists and evolutionary biologists that – wherever self interest conflicts with group outcomes – cooperation will emerge only if defectors are punished.

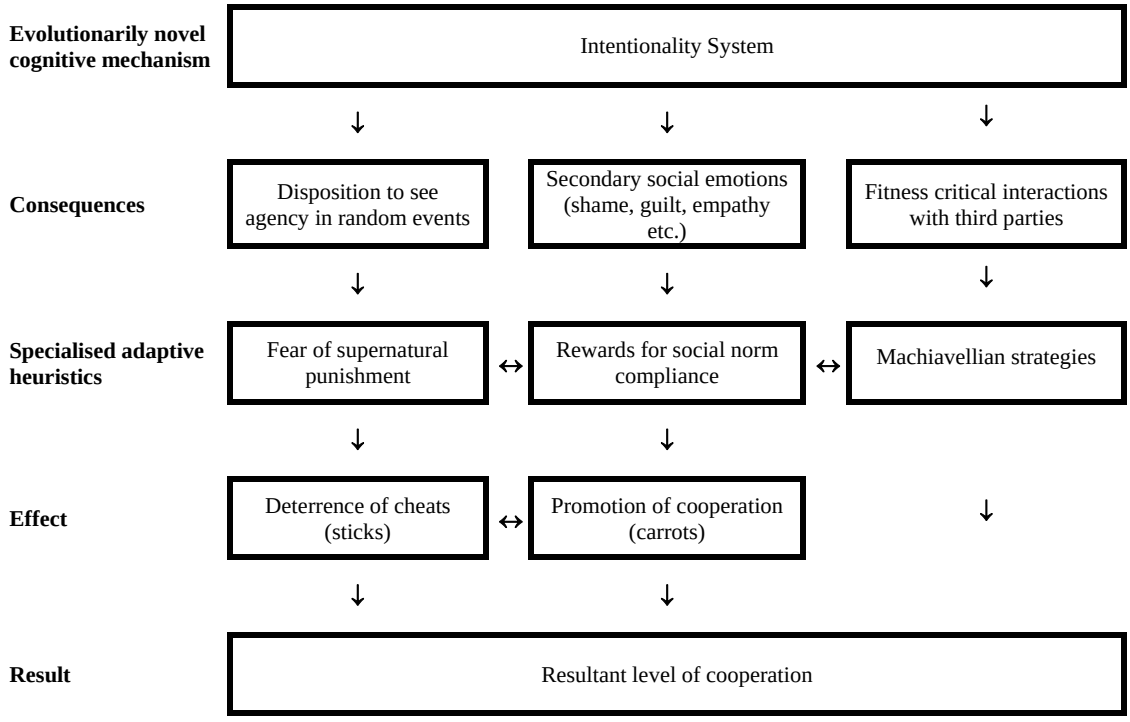
Rewards may contribute to promoting cooperation, but it is the weaker of the two complementary forces: punishment has an intrinsic leverage. While rewards clearly play an

important part in religious behaviour (a Christian, for example, may be motivated by eternity in heaven as much as by the fear of hell), the punishment aspect is likely to have the more potent influence on the dynamics of cooperation. As one theologian pointed out: "The very proclamation of hell indicates that the defenders of religion found it necessary to balance the attraction of its promise with a threat for the 'others', who rejected it or failed to meet its tests" (Bernstein 1993, p. x). This resonates with the observation that while there are many pre-industrial societies in which the only supernatural agents are antagonistic, there are few, if any, whose only supernatural agents are beneficent. The effectiveness of sticks over carrots also concords with accumulating evidence that negative psychological events and phenomena are much more potent in their effects than positive ones (Baumeister et al. 2001).

Human cognition and supernatural agency

The supernatural punishment theory, outlined above, offers the plausible hypothesis that a fear of supernatural punishment is the proximate mechanism that maintains cooperation, but it begs the all-important question of how the system initiates in the first place. Johnson and Kruger (2004) suggested that the mechanism could originate via the "green beard effect" (Dawkins 1986; Hamilton 1964), via a purely cultural innovation, or via group selection processes (Sober and Wilson 1998; Wilson and Sober 1994). None of these mechanisms may be necessary, or sufficient, however.

Figure 1. The human intentionality system has three key consequences: disposition to see agency in random events; secondary social emotions (shame, guilt, empathy etc.); and fitness critical interactions with third parties. These lead to a fear of supernatural punishment (which deters potential defectors), rewards for social norm compliance (which promotes cooperative tendencies), and Machiavellian strategies (which exploit the intentionality system). In combination, these effects determine the resultant level of cooperation.



As has been recently pointed out, "supernatural punishment can only be an effective deterrent insofar as individuals are capable of reasoning that negative life events are caused by supernatural agents who have explicit *reasons* for bringing about such events" (Bering 2004, p. 434). Recent work by Bering (2002; Forthcoming) offers precise theory and evidence suggesting that humans do indeed reason in this way about negative events. We appear to have an inherent cognitive tendency to search for reason and intentionality in life events, and to attribute positive and negative outcomes to supernatural agency. Keleman suggests that children are "intuitive theists" because of their commonplace teleological reasoning that things usually exist "for" something (e.g. clouds are for raining, see Keleman 2004). Bering and Bjorklund's (2004) study on children's reasoning about the psychological states of dead agents also hints at a default "afterlife" stance that may only be usurped by explicit scientific understanding about biology and death – knowledge that was of course limited in our pre-scientific evolutionary past. Such tendencies, we argue, may have specific selective advantages at the individual level. The logic is set out below and illustrated in Figure 1.

Novel selective pressures on human sociality

Unlike other primate species, humans possess a sophisticated "theory of mind" and, in particular, an "intentionality system," which is the capacity to represent mental states as the unseen causes of behaviour (Bering 2002; Povinelli and Bering 2002). This intentionality system is foundational for a uniquely human cognitive specialization: "second- and third-order representation" – the ability to know what others know, and to know that they know what we know, or did (i.e. A knows that B knows what A knows, or did). Humans also differ from other species in having complex language (allowing information about specific social behaviours to spread among the group). Consequently, B can inform C by word of mouth about A's actions, information that can profoundly influence the nature of subsequent interactions between A and C, with significant fitness consequences (for example, if A stole from B in the absence of any social others, then retaliation against A might come from C,D, or E and so on, perhaps days, weeks, or months later). Through the lens of this evolutionary novelty, many higher-function and premeditated human behaviours take on great adaptive significance, including the murder of witnesses, revenge, suicide, and generosity (Bering and Shackelford 2004). With humans, therefore, natural selection has a new workbench to shape behavioural adaptations that Darwin did not consider. No other species are subject to its effects.

Before the evolution of the intentionality system and complex language, selfish behaviour would be consistently selected for as long as it conferred a net payoff (even when this occurred in full view of others). For example, chimpanzees can be selfish in front of other chimpanzees without their behaviour being reported to absent others. There can therefore be no negative repercussions from absent third parties because such individuals could not entertain others' knowledge states (nor could they learn such complex information by communication).

After the evolution of the intentionality system and complex language, by contrast, it was in the genes' interests to *avoid* selfish behaviour in contexts that could bring negative repercussions (now, one had to worry about the consequences of other actors, wholly removed from the scene of the crime, learning of the act and responding later). People could hear, discover, infer, remember, report, hypothesize, plan and act on others' behaviour – even long after the event. What are the consequences?

God-fearing strategies

Selfish behaviour is evolutionarily ancient, whereas the intentionality system and complex language are evolutionarily novel. So while selfish behaviours might have paid off in the simpler social life of our prehistoric ancestors, many of them (or too many of them) would bring a net fitness loss in a cognitively sophisticated, whispering society. The advent of these novel cognitive abilities increased the likelihood of public exposure for selfish behaviour, which could bring high costs of retaliation by other group members (involving social sanctions, seizure of property, physical harm, ostracism, imprisonment, punishment of kin, or death).

Specific mechanisms might have evolved to rescue inclusive fitness *after* the individual committed a social offence in this new "big-brother" society (e.g. cognitive processes underlying confession, blackmail, killing witnesses, suicide and so forth (Bering and Shackelford 2004)). However, these de facto strategies tax reproductive success, so natural selection would favour more efficient traits that constrain selfishness to some extent in the first place (indeed we see such traits in human interaction every day – restraint, self-control, sacrifice, sharing, patience etc.). Those that carried on being indiscriminately selfish would be out-competed by prudent others who were able to successfully inhibit their more ancient selfish motives and refrain from breaching social rules to begin with.

According to Bering and Shackelford (2004), the human intentionality system allowed the selection of traits that militated against public exposure. Because the temptation to cheat remained, however, we add that something extra – a belief in *supernatural* punishment – was an effective way to caution oneself against transgressions and thereby avoid "real" worldly retribution by other group members. God-fearing people may, therefore, have had a selective advantage over non-believers because the latter's more indiscriminately selfish behaviour carried a higher risk of real-world vengeance by the community.

Machiavellian strategies

So far we have focused on the disadvantages of the novel intentionality system and complex language – selfish actions now bring an increased risk of detection and retaliation. However, these cognitive innovations also brought opportunities: selective pressures for traits that *exploit* them. One can manipulate others' knowledge as well as suffer from it (as a result of these two mechanisms, the overall selective effect might be expected to be quite strong, effectively "pushed" and "pulled" simultaneously in the same direction by evolution – exposed transgressors are selected out, prudent exploiters of the social cognitive system are selected in). As an example of manipulation, one can conceal the transgressions of kin, or preferentially cooperate with those who have established a good reputation with others – examples which hint at significant implications for the evolution of kin-selection and direct or indirect reciprocal altruism among humans (Johnson and Bering In prep). In short, these new psychological forces gave humans, for better or worse, a new capital stock to trade in – social information. Our ancestors became highly invested in this stock because it exerted a significant influence on reproductive gain. Profits came from effectively gathering, retaining, and regulating (through whatever means possible, including deception, threats, and violence) the flow of social information that had the potential to impact inclusive fitness. One may therefore postulate Machiavellian strategies that did exploit the human intentionality system for personal gain, but which were not god-fearing.

Table 1. Three strategies come into competition with the advent of the human intentionality system (IS) and complex language. Grey-shading indicates consequences that act *against* genetic fitness. Machiavellians outcompete ancestral individuals, and god-fearing strategists outcompete Machiavellians as long as $pc > m$. See text for further details.

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

Which strategy wins?

Table 1 compares the performance of the above two strategies (God-fearing and Machiavellian), and the ancestral state, following the advent of the intentionality system and complex language. Machiavellians would clearly outcompete ancestral individuals because, while everything else is identical between them, ancestrals cannot exploit these new cognitive features for personal gain. More importantly however, Table 1 indicates that god-fearing strategists can outcompete Machiavellians. They differ in just two respects: god-fearing strategists have a lower probability of detection, but miss out on some opportunities for selfish rewards. Therefore, god-fearing strategists will outcompete Machiavellians *as long as* the total expected costs of punishment (i.e. the probability of detection (*p*) multiplied by the cost of punishment (*c*)) is greater than the cost of missed opportunities for selfish rewards (*m*). In other words, when the inequality $pc > m$ is true. This would occur wherever the rewards of selfishness were relatively small compared with the costs of public exposure (which may include social sanctions, seizure of property, physical harm, ostracism, imprisonment, punishment of kin, or death). Even a small *p* can mean selfishness does not pay on the average. Moreover, *Error Management Theory* predicts that, where $pc > m$, we should expect *exaggerated* estimates of *p* (such as a belief that supernatural agents are watching) to outperform *accurate* estimates of *p*, given that the latter will engender more mistakes (Haselton and Buss 2000; Haselton and Nettle 2006; Nettle 2004). Interestingly, recent criminal evidence indicates that offenders tend to *underestimate* the probability of being caught and the costs of punishment (Robinson and Darley 2004).

Summary of the model

Humans often act on selfish motives (and sometimes inadvertently due to emotionally charged situations) – acts which, thanks to the human intentionality system and complex language, carry a far greater chance of social exposure than in previous stages of evolution. If the costs of exposure are high enough, individuals that were more likely to refrain from

cheating for fear of supernatural agents concerned with group norms (indeed, such agents are often the proposed authors of these norms), and who are believed to punish defectors by inflicting misfortune (on both the self and innocent others), could have out-reproduced otherwise equal – and more indiscriminately selfish – individuals. Of course, Machiavellian, non-believing cheats who do not get caught would do best of all, but we suggest that the heightened costs of exposure by virtue of human cognitive sophistication favoured the evolution of traits that suppress selfish behaviour, and favoured instead the kind of moralistic behaviour that is, after all, empirically common among human societies (Alexander 1987; Trivers 1971).

Conclusions

The supernatural punishment theory of Johnson and Kruger (2004), combined with the powerful implications of the human intentionality system and complex language (Bering and Shackelford 2004), offers a novel theory for the origins of human cooperation – a solution that has a specific proximate mechanism, and that precisely defines the cognitive processes involved. Our proposition is not mutually exclusive of other theories of religion, nor of other theories of cooperation. The mechanism we describe would complement many of them. However, our proposal offers a more complete and plausible mechanism than some, and an intuitive and circumstantially supported one. Although we have highlighted a central role for individual selection in our theory, which we believe could drive the system on its own, any inter-group advantages leading to the group selection of such morally bound cooperative behaviour would augment the process (as per Sober and Wilson 1998; Wilson and Sober 1994; Wilson 2002). Indeed, group selection would lead to a much more rapid dominance of god-fearing strategies, since groups with Machiavellians will suffer by comparison.

An additional lever in our proposed mechanism comes from a consideration of third parties. Over and above any personal experience linking one's own actions to one's fortunes, people can draw lessons from supernatural agency apparently befalling others (again, a faculty made possible by the intentionality system). Someone else's misfortune or fortune (e.g. illness, gifted children) may tend to be seen as evidence of wrongdoing or virtue (e.g. selfishness, generosity). Whether the victim really *is* bad or virtuous is of little consequence for selective pressures to operate if the events themselves are perceived as the "evidence" (especially where other group members corroborate that interpretation; cultural learning is clearly important here). Such perceived connections will steer onlookers away from behaviour that would bring the same fate – not just because of the fear of supernatural punishment (as we proposed in our general argument above), but also because of learning how such negative life events would be viewed and treated by other group members. Thus, supernatural agents are seen not only as communicating to the self through life events, but in so doing, they are also seen as communicating to other group members about the moral (in)aptitude of the self. The gods effectively call out the wicked, exposing them to the group to impose its own social punishments.

How does our theory fit with existing literature? Sosis and Bressler (2003, p. 227) found that, on the basis of their comparisons of secular and religious communes, the costly signalling theory of religion fails to "capture some critical elements of religious belief that distinguish it from belief in a secular ideology." In their study, variation in costly signalling explained variation in *religious* commune survival. However, variation in costly signalling did *not* explain variation in *secular* commune survival. The underlying reason for this, they suggest, is the special "sanctity" of religious rituals, which simply cannot be matched by secular rituals (see also Whitehouse 2000). Religious rituals are superior to secular ones in

their ability to build solidarity among group members, it appears, *because* they are directed towards a supernatural being, which authenticates them beyond logical analysis – a critical component of their success (Rappaport 1999). Sosis and Bressler predict as a consequence that, among different religious doctrines, those that are more reliant on supernatural agency should exhibit higher levels of cooperation (a prediction partially supported by a recent empirical analysis of 186 pre-industrial cultures, Johnson 2005). Other evolutionary studies arrive at similar appeals to some as-yet-unexplained, special feature of religion: Dogon women in Mali, for example, are obliged to visit "menstrual huts" to advertise their fertility cycle and thereby reduce cuckoldry. Although this conforms to theories based on ritual, the study's author noted that "the threat of supernatural sanctions is crucial for enforcement" (Strassmann 1992). We offer an explanation for *why* such a supernatural component may be so fundamental to understanding the power of religion in achieving cooperation.

Clearly, cooperation in the modern world cannot be explained solely by any religious theory because cooperation is prevalent among atheists as well as believers (although we must remember that 79% of Americans expect a day of judgment when God decides whether they will go to heaven or hell and, depending on religious affiliation, 74% or more believe in an afterlife, as do 58% of adults *who have no religious affiliation* (Pinker 2002; Religioustolerance.org)). Many instances of social cooperation today are no puzzle at all, because governments and other organizations impose strong social contracts to cooperate (and punishment if one does not). However, many of these modern institutions, and their founding morals, ethics and norms, are in fact deeply rooted in local traditions that are essentially religious. Indeed, religious traditions continue to underlie fundamental aspects of law, political discourse, appeals to public action problems, and social life, even if the modern proponents are no longer themselves believers (consider marriage, swearing on the bible in court, charity, many national constitutions, and calls for U.S. unity against an "evil empire" or "axis of evil" – it is not inconceivable that these norms originated and persevered because of their selective success and cognitive salience over human history). Certainly, most people today – even atheists – continue to behave in accordance with a set of values which, although they may appear as self-evident, are directly analogous to many religious codes (and evoke the same secondary emotions of shame, empathy, guilt etc., that supervise one's own actions). Our proposed mechanism can be generalized to suggest similar adaptive advantages in superstition, folklore, or just world beliefs.

Speculations about modern society aside, the real puzzle is still the *evolutionary origins* of cooperation behaviour – independent of the forces governing cooperation today. How did early human societies achieve cooperation? Future studies of the evolutionary origins of cooperation must focus on analogues of that point in our history, the best window onto which comes from evidence on contemporary hunter-gatherer societies. Further cross-cultural and within-culture empirical tests would be tremendously useful (Johnson 2005; Wilson 2002). We also need to know more about how religious beliefs and behaviours are transmitted across generations. There is exciting new work along these lines (Alcorta and Sosis 2005), and recent studies of twins indicate that aspects of religiosity are heritable, and that this influences adult behaviour over and above influences in environmental conditions while growing up (Koenig et al. 2005). Certainly, there is something deep in biology and human nature that predisposes us to religious beliefs, offering a wealth of opportunities for future research (Atran and Norenzayan 2004; Barrett 2004; Bering 2002; Bering and Bjorklund 2004; Bering and Johnson 2005; Boyer 2001; Keleman 2004; Wilson 2002).

Much of the literature on religion and cooperation focuses on squaring religious behaviour with economic "rational-actor" assumptions or, at the other extreme, the physiological responses of the brain. What is lacking, however, is a careful consideration of the "black box" in between – the human mind itself, and how cognitive processes interact

with the natural selection of behaviour. We suggest that, by virtue of our unique social cognitive abilities, the evolution of cooperation may have been influenced more than currently appreciated by the hand of God at work in the mind of man.

Acknowledgements: We thank Candace Alcorta, Justin Barrett, Jeffery Boswall, Terry Burnham, Oliver Curry, Gordon Gallup, Stewart Guthrie, Brian Hare, Roger Johnson, David Kydd, Gabriella de la Rosa, Ian Pitchford, Alvin Plantinga, Todd Shackelford, Jeffrey Schloss, Richard Sosis, Dan Sperber, David Voas, Harvey Whitehouse, David Sloan Wilson, and Richard Wrangham for comments and criticisms on the ideas in this paper. We also sincerely thank Jeffery Schloss and Alvin Plantinga for the invitation to present this work at their conference on "The 'Nature' of Belief," at Calvin College, Grand Rapids MI, in November 2005.

Received 15 September 2005; Accepted 21 March, 2006.

References

Alcorta, C. S. and Sosis, R. (2005). Ritual, Emotion, and Sacred Symbols: The Evolution of Religion as an Adaptive Complex. *Human Nature*, 16: 323-359.

Alexander, R. D. (1987). *The Biology of Moral Systems*. Aldine, N.Y.: Hawthorne.

Andreoni, J., Harbaugh, W. and Vesterlund, L. (2003). The Carrot or the Stick: Rewards, Punishments, and Cooperation. *American Economic Review*, 93: 893-902.

Atran, S. and Norenzayan, A. (2004). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioural and Brain Sciences*, 27: 713-730.

Axelrod, R. (1984). *The Evolution of Cooperation*. London: Penguin.

Barkow, J. H., Cosmides, L. and Tooby, J. (Eds.). (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Barrett, J. L. (2004). *Why Would Anyone Believe in God?* Altamira Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. and Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5: 323-370.

Bering, J. M. (2002). The existential theory of mind. *Review of General Psychology*, 6: 3-24.

Bering, J. M. (2004). The evolutionary history of an illusion: religious causal beliefs in children and adults. In Ellis, B. and Bjorklund, D. (Eds.), *Origins of the social mind: Evolutionary psychology and child development* (pp. 411-437). New York: Guilford Press.

Bering, J. M. and Bjorklund, D. F. (2004). The natural emergence of reasoning about the afterlife as a developmental regularity. *Developmental Psychology*, 40: 217-233.

Bering, J. M. and Shackelford, T. (2004). The causal role of consciousness: A conceptual addendum to human evolutionary psychology. *Review of General Psychology*, 8: 227-248.

Bering, J. M. and 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.

Bering, J. M. (Forthcoming). The folk psychology of souls. *Behavioural and Brain Sciences.*

Bernstein, A. E. (1993). *The Formation of Hell: Death and Retribution in the Ancient and Early Christian Worlds*. Ithaca, NY: Cornell University Press.

Bonsu, S. K. and Belk, R. W. (2003). Do not go cheaply into that good night: Death-ritual consumption in Asante, Ghana. *Journal of Consumer Research*, 30: 41-55.

Boyd, R., Gintis, H., Bowles, S. and Richerson, P. J. (2003). The Evolution of Altruistic Punishment. *Proceedings of the National Academy of Sciences*, 100: 3531-3535.

Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Burnham, T. and Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik*, 27: 113-135.

Chagnon, N. A. (1997). *Yanomamo*. Fort Worth: Harcourt Brace.

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

Coward, H. (2003). *Sin and Salvation in the World Religions: A Short Introduction.* Oxford: Oneworld.

Cronk, L. (1994). Evolutionary theories of morality and the manipulative use of signals. *Zygon*, 4: 117–135.

Dawkins, R. (1986). *The Selfish Gene*. Oxford: Oxford University Press.

Dudley, M. K. (2003). *A Hawaiian Nation I: Man, Gods, and Nature*. Kapolei, Hawai'i: Na Kane O Ka Malo Press.

Dugatkin, L. A. (1997). *Cooperation in Animals*. Oxford: Oxford University Press.

Earhart, H. B. (Ed.). (1993). *Religious Traditions of the World*. New York: Harper Collins.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415: 137-140.

Fehr, E. and Fischbacher, U. (2003). The Nature of Human Altruism. *Nature*, 425: 785 - 791.

Fehr, E. and Gachter, S. (2003). Reply to Johnson et al. *Nature*, 421: 912.

Fishman, A. and Goldschmidt, Y. (1990). The Orthodox Kibbutzim and Economic Success. *Journal for the Scientific Study of Religion*, 29: 505-511.

Gadagkar, R. (2001). *Survival Strategies: Cooperation and Conflict in Animal Societies*. Cambridge, Mass: Harvard University Press.

Gilbert, D. T., Brown, R. P., Pinel, E. C. and Wilson, T. D. (2000). The illusion of external agency. *Journal of Personality & Social Psychology*, 79: 690-700.

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology*, 206: 169-179.

Gintis, H., Smith, E. and Bowles, S. (2001). Costly signalling and cooperation. *Journal of Theoretical Biology*, 213: 103-119.

Gintis, H. (2003). Solving the Puzzle of Prosociality. *Rationality and Society*, 15: 155-187.

Hackathorn, D. D. (1989). Collective action and the second-order free-rider problem. *Rational Society*, 1: 78-100.

Hamilton, W. D. (1964). The genetical evolution of social behaviour, I & II. *Journal of Theoretical Biology*, 7: 1-52.

Harrison, E., Bromiley, G. and Henry, C. (Eds.). (1960). *Wycliffe Dictionary of Theology*. Peabody: Hendrickson.

Haselton, M. G. and Buss, D. M. (2000). Error Management Theory: A new perspecitive on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, 78: 81-91.

Haselton, M. G. and Nettle, D. (2006). The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review*, 10: 47-66.

Henrich, J. and Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208: 79-89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E. and Gintis, H. (Eds.). (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.

Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In Nesse, R. (Ed.), *Evolution and the capacity for commitment* (pp. 292-309). New York: Russell Sage Foundation.

Johnson, D., Stopka, P. and Knights, S. (2003). The puzzle of human cooperation. *Nature*, 421: 911-912.

Johnson, D. D. P. and Kruger, O. (2004). The Good of Wrath: Supernatural Punishment and the Evolution of Cooperation. *Political Theology*, 5.2: 159-176.

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. and Bering, J. M. (In prep). A cognitive revolution for the evolution of cooperation.

Keleman, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological Science*, 15: 295-301.

Koenig, L. B., McGue, M., Krueger, R. F. and Bouchard, T. J. (2005). Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *Journal of Personality*, 73: 471-488.

Morris, B. (1987). *Anthropological studies of religion: An introductory text*. Cambridge Cambridgeshire; New York: Cambridge University Press.

Murdock, G. P. (1980). *Theories of Illness: A World Survey*. Pittsburg: HRAF, University of Pittsburgh Press.

Nettle, D. (2004). Adaptive illusions: Optimism, control and human rationality. In Evans, D. and Cruse, P. (Eds.), *Emotion, Evolution and Rationality* (pp. 193-208). Oxford: Oxford University Press.

Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393: 573-577.

Ostrom, E., Walker, J. and Gardner, R. (1992). Covenants with and without a sword: self governance is possible. *American Political Science Review*, 86: 404-417.

Pals, D. L. (2006). *Eight Theories of Religion*. New York: Oxford University Press.

Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature*: Penguin Putnam.

Poundstone, W. (1992). *Prisoner's Dilemma: John von Neumann, Game Theory and the Puzzle of the Bomb*. Oxford: Oxford University Press.

Povinelli, D. J. and Bering, J. M. (2002). The mentality of apes revisited. *Current Directions in Psychological Science*, 11: 115-119.

Rappaport, R. A. (1999). *Ritual and religion in the making of humanity*. Cambridge, U.K.; New York: Cambridge University Press.

Religioustolerance.org. (2005). www.religioustolerance.org/chr_poll3.htm.

Robinson, P. H. and Darley, J. M. (2004). Does criminal law deter? A behavioural science investigation. *Oxford Journal of Legal Studies*, 24: 173-205.

Roes, F. L. and Raymond, M. (2003). Belief in moralizing gods. *Evolution and Human Behaviour*, 24: 126-135.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard: Harvard University Press.

Schloss, J. P. (2004). Evolutionary ethics and Christian morality: surveying the issues. In Clayton, P. and Schloss, J. (Eds.), *Evolution and Ethics: Human Morality in Biological and Religious Perspective* (pp. 1-24). Grand Rapids, Mich.: Eerdmans.

Sigmund, K., Hauert, C. and Nowak, M. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences USA*, 98: 10757-10761.

Sober, E. and Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behaviour*. Cambridge: Harvard University Press.

Sosis, R. (2003). Why aren't we all Hutterites? Costly signaling theory and religious behavior. *Human Nature*, 14: 91-127.

Sosis, R. and Alcorta, C. (2003). Signaling, Solidarity, and the Sacred: The Evolution of Religious Behavior. *Evolutionary Anthropology*, 12: 264–274.

Sosis, R. and Bressler, E. R. (2003). Cooperation and commune longevity: A test of the costly signaling theory of religion. *Cross-Cultural Research*, 37: 211-239.

Sosis, R. and Ruffle, B. (2003). Religious ritual and cooperation: Testing for a relationship on Israeli religious and secular kibbutzim. *Current Anthropology*, 44: 713-722.

Strassmann, B. I. (1992). The function of menstrual taboos among the Dogon: Defense against cuckoldry? *Human Nature*, 3: 89-131.

Thomas, W. I. and Thomas, D. S. (1928). *The Child in America: Behaviour Problems and Programs*. New York: Alfred A. Knopf.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46: 35-57.

Weber, M. (1922/1978). *The Sociology of Religion*. Berkeley: University of California Press.

Whitehouse, H. (2000). *Arguments and Icons: Divergent Modes of Religiosity*. Oxford: Oxford University Press.

Wilson, D. S. and Sober, E. (1994). Reintroducing group selection to the human behavioural sciences. *Behavioral and Brain Sciences*, 17: 585-654.

Wilson, D. S. (2002). *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago: University of Chicago Press.

Wilson, E. O. (2000). *Sociobiology: The New Synthesis*. Harvard: Belknap Press.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51: 110-116.

Zahavi, A. (1995). Altruism as Handicap: the Limitations of Kin Selection and Reciprocity. *Journal of Avian Biology*, 26: 1-3.