Joseph Bulbulia

# Religious Costs as Adaptations that Signal Altruistic Intention


## Abstract

*Most cognitive psychologists explain supernatural understandings as cognitive by-products acquired in specific but fairly common cultural circumstances. This paper uses evolutionary game theory and the biology of animal signalling to promote a contrary view. It explains religious cognition as an exquisite adaptation that enhances individual reproductive success by facilitating reciprocal altruism. The key to understanding the design innovation of religious altruism lies in the specific costs that religious thought and practice impose on the believing organism. These costs play a strategic role in displaying authentic commitment to policies of social exchange, applying critical safeguards to defection from co-operative ventures. The following account explains a suite of otherwise anomalous cognitive features associated with religious thought, such as strong emotional responses to unseen persons and forces; belief in supernatural punishments and reward; illusions about the moral goodness of co-religionists and the vices of heretics; and dispositions to invest in expensive and wasteful ritual displays. The paper offers some testable predictions about the psychological architecture that generates religious thought and suggests some new horizons for psychological exploration.*



## Key words

*Altruism; cognitive psychology; evolutionary psychology; game theory; NASH equilibrium; prisoners dilemma; reciprocal altruism; religion; religious cognition; ritual; social exchange; the handicap principle.*


## Opening

Below I use research in theoretical biology to shed light on the nature and function of supernatural cognition, and suggest some promising new paths of psychological inquiry these new understandings open. By supernatural cognition I mean beliefs[1], emotions, and practices relative to supernatural beings and powers like Yahweh, The Amida Buddha, Shiva, Jesus Christ, Zeus, grace, mana, soul, num, and other supernatural beings and forces to which individuals are committed. To make exposition easier, I have adopted the convention of using "religion" to denote supernatural cognition and the term "gods" to denote supernatural beings and powers.[2]

I argue that a central function of supernatural cognition is to facilitate efficient solutions to otherwise difficult or intractable co-ordination problems. These emerge when individuals acting collectively further their individual interests through sacrificing on each others behalf. Religion is an adaptation that generates a special kind of reciprocal altruism, which may be called religious altruism. An altruistic act is one that benefits the inclusive fitness (RS) of the recipient at some cost to the inclusive fitness (RS) of the benefactor.[3] Religious altruists are motivated to altruism from a belief in supernatural powers capable of rewarding altruism and punishing defection (failure to reciprocate). The paper explains religious beliefs, emotions, and behaviours as products of an adaptive signalling system designed to propagate religious altruism.

## The Costs of Religion

The main problem facing any adaptationist theory is how to account for religions reproductive cost (–RS) (ATRAN 2002).[4] The most immediate expense of religious belief is a distortion of reality. Religious persons live under a sky crowded with supernatural beings and must adjust their activities to these beliefs. Selection generally acts against tendencies to misperceive.[5] At a physiological level, the devotions and pieties of religion, bowing and scrapping before gods deplete metabolic reserves. Religious persons will often assume strenuous bodily postures and enact ritual movements that burn calories and require long training regimes. Religion also incurs material costs. To stage a ritual means gathering precious objects and animals and manipulating them in ways that may lead to their

destruction, another form of waste. Inevitably there are opportunity costs that flow from any religious practice. In venerating gods, the religious delay or forgo hunting, gathering, breeding and rearing, leaving potential fitness enhancement unrealised. Religion is frequently hazardous. The devout risk infection by ritual piercing; malnourishment through starvation; malling when hunting sacrificial carnivores; burning in trials by fire, and other harms.

Given the price of religion, it is interesting to inquire why selection did not recoil from sticker shock at the religious option. Biologists and cognitive psychologists have proposed two general answers, which though promising, leave some critical questions unanswered.

### Standard adaptationism and cost

Traditional adaptationist accounts hold that religious cognition is evolutionarily enforceable because whatever its costs, the reproductive payoffs are higher (WILSON 1998). Selection will favour traits that incur metabolic, material, or opportunity costs, sometimes substantial, if the average inclusive fitness of an individual disposed to this kind of biological spending thereby increases. Most adaptationist accounts look for the advantages of religion in enhancements to social exchange (WILSON 2002).

> Standard Adaptivism: $RS cost of religion < +$RS value of enhanced social exchange

The approach is helpful because it places religious cognition within the natural history of our species, a perspective that in other domains has shed much light on the nature and function of cognition (PLOTKIN 1998; BARKOW/COSMIDES/TOOBY 1992). The analysis however leaves obscure why selection did not endorse cheaper versions of co-operation. The reproductive costs of religious practice must be subtracted from reproductive benefits of social exchange. It is unclear how sacrificing to imaginary sky beings instead of promoting ourselves and others who will promote us increases average RS. Given the added expense of god commitment, an optimal system would eliminate costly religious inclination.

**Problem: +$RS value of religious social exchange < + RS value of non-religious social exchange.** Given that non-religious altruism is reproductively cheaper than religious altruism, it is unclear what evolutionary force holds god-belief in place.[6]

### Spandrel theory and cost

It may be that there is no overall +RS advantage to religious cognition. Perhaps religion is not itself a design artefact but rather a cognitive by-product of other adaptive systems, a spandrel (GOULD/LEWONTIN 1979). If so, the costs of religion do not need to be explained. Religion is merely noise made by functional cognitive machinery (BARRETT/KEIL 1996; BARRETT/NYHOF 2001; BOYER 1992, 1994; BOYER/WALKER 2000).

> Spandrel Theory: $RS cost of religion < +$RS value of cognitive systems that accidentally generate it

Spandrel theories are desirable because they minimise assumptions about the complexity of the cognitive system. It is better to avoid postulating an intricate functional design when a simpler explanation, appealing to a rag heap of cognitive mechanisms, adequately explains religious thought (BOYER 2001).

Moreover, the connection between religious understandings and social exchange remains explicable. The social exchange system, itself exquisitely adaptive, may simply incorporate religious information as cultural input, weaving it into the relevant exchange outputs. Suppose, plausibly, that altruists act in virtue of group identities (HARDIN 1995). It may be that social exchange system takes religious understanding and affiliation as the relevant marker. In other cases, sex, race, age, family, trade, sporting affiliation or something else may serve as the relevant marker. That exchange partners sometimes use religion as a shibboleth does not imply a dedicated faculty that systematically distorts the world as god ridden (BOYER 2001).

Finally, spandrel theory accords well with the observation that religion is not universal in the way that vision or language is (SPERBER 1990, 1996). Some normal people claim not to believe in gods while presenting no cognitive deficiency. They are not like language speakers without verbs.

Most cognitive psychologists argue that religious understandings are acquired and transmitted in virtue their intrinsic recall properties and/or features of ritual settings in which they are learned.[7] Even committed evolutionary psychologists agree there are no *dedicated* faculties designed to produce religious thought (ATRAN 2002; PINKER 1997, pp554–558).[8] Rather, religion emerges through an interaction of specific cultural material and the

evolved mind. Remove those specific materials and watch religion disappear.

In spite of its explanatory virtues, it may be that spandrel theory undervalues the costs of religious commitment. If religion is cognitive noise, its sounds seem to be reproductively deafening. It is critical not to lose sight of the broad spectrum of human investment that lies behind a god-centred reality, one extreme of which is occupied by convulsing shamans, celibate priests, and suicide bombers. It seems selection should have placed mufflers over the relevant cognitive systems that produce such understandings and practices as by-products. On the contrary, selection seems only to have amplified religious distortions with powerful emotional responses and motivations.

> Spandrel Theorys Problem: $RS of religion is sufficiently high for selection to have acted against it.

Spandrel theory only works if religion is reproductively inexpensive, yet religious believers seem to pay a high reproductive premium to inhabit castles in the air.

In what follows I will not attempt to prove that religious cognition is costly. Rather the paper explains religious cognition as a signalling system that generates cost as an adaptive feature. If there are such costs, as there seems to be, the present model can explain them.

#### Solution: religious cost as adaptive

In order to understand how religion works it is important to look beyond the religious believer to her audience. Religion facilitates altruism. For altruism to evolve, altruistic individuals need to find each other. To do this they must reliably signal their willingness to reciprocate to others. Religion differentiates genuine altruists from impostors by imposing specific costs on altruists that only they are willing to pay. Cost is an adaptive feature of this signaling system*: if supernatural cognition were not expensive it would not have evolved*.

The following account explores how religion reliably encodes altruistic commitment. First I explain how the intentional structure of religious belief motivates altruistic exchange. I show how:

> Religious belief + (other beliefs and desires) → reliably motivates → commitment to altruistic exchange

Next I briefly examine how certain emotional expressions of religious commitment reliably mark the presence of specific motivations for altruistic sacrifice. I show how:

> Religious emotion → reliably signals → [Religious belief +… → commitment to altruistic exchange]

Thirdly I explain how the specific costs of ostentatious religious practice, especially rituals, certify religious commitment, and hence altruistic exchange. I show how:

> Participation in religious ritual → reliably signals → [Religious Belief +…→ commitment to altruistic exchange]

If religion is a cost inducing signalling mechanism, it may be possible to reverse engineer important but concealed features of its cognitive design. I close with a discussion of key distortion mechanisms that generate commitment to supernatural beings and other illusions; signalling and detection mechanisms that enabling individuals to distinguish co-religionists from defectors; and altruistic mechanisms to reward and punish persons believed to hold similar supernatural commitments.

## Religious Belief and Reciprocal Altruism

#### Convention

In an important paper on the evolution of reciprocal altruism, AXELROD and HAMILTON point out that the benefits sought by living things are disproportionally available to co-operating groups (AXELROD/HAMILTON 1981, p1391) a statement that seems a truism, but whose truth, when analysed, seems far from guaranteed (AXELROD 1997).

Where interests exactly converge it is easy to see how co-operation can evolve (LEWIS 1969; SKYRMS 1996). HUME considers the example of two rowers positioned in the same boat (HUME 1739). Left to a random stroke pattern, the boat will jerk forward inefficiently. Assuming speed and fluidity of motion as desiderata (the rowers are not playing some game of disruption or assaulting each other) each rower shares the identical interest: to synchronising their strokes. Assuming zero establishment costs and an arbitrary benefit for convergence of one utile, the rowers are bound by the following payoff matrix:

| | Rower 1 | Rower 1 |
|---|---|---|
| Rower 2 | Unsynchronised 0<br>0 | 0<br>0 |
| Rower 2 | 0 | Synchronised +1<br>+1 |

**Table 1:** Rowing Problem

Each rower therefore has an interest in adopting the same standard.

Similarly, consider two individuals who want to meet for a beer. There are three options: *Tupelo, Motel, and the Q-bar.* Neither cares where to meet. Their payoff matrix is:

| | Trundle 1 | Trundle 2 | Trundle 3 |
|---|---|---|---|
| Ed 1 | Tupelo 1<br>1 | 0<br>0 | 0<br>0 |
| Ed 2 | 0 | Motel 1<br>1 | 0<br>0 |
| Ed 3 | 0 | 0 | The Q-bar 1<br>1 |

**Table 2:** Meeting Problem

Trundle and Ed gain by going to the same place. To improve the odds, they need to signal their intentions. They have no interest in lying, so clear and accurate signals work best: When Ed tips his hand to his mouth, then meet at Tupelo.

Assume the benefits of convergence in these cases enhance reproductive success, that the scenarios are repeated often in life, and that they arise consistently generation after generation. Selection will tend to advance any psychological disposition that improves synchronisation. Over evolutionary time, selection will favour cognitive systems that foster conventional signalling in pure co-operative ventures.[9]

It is easy to see how the benefits sought by living things are disproportionally available to social creatures able to co-ordinate on matters of common benefit with zero loss. But in nature such interactions are rare because benefits are distributed unevenly (MAYNARD SMITH 1982). Often, an individuals best move depends critically on how others with conflicting interests will act. Where interests overlap imperfectly action is strategic. Frequently it is in an organisms best interest to effectively misrepresent how it will act. For example, in species where parental investment falls heavily on females, males have an incentive to present themselves as willing investment partners to as many females as will copulate with them, even if a strategy of promiscuity prevents such global parental investment (TRIVERS 1972).

Economists have developed methods for describing optimal strategies in uncertain conditions. Because selection endorses strategies optimised to maximise RS, it is possible to use economic theory to generate hypotheses about an organisms cognitive design (MAYNARD SMITH 1982). Understanding how organisms bypass strategies of deception in uncertain conditions sheds much explanatory light on how religion works.

### Strategy and social exchange: Classical game theory and the concept of a Nash equilibrium

Call a *game* an interaction between two or more persons called *players* whose outcome depends on the interactive strategies of each player. Allow that each players motives may coincide, conflict, or fall somewhere between these extremes. A *NASH equilibrium* is a strategy or set of strategies in a game in which no player can benefit by changing his or her strategy while the other players keep their strategies unchanged. In a two-player game, a NASH equilibrium occurs where each players strategy is optimal, given the other players best strategy (NASH 1951; SCHELLING 1960).

Consider the prisoners dilemma: Ed and Trundle are captured by the authorities for a crime they jointly committed. If neither complies with the authorities (that is if both players cooperate with each other), both will go to jail for three years. If one defects by turning in the accomplice, who co-operates by remaining quite, the snitch will be set free, while the accomplice will be sentenced to twenty years in jail. However if both comply with the authorities each will spend ten years in jail. What should the prisoners do?

| | Trunde defect | Trundle co-operate |
|---|---|---|
| Ed defect | –10<br>–10 | –20<br>0 |
| Ed co-operate | 0<br>–20 | –3<br>–3 |

**Table 3:** The Prisoners Dilemma

Defection is the single NASH equilibrium for this game. Both Ed and Trundle could do no better by remaining silent, given the best strategy of the other is to remain silent. In fact, defection is an example of a dominant NASH equilibrium: it is the best strat-

egy no matter what the other may do, that is, even if the other co-operates.

Notice that co-operation, the irrational option, is *strictly efficient*: it is the strategy whose interaction with itself brings the greatest payoff. Yet neither prisoner has an incentive to follow the strictly efficient strategy. Defection dominates strictly efficient moves, hence the dilemma. A large class of social interactions involve conflicts of interest of this kind (FRANK 1988; SCHELLING 1960).[10]

### Evolutionary game theory and strictly efficient strategies

Natural selection is a theory of reproduction and differential survival of individuals living in breeding populations. Selection favours alleles (gene sequences) that generate reproductively successful phenotypes, organisms with high average (+RS). The relevant traits may be physiological or psychological, with no sharp distinction between the two. Assuming strategic innovations can be inherited, the expectation is for individuals in species with high neural capacity to demonstrate elaborate sophistication in strategic thinking, particularly in multiparty interaction.[11] Even organisms as simple as *Rhizobium* bacteria seem to exhibit strategic responses to their environment (AXELROD/HAMILTON 1981, p1392). Where players are not closely related, selection would seem to promote strategies of defection in interactions resembling the Prisoners Dilemma. Defection is an evolutionarily stable strategy, because it cannot be invaded by mutants adopting a different strategy.[12]

Where players can match themselves to other players, the case is different. Given high correlation between co-operators, strictly efficient strategies, once they appear in a finite population, will move to fixation. Brian SKYRMS has used computer modelling to show that in cases of perfect correlation, evolution will carry co-operation to fixation. What evolves is a DARWINIAN version of KANT's categorical imperative: *Act so that if others act likewise, fitness is maximised* (SKYRMS 1996, p62). If co-operators can secure encounters with like-minded co-operators, then the costs imposed by co-operative behaviour are more than repaid by its benefits. This is true even if co-operative strategies are strictly dominated by other strategies.

The trick in this instance is for co-operators to reliably signal their strategy to others and to avoid exchanging with defectors who have an interest in mimicking the signal. Because defectors stand to reap even higher rewards by imitating co-operative signalling, defectors should be willing to pay the price for any arbitrary (conventional) signal a co-operator would pay. Where co-relation is unreliable, defection can invade. Hence, for cooperation, signalling needs to be secure.

Reliable signalling is essential to the model of religion promoted here and I explore the evolution and functional nature of this capacity more formally below. Before turning to signalling behaviour, however, it is important to examine how brute force can ensure obedience to co-operative strategies, an aspect of social exchange that will prove critical to understanding religious cognition.

### Enforced co-operation

Clearly external systems of reward and punishment, if widely advertised, may enhance cooperation. Call such systems external enforcement systems. If the relevant costs of such systems do not exceed the average benefits to each individual supporting it, selection will ratify psychological dispositions that favour establishing a police force. Enforcement works by altering the punishment structure for games that invite defection, converting strictly efficient strategies into NASH equilibriums, thus eliminating defection as a rational option.

Imagine that the jailed Ed and Trundle are members of a Mafia family, the Agaronis. Though turning states evidence on an accomplice brings freedom, the advantage is short lived. The Agaroni family promptly hunts down all defectors and outfits them with swimsuits and matching cement shoes. Perfectly credible threats of punishment alter the actual structure of the original game where defection was the NASH equilibrium.

| | Trunde defect | Trundle co-operate |
|---|---|---|
| Ed defect | –10<br>–10 | –20<br>– infinity |
| Ed co-operate | – infinity<br>–20 | –3<br>–3 |

**Table 4:** External Enforcement Game: Agaroni Family

Co-operation in the Agaroni punishment game is the single NASH equilibrium. The new payoff schedule, when advertised, enforces cooperation.

Similarly, credible promises of reward for co-operative play adjust the pay-off schedules of individual players. If each prisoner possessed the assurance that co-operation will bring specific new

value, say a million dollars upon release, then strictly efficient co-operation becomes the single NASH equilibrium in this game. Co-operation in the Reward Game is rational because the game itself has been altered to favour co-operation.[13] Variations of the Reward Game are variations of external enforcement. Whether punishments or rewards, extra inducements alter rational play by altering the game itself.

### The costs of external enforcement

Any system providing the relevant incentives, however, will necessarily impose further costs on the organisms that produce, manage, and enforce such a system. First, we must pay our police. The levies from which enforcers are paid must be enforced, as each individual has an incentive to avoid taxation. Second, it is easy for corruption to enter into a system of exchange that relies on enforcement at precisely those points where players should want to forestall itnamely where the gains from defection are potentially massive. In such cases defectors need only bribe the police more than their salaries. Dynamic interaction with enforcement agents sets up additional co-ordination problems. Once deputised, the police have an incentive to turn power to their advantage. Once corrupted, there will be no one to guard us from our guardians.

Mechanisms that effectively police co-operative exchanges are costly in proportion to their efficacy. That crime pays brings fresh incentive to extend the arm of the law. As that arm grows longer to meet defection incentives, enforcement cost (and thus negative utility) thereby rises.

Strong deterrence may bring the desired effect because strategic planning balances probable outcomes against expected utility (FRANK 1988; SCHELLING 1960). The prospects of torture, mutilation, and a prison boyfriend when factored into the expected utility equation can enforce co-operation, even where the likelihood of getting caught is low. Excessive rewards may act as similar inducements. However, deterrence imposes fresh costs. Enforcement must avoid erroneous discipline lest beneficial co-operators get locked away or coalitions arise to combat the harsh regime. Accuracy is expensive and even then not assured. A system of terror is moreover open to fresh internal corruption of various kinds. There are, for example, strong incentives to bribe those charged with distributing justice. As before, self-interest may dictate abusing power for gain.

Generalising, the establishment of a reliable policing system may prove to be too expensive to be worth while. Each cost counts against any gain from co-operation.

### Enforcement through supernatural causation

Shifting the *actual* payoff matrix of the game through external inducement may prove inefficient because the relevant costs involved in the adjustment are too high. Suppose that we cannot afford to pay our police. Does this place co-operation out of reach? Curiously not if there are irrational players who opt for strictly efficiency play, the irrational option. in iterated exchange, if irrational players were to interact only with other irrational players, then each irrational player would fair better than the rational economist would. Moreover the extra expense of paying for enforcement is avoided.

Return to Ed and Trundle in the prison. Imagine the authorities are coming down hard on them to turn states evidence. The authorities have laid out the options, but imagine that both Ed and Trundle have poor hearing. The authorities say:

"Twenty years if the other talks and you don't."

Instead they hear:

"You go free if you don't talk you bloke."

An improbable sentence, but not impossible (especially if the prison officials speak with heavy accents.)

Here *misunderstanding* the payoff matrix brings strictly efficient rewards to both players. The assumption is that both Ed and Trundle are rationally self-interested agents. But they both get the problem wrong, and for this fortunate mistake each is better off. Of course, the reward is conditional on interaction with another player who also gets it wrong, but no less real.

The example highlights how incorrect assignments of reward value may generate substantial payoffs when confused players interact with others who similarly misunderstand. In iterated play, discrete groups of befuddled players will fair better than economic rationalists do.[14]Consider how belief and commitment to supernatural agents with specific properties constitutes a fitness enhancing illusion. If individuals believe in gods who can alter fortune in accordance with strictly efficient play such god-fearers will benefit from co-operative exchange with each other. The religious belief induces altruism. Supernatural rewards may be in kind: do good and good will be done to you; or of some equivalent value, blessed are the poor for they shall

inherit the earth, the concept of reward in the latter instance relying on other supernatural elements. A property of the relevant gods is that they trouble with mortals by imposing a payoff matrix that clearly favours co-operation. Those who believe in such gods are like the befuddled prisoners who misunderstand the payoffs of co-operation. For them, *strict efficiency* is NASH. When those who believe in gods of fortune co-operate exclusively with each other (or with other reliable altruists) the players flourish.

Imagine that Ed and Trundle both believe in the great god Zugroo. The eye of Zugroo observes all and the hand of Zugroo dispenses riches to those who act by His law. The sword of Zugroo vanquishes those who transgress His way.

| | Trunde defect | Trundle co-operate |
|---|---|---|
| Ed defect | – infinity<br>– infinity | + infinity<br>– infinity |
| Ed co-operate | – infinity<br>+ infinity | + infinity<br>+ infinity |

**Table 5:** God of Fortune Game: Submission to Zugroo game

If Zugroo exists, then co-operation is the single NASH equilibrium for the game. From the perspective of strict efficiency, however, the gods *actual* existence is an inessential detail. Zugroo himself is an imaginary being, a tissue of confusion. Yet motivations to co-operate follow directly from Ed and Trundles belief in the god. Commitment to Zugroo adds +RS value when both exchange partners share it.

Generalising, once agents believe in gods who render fortune commensurate with co-operation, religiously motivated altruism is possible.

> *Religious Altruism*: belief in supernatural causation + belief that causal agents enforce strict efficiency + [natural beliefs and desires] → motivates → commitment to altruism.

A cognitive design that distorts information flow within individuals, altering expected utilities in accordance with strictly efficient exchange could evolve alongside perceptual and motivational systems that bring such players together and keep defectors out. Religious causation and the motivational systems that underlie religious life seem to fit this description. Beliefs in the existence of supernatural agents serve to enhance strictly efficient exchange in communities of shared commitment. Belief in gods capable of altering individual fortune promotes efficient play by prompting the motivational structure to produce *strategically* co-operative behaviour. The sacrifice of the defection payoff is understood as a kind of investment, the god acting to insure desirable outcomes through supernatural causation. *The strategy works because it is based on an illusion, not in spite of any illusion*. Selection will reinforce tendencies to this illusion along side other co-relational mechanisms. If belief in supernatural causation is to evolve, there clearly need to be further constraints on the cognitive design of individuals disposed to this belief. These includes a system of projection and denial that generates supernatural commitment with zero empirical evidence; the desire to seek out con-specifics who are of a similar mind about the gods; careful attention to displays that authenticate commitment; a willingness to publicly manifest and present evidence of god commitment; mistrust of heretics; and moralistic aggression against unbelievers where the costs of defection are high.

Before exploring these and additional aspects of religious altruism, it is critical to examine how religious believers reliably signal the presence of religious commitment to others.

## Signalling Religious Commitment

Religion appears to be an efficient means for policing the social exchange. But how can players harbouring the relevant illusions find each other?

Defection pays better than co-operation, so it is always in a defectors interest to attempt to imitate a signalling behaviour. But a signal that can be imitated is worthless as a signal. More formally:

> –$RS cost of a strictly efficient play signal < +$RS value of reciprocity < +$RS value of (unpunished) defection

Hence,

> Σ [+RS value of reciprocity – $RS cost of a strictly efficient play signal] < Σ [+$RS value of (unpunished) defection – $RS cost of the strictly efficient play signal]

How then does reliable signalling evolve? With ordinary (non-religious) reciprocal altruism, the signalling of altruistic tendencies among those who are not closely related comes from an ability to (1) observe and remember past play (2) gather informa-

tion relative to past play not directly observed, and (3) follow the rule "past is precedent." AXELROD and HAMILTON have shown that simple "tit-for-tat" strategies are robust and stable over iterated play (AXELROD/HAMILTON 1981). The strategy is simple: co-operate first and imitate an exchange partner's the last move. By helping those who have helped in the past, and not helping those who with a record of defection, altruists can be reliably identified and co-operation becomes evolutionarily stable. Notice the altruistic signal here is intrinsically connected to its meaning. It is difficult for defectors to invade without acting altruistically, that is without *becoming* altruists.

Much of human social thought can be explained by placing altruistic signalling in contexts where individuals must frequently interact with many different players. Under such conditions, the theory of reciprocal altruism accounts for many aspects of social cognition, including an interest in past reciprocity, tendencies to gossip, the desire to seek and defend reputation; dispositions to advertise past altruistic efforts; the desire to enhance the status of altruistic players as well as to disguise one's own indiscretions; dispositions to falsely present oneself in an altruistic light; the tendency to self-deception about one's moral goodness in order to better deceive others, and much more, to a high order of intricacy (TRIVERS 1971, 2001). The elaborate productions of the psychological system that generates human altruism can be traced to the simple fact that altruists signal authentic altruism merely by acting altruistically. Reciprocity is what reciprocity has done.

With respect to religious altruism, the signalling of altruistic intention is harder to explain. Religious altruists cannot simply look to past examples of religious behaviour as a signal without already knowing how to detect religious behaviour. Audiences need to know what makes some behaviour a reliable signal of religious (and therefore altruistic) commitment. Crucially, linguistic utterances—declarations of faith, pious professions, etc.—are poor vehicles for signalling commitment. Atheistic defectors could merely lie their way into exchange with the god-fearing, repeating the rewards of social existence without paying any price.[15]

I have suggested that the *costs* imposed by religious cognition are themselves adaptive because they certify authentic commitment to the gods, and hence to altered expected utilities. Words are cheap, but more costly expressions may do the trick. *Crucially, not any wasteful display can ensure the reliability of a signal of religious commitment*. Suppose that growing to a height of six meters and producing colorful feathers from one's forehead (an arbitrary costly signal) emerges as a cue enabling audiences to separate the religious wheat from atheistic chaff. Because defection pays better than co-operation, defectors have an incentive to match these, or any other, arbitrary cues. As with ordinary altruism, there must be an *intrinsic* relationship between a signal of religious altruism and its meaning. It is critical to understand how this intrinsic signaling relation may be forged.

### The handicap principle

Consider religious signaling in its wider biological context. The Israeli biologists Amotz and Avishag ZAHAVI have shown from an analysis of a broad range of organic communication devices that where deception pays signals are always self-certifying. Authentication comes by way of handicaps built into the signal that strategically target specific information about the signaler. Handicapping costs disadvantages signalers as only authentic signalers can endure. In doing so, a signal's cost is always linked to the nature of the information transmitted. The ZAHAVIS call this rule "The Handicap Principle"(ZAHAVI 1975, 1977, 1987, 1993; ZAHAVI/ZAHAVI 1997) see also (Grafen 1990a, 1990b; LOTEM 1993; MAYNARD-SMITH 1993).[16]

There are innumerable examples in nature of signals that strategically handicap organisms. When approached by a predatory wolf, fit gazelles will often leap in to the air (stot), a highly puzzling action given the predatory-prey relationship. Stotting makes the gazelle both more visible and requires aerobic expenditure, flushing its muscles with lactic acid. Why would a gazelle signal its presence and then exhaust itself before a life-threatening chase? The answer is that it can afford the expense. Less fit gazelles are incapable of such feats, and so must conserve resources for effective flight. The less fit cannot afford catch-me-if-you-can signaling. Observers of gazelle/wolf interactions note that the predators rarely chase stotting prey (FITZGIBBON/FANSHAWE 1988). Stotting has evolved as an effective signaling system that enables both predators and prey to avoid pointless pursuits that impose significant costs on both organisms. Bright coloration, complex and difficult mating rituals, exposure to risk through stretching or stotting, warning cries, threats and mock fights, markings that accentuate features, song and howling—these and other costly

aspects of animal appearance and behavior all encode specific meanings. The signals handicap organisms in ways *directly related* to the signal's meaning.

By placing religious signaling in this wider biological framework, it is possible to inquire how the costs of religion are intrinsically connected to the meaning of the message conveyed. Because, presumably, the meaning of a religious signal is, "I am committed to a god of reciprocal justice," the costs that signal religious commitment must simultaneously test the bonds of precisely that commitment. The assessment should be such that those lacking commitment to a god of reciprocal justice will find it very difficult to pass.

## Religious Emotions as a Signal of Religious Commitment

The economist Robert FRANK argues that all emotional states share a common functional design. According to FRANK, emotions act as commitment devices that strategically enhance individual prospects in co-operative exchange. Paradoxically, they do this by pre-committing individuals to certain policies that may run against their strategic interests. Emotions lock people into moves that depart from NASH, in ways that tend ultimately to benefit those driven by emotion (FRANK 1988).

Emotions seem to be private affairs of the heart, but if emotions were *merely* internal guides to act in irrational ways, they would have no functional value. In order for emotions to work they must be displayed. In our own species emotions have physical manifestations, in the subtle expressions of my face, in my stride and posture, in the timbre of my voice, through blushing and tremors, each manifestation when combined with others provides information about my motivation states.

Emotions function as signalling devices by linking motivational states to physiological responses whose characteristic manifestation identifies the presence of these states. A solitary organism would have no need to wear her heart on her sleeve. But the automatic display of emotion certifies the presence of specific commitments to an audience, to better manipulate them. Manipulation is possible only if emotional displays accurately predict future responses emotions work because they are oracles.

The theory of emotions as commitment devices is deepened when viewed in light of the ZAHAVI's Handicap Principle. A cost based signalling theory predicts that emotional displays will be intrinsically related to their message. For this intrinsic relation to hold, signals must be expensive such that only a truthful signaller could produce them.

With respect to emotional display, commitment is authenticated because emotional signals 1) remain largely out of a signallers conscious control and 2) provide information about an organisms motivational state.[17] Typically, emotions generate extremely subtle and complex physiological manifestations, which are largely invariant across cultures: the dilation of pupils, perspiration, atypical facial coloration, rapid bodily vibrations or shudders, intricate facial manoeuvring, and other characteristic exhibitions denoting particular emotional states. Critically, emotions are processed in areas of the brain outside the neo-cortex, the region that governs conscious motor control. Rather emotions involve regions of the limbic system, which controls motivation and autonomic responses. The link of emotional display to motivation is so obvious that it is easy to overlook. Yet this relation is critical to the oracular function of emotional display. Knowing an organisms true motivations an audience can better predict what it will do.

Comparing forced smiles prompted by command with natural smiles, the neuroscientist V. S. RAMACHANDRAN writes: "Despite its apparent simplicity, smiling involves the careful orchestration of dozens of tiny muscles in the appropriate sequence. As far as the motor cortex (which is not specialised for generating natural smiles) is concerned, this is as complex a feat as playing Rachmaninoff though it never had lessons, and therefore fails utterly" (RAMACHANDRAN/BLAKESLEE 1998, p14). Were emotional displays easy to consciously manipulate, would lose their value as signals. Were the displays not intrinsically linked to motivation, their informational content would be uninteresting to observers.[18]

Selection could enhance the ability of organisms to consciously align emotional display with self-interest, each advance in mimicking ability in turn followed by refinements in detection. FRANK has shown that where the costs of false signalling are high, the detection ability will outpace lying ability, though when costs are lower successful mimicking can evolve (FRANK 1988, ch3; ZAHAVI/ZAHAVI 1997). The expectation is for audiences to scan signallers for subtle signs of deception (and self-deception) integrating the analysis of an emotional display with other strands of information, as for example come from the observance of past play, gossip, reputation, and so on.[19]

Turning to the religious emotions, the theory predicts emotional displays signalling authentic commitment to the gods (and so, to the altruistic group morality god belief motivates). The simplest system would link conventional emotional display to god belief. It is not surprising therefore that religious emotions are manifested as ordinary emotional displays directed to supernatural beings: hard to fake expressions of gratitude, shrinking before great authority, maternal and filial piety, fear of reprisal, hopeful expectation, sibling love for co-religionists, and so on. These emotional signals, an others, are intrinsically linked to behavioural trajectories via the motivations they assess: gratitude denoting an accumulation of debt, filial piety indicating fidelity to god's way, fear signalling an avoidance of danger (that of unavoidable supernatural punishment), and joy marking the expectation of heavenly rewards. The model predicts that:

> Religions emotions → reliably signal → [Religious belief +… → commitment to altruistic exchange]

Religious emotions, like all emotions, admit of gradations in intensity. I may love a little or fall into loves bottomless abyss, hate a little or loathe my place on the spectrum between extremes showing in my responses. In my view, the intensity of religious feeling suggests that religious altruism played a vital role in the evolution of our species.

Generalising, religious emotions reliably convey strategic information about how an agent will act in the future by exposing her religious motivations. These emotions tell and audience that the agents actions are informed by a specific conception of reality; the belief that supernaturally enforced justice holds ultimate sway.

## Ritual Action as a Signal of Religious Commitment

For religious commitment to facilitate altruistic exchange there must be public occasions in which supernatural commitments are put to test, especially for potential exchange partners who are not closely related, and so against whom there is often a special temptation to cheat. Displays of religious emotion would be worse than useless were they to occur only in private, their costs bringing no strategic advantage through the manipulation of others. The theory therefore predicts not only the display of religious commitment but also ostentatious display.

When will religious commitment be signalled? One of the paradoxes of god belief is that, on the one hand, it produces epistemic certainty and strong emotional responses, this confidence and passion certifying religious commitment as genuine. On the other hand, religious commitment must be safely contained from the business of ordinary life. Those who rely on imaginary beings to provide for their daily bread will have no daily bread (see discussion below). So audiences cannot look for evidence of religious commitment from practical dealings. Private life (in the ancestral environment as now) centres on family existence and close friendships, areas where independent measures of trust are normally available. We need no gods to love family and friends. Religion may be displayed in private, but were individuals *only* to display religious emotions outside of public view they could not manipulate others with them.

The theory of religion as a signalling system accords well with evidence of panhuman dispositions to produce and participate in rituals (Brown 1991). Selection leaves nothing important to chance. It is no accident that displays of religious commitment are prompted in special collective encounters, where emotional responses may undergo public scrutiny, where tears are matched to crocodiles. Occasions set apart from ordinary life where religious belief is publicly tested reduce uncertainty about who believes and how strongly. Such occasions provide immediate information about the relevant mental and motivational states of individuals in a community.

Given that rituals function as commitment assessments, the model predicts structural regularities beyond public display. In spite of the emphasis on creeds in some religions, transacting in verbiage is insufficient to tests the bonds of commitment, and should not be relegated to a central role in any ritual test. Defectors could merely lie, mouthing the relevant words and adding strategic "nots" to their actual selfish commitments and intentions.

The model predicts that where commitment is critical, ritual participation generally will be understood as obligatory, with failure to participate judged a species of defection, the unwillingness to be tested a sign one would likely fail.[20] If non-participation were unpunished, rituals could not serve as reliable gages of religious commitment. The temple becomes an imprecise instrument when empty.

It goes without saying that religious rituals will be geared to prompt explicit emotional reactions to the gods, mining the wells of feelings and calling up specific physiological responses. This include displays of

love and adoration, otherworldly stares and transports, the distinctive look of ecstasy, the quivering of fear, tears of joy, submission postures and others—the ritual body serving as a billboard to the believer's soul. It would be of little benefit to those interested in theological commitment to know how their fellows brush their teeth or cook meat (though this practical information may prove valuable in other contexts). Moreover, religious rituals will rarely assess emotional information unrelated to god commitment.

Critically, rituals may provide information about religious commitment through methods other than emotional prodding. The presence and strength of religious commitment can be tested by subjecting ritual goers to various traumas and ordeals. Such costs test subjects by rendering expected utilities explicit in ways directly related to supernatural belief. The trials need to be arranged so that only those actually committed to the relevant gods would be willing to subject themselves to the trials.

Consider Ed the believer deciding whether to partake in the strenuous rituals of his tradition. The costs of participating in the ritual times their frequency are discounted by the conditional probability that supernatural causation will bring about some better outcomes outweighing the costs. If Ed genuinely accords a high probability to future supernatural beneficence then for Ed:

> Cost of ritual participation x frequency < Conditional probability of value from pleasing the gods.

Consider Trundle the selfish atheist. Trundle would like to receive the spoils of defection from social exchange, but he must discount those benefits from the costs of ritual participation multiplied by their frequency. Trundle expects zero future returns to make up for these costs. Rather he anticipates only more ritual drudgery everlastingly, etc. Beyond this expense, there is the real possibility that Trundle will be caught out as defector—given this is her plan—and hence the requirement to factor in additional risk. It is easy to see that the expected utility from costly ritual action can exceed the likelihood of any advantage from cheating the devout.

So for Trundle:

> Conditional probability of value from cheating the devout < Costs of ritual participation x frequency

Notice that the ritual costs are not arbitrary. For ritual to be an effective test, it must accurately measure religious commitment. It must reliably reflect the belief in a system of supernatural causation capable of altering outcomes favourable to those who believe in it (and so act altruistically towards others similarly committed.) The logic is simple: if Trundle does not believe the gods will repay his ritual sacrifice then why should he believe they will repay his altruistic sacrifice? Whatever Trundle may say about his conviction, rituals assess whether he is willing to put his money where his mouth is.

The nature of a ritual ordeal may very widely. At one end of the spectrum it may involve exposing persons to settings that please the committed but which vex and bore the un-devout. What might be called "trials of unendurable tedium" tests authentic commitment by inflicting ennui on ritual participants who do not believe.[21]

The opposite end of the cost spectrum is distinguished by ordeals of extreme risk and denial: severe ascetic privations, the battle with carnivorous animals, immolation of expensive objects, leaping from extreme heights, trials by fire, and so on. The more arduous the test, the more effective it is at screening out those uncommitted to supernatural powers, or those whose commitments are weak.

The model predicts that where the costs of defection from religiously motivated altruism are high—as in war or famine—the more common and frequent will be rituals of extreme ordeal.[22] Partners in co-operative ventures will want strong evidence of enduring religious conviction before undertaking the risks of reciprocal exchange.

Summarizing, religious belief and emotion are insufficient by themselves to produce religious altruism. For religious altruism to evolve, individuals will have to hold their commitment open to public scrutiny. A more precise assessment of religious commitment comes when it is put directly to test. Rituals serve this function. They are public forums that prompt and assess religious understandings. Strong commitment to a system of supernatural causation implies a willingness to invest in activities that would otherwise appear pointless or dangerous. For believers, however, the ordeal is evidence of a secure investment bringing future advantage through sacred channels. For those who do not believe, or who believe only weakly, these trials are best avoided. Without any gods, the expected returns from such rituals are cannot justify their costs.

> Participation in Religious Ritual → reliably signals → [Religious Belief +… → commitment to altruistic exchange.]

It may be, of course, that rituals serve other functions as well.[23] Moreover, there may be other cognitive features that further constrain possible ritual structures.[24]

## Task Analysis: Reverse Engineering Religious Cognition

I suggested at the outset that by providing an evolutionary rational for a psychological design that actively distorts information about the world in costly ways, it is possible to open new lines of inquiry into the nature and function of religious cognition. I have argued that supernatural cognition distorts information flow within individuals in ways that enhance reciprocal exchange with their audience. It does this by making defection seem more expensive than co-operation, an illusion that fosters individual RS in communities whose members share this illusion. Altered expected utilities follow directly from the belief in supernatural agencies that dispense rewards and punishments commensurate with altruistic sacrifice. Around these beliefs various signalling and detection mechanisms producing costs that clearly identify and display religious commitment have arisen. It is therefore possible to explain the expensive illusions, feelings, and behaviours intrinsic to the religious life as signals of altruistic commitment.

It may be possible to take explanation even further. Reflecting on the optimal design of such a system may reveal more intricacy and specialisation at the level of systems dedicated to processing information relevant to religious altruism. With respect to the systems controlling the content and acquisition of religious information, the following seems likely, and worth pursuing in greater empirical detail. However, should many of these avenues lead to empirical dead ends, it may be necessary to substantially revise or abandon the theory that religion is a signalling system that propagates altruism.

## Gods

*Gods of fortune*: I have noted that the gods take an interest in human affairs, and possess powers to alter the future as it relates to the prospects of individual players, manipulating expected outcomes to motivate strictly efficient exchange. What is demanded, of course, are dispositions to believe in a particular kind of supernatural causation, one even *more specific* than category violation of the relevant intuitive kinds, as many contemporary cognitive psychologists suggest (Barrett/Nyhof 2001; Boyer 1994; Boyer/Ramble 2001). The theory predicts that the relevant supernatural causation will 1) bear on individual fortune in such a way that 2) rewards co-operation and punishes defection. Zugroo has the power to bring infinite reward and punishment, but such incentives exceed requirements of the system. Small benefits and punishments may be all that is needed to induce altruism. Moreover, the gods may be fallible yet deter defection by altering probable outcomes and hence expected utilities.

*Unjust gods*: taking the notion of fallibility further, gods need not always be imagined as perfectly just. Consider the Biblical story of Job who is a paradigm of righteousness and devotion, a man blameless and upright, one who feared God, and turned away from evil (Job 1: 1). Notably, God visits plagues transfiguring disease, and financial ruin upon Job. God also kills Jobs children. These are hardly optimal reproductive outcomes for righteous Job. Generalising, if the Gods are perfectly just, then why do bad things happen to good people? One way of avoiding theodicy is to drop the assumption that supernatural justice is perfect. Inevitably the vicissitudes of life bring tragedy to religious altruists, occasionally massive tragedy of the kind Job endures. Life also brings riches to defectors who by their deeds worship Nash. But the gods need not be perfectly just to alter expected utilities in the relevant ways. The representations should entail only that the Gods bring better lives *on balance* to those who act righteously, those who in committing to gods commit to relevant others. Those better lives could be worse than lives with no gods at all. Perhaps capricious Zugroo merely injures altruists less than defectors:

| | Trunde defect | Trundle co-operate |
|---|---|---|
| Ed defect | – infinity<br>– infinity | –1000<br>– infinity |
| Ed co-operate | – infinity<br>–1000 | –1000<br>–1000 |

**Table 6:** Evil Zugroo reward Game

Another solution is to project rewards into the future, perhaps after death. It is noteworthy that Job, in fact, is rewarded at the end of his life, where:

"The Lord restored the fortunes of Job, when he had prayed for his friends; and the Lord gave Job twice as much as he had beforeand the Lord blessed the latter days of Job more than his beginning; and he had fourteen thousand sheep, six thousand cam-

els, a thousand yoke of oxen, and a thousand she-asses. He also had seven sons and three daughtersJob lived a hundred and forty years, and saw his sons, and his sons sons, four generations. And Job died, and old man, and full of days" (Job 42: 10–14).

This is exactly as the model predicts, a Hollywood ending for Job.

*Impersonal gods*: It may be that the supernatural persons like Zugroo are the typical vehicles by which persons imagine their actions are held accountable. But religion understood as altruistic self-deception does not require cosmic individuals, and there are many instances of supernatural belief where the guiding forces are imagined as impersonal. The wheels of Karma, magical substances like sin and grace, astrological forces, the powers of witchcraft, and so on, when commonly assumed by partners in social exchange facilitate reciprocity. It is interesting that a shared belief for example in Karma, the idea that what goes around comes around—actually brings such a system of reward into existence. What goes around *really does* come around for those who exchange according to their belief in Karma, though for entirely natural reasons. Karma ensures strictly efficient action which delivers rewards to partners in exchange that are only possible when individuals make sacrifices on each others behalf.

*Indifferent gods*: There are many representations of gods who are not interested in human affairs or who though interested are impotent to alter individual fortune. The creator gods of many tribal religions certainly fit this description intellectually and morally imperfect beings, easily duped by human agents who on other occasions are charged with the task of educating them (KATZ 1984). The model proposed here does not exclude the possibility of such concepts emerging as objects of supernatural belief. It does however predict that uninterested and ineffectual gods are less likely to be candidates for extreme piety and devotion. Here the data are important: if such gods are imagined *never* to influence fortune, come what may, then costly sacrifice to them does not signal the relevant commitment to altruistic exchange. Sacrifice to an inert god is merely undirected conspicuous consumption. The model predicts that such gods, though perhaps discussed, will not provide the co-ordinating link that binds individuals together in common efforts. Emotional and ritual displays of commitment to them will be rare. Other gods will arise to fill the relevant functional roles.[25]

*Supernatural dessert and the Soul*: Given that natural justice imperfectly matches altruism and defection that defectors sometimes flourish and co-operators sometimes suffer the projection of supernatural agencies will involve the simultaneous projection of supernatural rewards and punishments that impinge on the believer beyond this world. It goes without saying that supernatural desserts will match altruistic decisions. From a long record of imperfect natural justice, it is easy to see how beliefs in a non-bodily essence, a spirit or soul, as well as in an afterlife, could emerge. Whereas an altruist may suffer material harm, there is another invisible side to an individuals existence, the life of the soul or spirit, which transcends this poor distribution. Rewards come mysterious through supernatural channels, affecting the soul in this world or in a supernatural world to come. Often, it will be possible to discern in an imaginary portrait of the gods, an image of a believers supernatural self and cosmic future.

*Gods with group effects*: Gods are frequently understood to effect group rather than individual destinies. The explanation for commitment to supernatural powers that primarily act by influencing individuals not genetically related to the believer is more complicated than belief in gods who directly influence individual fortune. Call such a deity a group god. The belief in a group god who will bring benefits *primarily to others* may seem to fall outside an explanatory framework that views supernatural belief in terms of distributary justice and individuals. In cases where the relevant supernatural causation benefits the group, as when a Yahweh restores his people to the Promised Land, rewards and punishments are not exactly visited upon individuals according to their exchange. It seems probable that expressions of conviction in gods who benefit a group are at least as common (perhaps more common) that convictions in gods who benefit individual worshipers, so they cannot be discounted as rare anomalies.

One obvious line of explanation would construe players as acting on the straightforward rational: what is good for my group is good for me, analysing group gods in terms of optimal individual strategies. Such an analysis quickly encounters trouble: a belief in a god that benefits me by benefiting my group raises the spectre of defection all over again. Why pay religious taxes to a god (and exchange fairly with others) as long as most of the others of my group will sacrifice? More problematic, it is difficult to see how selection would act on such beliefs. Whereas belief in individualistic gods evolves because audiences use the conviction of supernatural justice to certify commitment to reciprocal ex-

change, group gods seem to have the opposite effect. With a group god, there is scope for sacrificial laziness: Even if I defect, so long as others do not, the god of group fortune will help us. If one believes in a group god, then one believes in a system of justice that generates a payoff matrix that supports defection as the apparent NASH equilibrium. In a two-person group, where the group god is perceived to pay 1000 utiles for co-operation, and signalling costs 10 utiles, we have:

| | Trunde defect | Trundle co-operate |
|---|---|---|
| Ed defect | +1000<br>+1000 | +1000–10<br>+1000 |
| Ed co-operate | +1000<br>+1000–10 | +1000–10<br>+1000–10 |

**Table 7:** Group God Game

Why then, would religious commitment ever be expressed as a commitment to group-gods? A better understanding comes when we place convictions expressing commitment to such gods in the broader framework of human altruism. Many of the distortions produced by the altruism system are arrayed to make individuals appear more "beneffective" (GREENWALD 1980; TRIVERS 2001). In playing down past defections and viewing myself as genuinely good (distortions of denial and projection) I am able to more convincingly deceive others that my defects are small when compared with my virtue. Given my virtue they should want to exchange with me. Similarly, expressions of commitment to a collective god one empowered to help all of us (not merely me) will be favoured by a system established to make people appear beneffective. If my sacrifice to god is understood as a sacrifice to others (and not just self-interested way of sucking up to ultimate authority) then such sacrifices enhance my record of altruism.

Moreover, the theory advanced here predicts significant and testable constraints on the imagined nature of supernatural agency. Though capricious, group gods should not return individual piety with aggregate misfortune. The theory predicts that the good of the group generally will not come at zero-sum expense to the individually pious, and an individual's sacrifice, while benefiting the group, should not be expected to leave the sacrificer correspondingly worse off. And while the good of the individual may arrive through the good of the group, group gods should not let defection go unpunished. The model predicts that they will tend to distribute justice effectively to deter cheating. That is, group gods will also display an interest in individuals similar to that of purely individual gods.

*The evolution of god belief*: It is easy to see how selection acting gradually on genetic substrates could produce phenotypes that interpret their world as alive with cosmic agents, if small reproductive benefits accrued to these misapprehensions (GUTHRIE 1993). Given the presence of certain other psychological systems regulating altruism: an interest in social drama, fear of punishments, hope for reward, etc., it would be a small step to for selection to integrate anthropomorphic processing strategies with altruistic intuitions, enabling like minded proto-religionists to better achieve the benefits disproportionately available to social creatures. The precise steps evolution took to achieve the present design is a matter for speculation, and there are numerous possibilities. It cannot be ruled out that anthropomorphic tendencies initially performed functions unrelated to the policing of social exchange, perhaps facilitating healing through placebo like effects or promoting fitness-enhancing optimism and hope (MCCLENON 1997, 2002). There is some evidence that religious commitment still performs these functions (ELLISON 1991), and nothing in this account should be taken to rule out multiple functionality. However, once linked to altruistic sensibilities selection would have endorsed any disposition to detect the presence of god-belief in possible exchange partners, enhancing these signaling dispositions in successive generations of religionists. The expression of a signal, as I have noted, is intrinsically linked to its meaning. Any tendency in religionists to act as if the eye of a just god (or gods) sees all could serve as a candidate signal. Over time, and with the right neurological mutations, the effect would be to accentuate the vividness in passionate display to cosmic beings, reflecting heightened emotional depth and richness in the religious life. From here, selection would have endorsed any disposition to produce and maintain public ordeals rituals whose strategic organization targets the presence and level of religious commitment through the imposition of specific costs.

## Acquisition, Information Processing, and Bias

*The acquisition of religious understandings:* Recently there has been much intriguing research on the nature of religious concepts and their acquisition (BARRETT 2001; BARRETT/NYHOF 2001; BOYER/WALKER

2000; LAWSON 2000; MCCAULEY/LAWSON 2002; WHITEHOUSE 2000). Inferences concerning supernatural agents, substances, and powers seem to reflect the tacit understandings of ordinary agents, substances, and powers, with some minimal violation of intuitive expectations for the relevant kinds (BARRETT/KEIL 1998; BOYER 1999). That is, we think about gods in the same way we think about natural objects only the gods surprise us with a few extraordinary powers or properties. These results are interesting, among other reasons, because they suggest that the gods elaborated at length in theological tombs are very different from the gods of ordinary belief, and therefore serve as poor guides for psychological inquiry (BARRETT/KEIL 1998).

Is it possible to explain the acquisition and transmission of religious concepts by the vividness of such counterintuitive agents? Such an explanation leaves much unexplained. Jane may find Zugroo striking, but feel incapable of contemplating His existence, while Ed and Trundle can, and do. Jane may find Zugroo absurd, but dedicate her life to the Jesus Christ, an equally unnatural figure. With respect to Jane, a mature theory of religion needs to explain how Jane may simultaneously assent to the following propositions:

A. Zugroo violates an intuitive expectation, therefore does not exist.

B. Jesus violates and intuitive expectation and I dedicate my life to him.

Viewing religion as a signalling system that enhances altruism suggests dispositions to develop the religious understandings and practices of a social group may well be artefacts of natural design. Supernatural understandings do not spread simply because, as HUME writes, "the passion for surprise and wonder, being an agreeable emotion, gives a sensible tendency towards the belief in those events, from which it is derived" (HUME 1993, p90). In fact, roughly the opposite: our species possesses a passion for surprise and wonder and a tendency to believe in religious entities because we are designed to produce religion. While the nature of the systems that generates religious thought remains obscure, an optimal system would produce highly structured outcomes from impoverished informational inputs. The theory predicts that individuals (perhaps beginning in childhood) will take an active interest in the theological ideas of con-specifics, develop motivational commitments to these ideas, eventually producing emotional displays of these commitments and manifesting a willingness to engage in costly ritual activity to signal these commitments to others. Little is known about the initial state and development of these systems. Speculating, it may be, as with language, that the relevant information triggering their development arrives by way of an extreme poverty of stimulus, suggesting much of the structural and semantic architecture of the system is part of our genetic endowment (CHOMSKY 1988). Given the rest of what we know about the mind, it would be unsurprising if experience plays a minimal role in development, merely prompting and giving labels to pan-human religious understandings and strategies. Whether development proceeds along a fixed schedule is unknown. Given strong nativism in other areas of cognition, again looking to language as a model (CHOMSKY 2000), it is worth exploring the possibility children produce all possible religions, with experience fixing belief to some, causing children to forget certain of theseroughly the opposite of BOYER's memory theory. This much seems very likely, that in spite of massive apparent cultural diversity, all supernatural understandings are approximately the same, with variation on the margins. The view of mind expressed by DESCARTES as composed of innate understanding given in advance of any experience has been thoroughly vindicated after sixty years of research in cognitive psychology. It may be that DESCARTES will be shown correct on another and related score, namely that knowledge of Divinity is imprinted in on every human mind, though here the seas of speculation run high.

*Informational encapsulation*: The model predicts that belief in supernatural causation will be isolated to social exchange and display, and will be separated to a high degree from practical empirical understandings. Though religious commitment incurs cost, its costs should not be lethal. We should not look to the gods to build our houses, till our fields, or raise our children. Somewhat paradoxically, individuals will expect supernatural forces to be related to practical interest, that supernatural forces are mysterious connected to our lives. They will sacrifice on the basis of these beliefs, to signal commitment and on behalf of co-religionists in altruistic exchange. But they will not leave the exigencies of life up to the gods. They will believe the gods will provide, but their actions will speak differently. Outside of the altruistic system, the model predicts they will fight tooth and nail for their reproductive interests.

*Signaling costs*: As mentioned above, religious signals impose costs that assess religious commitment. In practice, the costs incurred by religious commit-

ment should be reduced to the threshold at which defectors are kept out. Again, the system that generates supernatural commitment must remain checked by and fully integrated with the systems that facilitate interaction with the natural world. Again the theory predicts its various costs should not be lethal, and modulate to the perceived payoffs of exchange.

*Receptiveness to religious signaling*: The model predicts that individuals will exhibit acute interest in whether partners in exchange are acting in ways consistent with belief in supernatural agency. That is, they will track costly behavior that signals commitment. The theory predicts, paradoxically, that they should not explicitly view their own behavior as a signal, something to be consciously manipulated to better manipulate others. Quite the opposite, discerning audiences will attend to signals that are buried from conscious control. Fully undertaking a distorted conception of reality that projects supernatural agents into the cosmos, together with capacities to identify similarly committed exchange partners, furnishes the most reliable and efficient solution to the problem of invasion by cheaters.

*Implicit denial of disconfirming evidence*: The confinement of supernatural commitment to altruism suggests there is no relationship between religious belief and the (non-social) empirical world. In developing and expressing religious conviction, religious cognition does not seek to align thought to the outlay of the natural world. Rather it actively distorts reality while prompting costly signalling behaviour. One such cost is the mistaken supernatural belief itself. Because social exchange hinges on the degree of certainty the devout accord to such beliefs (as mentioned, one fundamental aim of ritual is to assess that degree precisely) selection will act against any tendency to falliblism about ones religious conviction. Quite the contrary, it is expected that disconfirming evidence generally will be internally suppressed and openly denied. It is clear how active suppression and denial may benefit an organism exchanging with co-religions: self-deception produces more convincing displays of conviction in a religious truth. Religious belief is a distortion represented as a certainty. But its epistemic status is better understood as an output of the systems that regulates social exchange rather than of the perceptual systems that mediate the relation between organisms and non-social reality.

*Explicit moralistic denial:* The core elements of this system that produces religion are hidden from consciousness, indeed when presented as mechanisms of projection, moralistically condemned, perhaps aggressively. Explicit denial of the distortion is likely.

*The altruism system*: Given that religion facilitates altruism, an optimal psychological architecture will produce a high degree of functional mesh between the systems that generate religious signalling and the broader psychological architecture underlying non-religiously motivated social exchange. It is therefore predictable, for example, that individuals will gossip about the religion of conspecifics, describe commitment in the language of infamy and prestige, and punish and reward acts of religious defection and charity. Given this integration, discriminating between the functional domains of religious altruism and ordinary reciprocal altruism may prove difficult. For example, dispositions to make religious converts may belong to the altruism system, which takes religious commitment as the relevant exchange signal. Some of the aspects I describe as belonging to the domain of religious altruism may be more usefully understood as aspects of an altruism system working with religious material. Divergence between the outputs of religiously motivated altruism and those of the reciprocal altruism system may provide convincing evidence in support of the theory that group selection dynamics played a vital role in the evolution of our species. If religion modulates the outputs of reciprocal altruism in a way that enhances the success of groups (an in turn average RS of the genes of individuals living in strong groups) this may add empirical support to group selectionist theories (Wilson 2002). Quite apart from the group selectionist controversy, the analysis of distinctive elements of religious altruism remains critical to advancing a broader understanding of the human sociality. Again, these aspects remain almost entirely unexplored.

*Strict efficiency and god concepts:* It is important not to underestimate the difficulties in securing convergence in judgements over fair exchange. Intuitions must intersect to approximately common understandings across players whose individual interests diverge. When making a decision, a self-interested individual will ask what is in it for me? and will act on the expected utility of probable outcomes, a difficult problem, but one that pales in comparison to the determination of fairness. Gods cannot merely endorse individual self-interest, backing what individuals want to do anyway because such gods cannot secure altruism. The computational feats involved in arriving at such determinations are presumably massive, as greater degrees accuracy are

demanding in forecasting the consequences of action over longer stretches of time.[26] What view will be projected into the commanding mouth of a god?

In hunter-gather societies it is common to find explicit pronouncements about the nature of the gods, voiced by religious elites (often shamans or adepts) to whom special knowledge of the supernatural is accorded (McClenon 1997, 2002; Pearson 2002). Explicit pronouncement by religious elites, while perhaps facilitating religious conventions, cannot explain strictly efficient outcomes. In fact, the power to authorise religious conventions only makes such an explanation more urgent. Otherwise religious elites could *always* promote their individual self-interest, dressing it up to look like the will of the gods.

The solution to the problem comes by reflecting on religions broader functional role. I have suggested that religious cognition enhances reciprocal altruism it does not replace the broader system that regulates altruism. There seem to be multifarious cognitive processes through which we undertake cooperative ventures and determine fair exchange. Much of the operations of the altruism system are implicit in our sense of justice as fairness, aspects of guilt and shame, feelings of friendship and dislike, moralistic aggression, sympathy, trust and suspicion, some forms of self-deception and strategic dishonesty. Leaving possible effects of group selection to the side, the simplest assumption is that the gods are projected through the lens of ordinary reciprocal altruism, with the intuitive deliverances of that system of justice left largely in tact. As with ordinary altruism, individual conceptions of justice may vary (typically veering towards self-interested conclusions) and may be subject to explicit bartering. The model therefore predicts variation in the content of divine justice as tolerable. As with ordinary altruism, variation is bounded by the practical exigencies of exchange. Individuals will tend to infer outcomes commensurate with strict-efficiency: they will tend to infer policies of sacrifice for those who will return favour, and of punishment for cheaters.

*Convention*: The theory predicts the emergence of conventions through which gods are named and differentiated from fictions, norms of religious practice are explicated, and the standards of worship and piety are explained. As noted above, effective co-ordination frequently requires convention. Some division of religious labour will likely emerge in any small society to facilitate theological standardisation about the nature and expectation of the gods. In a species with the capacity for both for religious illusion and social differentiation, such as our own, a class of religious elites charged with instituting religious conventions may be expected arise. Again, even if the remarks of religious elites establish conventions necessary for efficient exchange, nevertheless the epistemological and motivational basis of religious intuitions that which makes these remarks plausible guides to action, must come from the implicit understandings of individuals. Utterances and marks on course merely direct a system whose internal design generates predicable outcomes.

*Religious experience*: Given the substantial sacrifice religious commitment imposes, on the one hand, and zero evidence for the gods, on the other, an effective cognitive design will generate confirming supernatural experience to support god-belief. In an optimal design, religious experience should be powerful but relatively rare and confined to areas of life not directly impinging on survival. Selection will not favour religious experiences that greatly impede the ability to hunt and gather food, or to seek out high quality sexual partners, and other practicalia. Special technologies for inducing religious experiences through music, drugs, the manipulation of bodily postures, and other means are also expected to be preserved and cultivated, if these foster religious commitment at a cost that justify altruistic returns.

*Real rewards and punishments:* Religious individuals will believe that the gods efficiently reward and punish, but will not leave it up to the gods to reward and punish. Here lies another paradox. The practical inference from a belief that the gods absolve the righteous and bring justice to enemies (viz. defectors) would ordinarily be, let the gods punish and reward. However, adopted as a general policy this inference would have disastrous effects on reproductive health of creatures prone to it, and selection will act powerfully against dispositions that favour it. If religion is to enhance survival, it must have material effects. Because the gods do not exist, they are not able to deliver the relevant punishments and rewards. For religious altruism to evolve then, real benefits must come through natural channels. The model therefore predicts that religion will not suppress the motivation to seek this-worldly justice. If defection is left entirely to non-extant gods, defection will spread.

*Doubting of Religion:* If religion is distortion in the service of altruism, it is not surprising that doubting the truth the distortion should be experienced subjectively as a kind of sedition against a group who one is enjoined to love. And it is. The expectation is

for cognitive protocols that quickly stamp out doubting, especially in cases where it is strategically unwise to defect form one's group.

*Expressions of doubt*: Verbal expressions of religious belief are insufficient to certify religious commitment. However verbal expressions of doubt warrant special concern, because presumably there is no advantage in lying about an intention to defect. The expectation is for members of religious communities to treat expressions of disbelief seriously, punishing them if the cost of defection is significant, as plots to commit any crime when discovered are punished. What is true for verbal expressions applies to lacklustre emotional and ritual performance, though here the prediction is for correspondingly less severe punishments if some display has been ventured.

*Punishments to perceived heretics*: Symbolic expressions of disbelief (in the relevant gods) may be interpreted as threats to social order, and punished. Punishments may range from simple avoidance and non-co-operation to more aggressive measures: mutilation, torture, and execution. Dispositions to punish the symbolic expressions of religion reflect the deeply social nature of religious commitment. We do not inflict such harm on those who doubt it will rain tomorrow or believe in the actuality of numbers. We do not torture the weatherman or mathematician as we do the heretic. The expectation is that expressions of religious disbelief, as for example advanced by this author, will be returned with genuine concern about his moral goodness.

*Apparent moral difference*: The theory predicts bias concerning the virtues of co-religionists and the vices of heretics. These moral properties may be given supernatural overtones, as when the faithful think of themselves as chosen or of their fate as predestined, or conceive of themselves as endowed with magical qualities—grace mana etc.—lacking in wicked heretics. Given that religious cognition is geared to produce altruism, these predictions of the faithful will generally forecast social interaction, a self-fulfilling prophecy: co-religionists will tend to sacrifice on each others behalf, and less reliably in the interests of out-group members, perhaps actively seeking their harm. This outcome actually produces real empirical evidence for the apparent moral differences separating groups, thereby fuelling inter-group bias.

*Apparent theological difference and synchronism*: If religious understandings serve to discriminate between exchange partners then an optimal system will systematically overestimate the differences between religious points of view. On the flip side, the theory predicts that individuals sitting next to each other at supernatural rituals will believe themselves to think along similar lines, even if there is wide variation in the details of theological belief. Beyond shows of religious altruism, it may be that much theological variation is permissible. All things equal, the model predicts a default bias favouring the uniqueness and particularity of religious traditions, and in turn, a default bias against universalistic theories of religion, such as the one advanced here. An optimal system would reverse these settings when new groups of formally theologically distinct communities merge into co-operative units. Agents should be expected to promote and commit to a new theological and ritual synthesis, perhaps constructed out of fragments of the older religious traditions.

*Theologians*: If the principal design function of belief in unseen realities is to ensure social co-ordination through reciprocal exchange, then it is plausible to describe theology, the practice of explicating religious commitment through analysis and ratiocination (counting angels on pins), as a kind of ostentatious religious expression. Mastery of a religious tradition shared by members of a group provides a reliable watermark of religious commitment—who wants to count angels on pins?—but it serves as a poor guide to the psychology of belief. Barrett and Keil (1998) have shown that laypersons make for poor theologians, drawing massively conflicting inferences. Again, were religion designed for empirical matters, say as a navigational system, such exposure to contradiction would prove lethal. A migratory animal cannot believe that North and South are both that-a-way (pointing to the same direction) and live long. From an engineering perspective, religion is optimised to bring dependable co-operators together at the exclusion of defectors. Belief in the gods as brokers of fortune, religious emotions, and rituals (such as theological practice) are expensive signals whose costs reliably certify commitment to altruistic exchange. The details of an individuals theological convictions do not matter very much to this aim, except when explicit as signals of commitment to a supernatural world that favours in-group reciprocity. Discrepancies at the level of religious doctrine only matter when explicated and when those explications threaten co-operative activity. Religious belief is what it does, not what it says.

*Theological correctness*: toleration for theological variation is possible within religiously circumscribed communities. Toleration ends only where

the expression of theological divergence signals defection from social exchange. In the simplest cases, toleration will end when the relevant theological convictions radically alter the payoff schedule in coordination games of partial conflict. Any reduction in accountability to supernatural law is an obvious example of such departures, and will not likely go unpunished.

*Interest in the religious conviction of others*: Given the significant role of religious commitment in social exchange, the theory predicts an interest in the religion of others, as well as dispositions to put those convictions to the test where exchange is at issue. We are interested in the natural beliefs of others, whether for example they think what I am eating is poisonous. However, the practical inferences that follow from an assessment of religious conviction should reflect moral understandings. I will not moralistically condemn someone who in good faith (but falsely) warns me that I have ingested poison, unless I think he is playing at my expense. But I may avoid and dislike a worshiper of Zugroo. Again, the expectation is for interest in religion to track strategic information relevant to exchange.

*Preference for co-religionists*: On the flip side, the model predicts that individuals will generally feel more secure and comfortable in company with those committed to the same gods, and should display a preference for arrangements that place them in association with them.

*Conversion:* The system that generates religion is expected to remain indifferent to empirical disconfirmation, but this does not mean that it incapable of change. Given the strategic nature of social exchange, an optimal system would adjust to local social circumstances, strategically modulating itself to changes in social arrangement in ways that tend to increase RS. Not only will the costs of religious display adjust to actual (non-supernatural) returns, but also the nature of the conviction displayed will be expected to change if these strategic advantages are obvious. An optimal system will motivate individuals to convert when presented with credible prospect of long-standing exchange with a social group bound to common motivating religious ideals. Reflection on the long-standing regularities (if any) in the ancestral environment, considered with reference to internal constraints imposed by architectural features of the religion system, should enable a more careful task-analysis of conversion strategies. The rhetorical techniques of missionaries and cultural anthropologists may shed further light on nature of these strategies.

*Missionary behaviour.* It may seem that a signalling based theory of religious cost cannot account for the expense of missionary activity, where (presumably) individuals display commitment to gods in front of audiences who do not believe in them. If anything, the model would seem to predict hostility and aggression in response to such displays, further escalating the costs of religious investment. What then accounts for the desire to convert others? If group selection were a strong force in the evolution of our species, then an argument could be developed that missionary investment builds stronger coalitions: missionaries sacrifice to enhance the power of their groups. But missionary cost can be explained from within the limits of ordinary selection, if genuine enhancements to the average inclusive fitness of signallers follow from their seeking of converts. Placing to the side any expected return generated through successful missionary activity (which may be far from trivial) it may be that missionaries and their kin are rewarded by *current religious affiliates*. Such benefits may come in the form of actual goods (payments and the like) or as prestige, a reputation for moral excellence. It is not hard to see how dispositions to evangelise could evolve as *signals of altruistic worth directed to ones group*. Like emotional commitment and ritual display, the costs of missionary sacrifice are intrinsically linked to their message. Only a committed believer would undertake the associated discomfort and risk of trying to convert cannibals, etc. Here there is a strong analogy to exhibitions of military courage where the risks of harm in battle are presumably balanced against the benefits generated by courageous signalling and against potential injury via one's local affiliates, who may not leave cowardice unpunished. A fully developed account of missionary psychology would need to factor all the expected benefits of sacrifice, carefully assessing the various strategic advantages of missionary behaviour and signalling. With respect to potential converts, the model predicts very little initial religious display directed to converts (where such display is likely to induce blank looks of incomprehension or hostility) and much material benevolence. Very likely, missionaries will work first to improve the concrete worldly circumstances of those they seek to initiate, signalling their altruistic value tangibly by their deeds. The pomp of religion will likely emerge later, only after the exact meanings of religious display have been explicated and conveyed.

*Agape*: Religiously motivated co-operation has its limitations. Individuals should not be prepared to make *any* sacrifice for their co-religionists. They

should rather tend to act in ways that maximise reproductive fitness, if all act as they do. This is a version of reciprocal altruism, not altruism at all costs come what may. As a strategy, selection will act against generalised love that is not reciprocated. It is theoretically possible that agape (universal love) could emerge as an optimal strategy if its costs were taken as signalling costs, and repaid indirectly by other Agapists. Agape may further advance conversion by prompting strategies of altruistic exchange in heretics. Roughly, the idea is that if the Agapist has given me something I must be indebted. Where heretics are genuine threats—especially when the benefits of agape come at risk of peril to Agapists, as in crime and warfare—selection will act against indiscriminate assistance to the out group as a form of signalling. We do not feed sharks with our hands.

There is much evidence to support this prediction. It is clear that most Christians in Northern Ireland or the Balkans do not turn the other cheek as their Gospel enjoins them to. One interpretation is that they are bad Christians. From a tactical point of view however their strategies conform to the expectations of a social mind engineered to enhance RS in ancestral conditions. A corpse cannot turn its cheek. When stakes are sufficiently high, then, the expectation is violence returned for violence and sanctioned by divine command.

As noted above, David Sloan WILSON has recently provided an intriguing group selectionist account of religious altruism consistent with altruistic sacrifice beyond reciprocity (WILSON 2002). The presence of supererogatory tendencies may be evidence that selection at the level of groups was a strong force in human evolution. Alternatively, explicit avowal of extreme sacrifice may be best explained as a strategy of self-deception. We sincerely and openly promote high- minded ideals while living by less stringent morals when this pays (ALEXANDER 1987; TRIVERS 2001). The precise extent to which individuals will sacrifice for their gods, turning their backs on established exchange partners and family, compels a more careful empirical study.

*Implicit Religion*: Religion is an artefact of natural selection, part of a panhuman psychological design and a species property. The simplest assumption is that individuals without gods should be as rare as babies lacking feet. Yet, clearly, a developmental outcome of our psychological architecture cannot exclude atheism and disenchantment with religious ritual. This seem paradoxical, a prediction that all will believe in gods that allows some may not. Paradox is resolved by noticing residual elements and vestiges of supernatural understandings are frequently expressed through secular thought and practice. "He believes in no God and worships him," William JAMES once wrote, a sentiment applicable to many contemporary materialist professions of faith. It seems clear that persons uncommitted to gods conduct their lives under the motivational influences of abstract entities that have some quasi-magical bearing on the self and its future. Such abstractions as: The Constitution, Justice, Law, Evolutionary Psychology, Philosophy, Progress, Freedom, Tradition, Community are exemplary. Even if such concepts can be reduced to the language of physics (doubtful) those who transact in them rarely bother with the details of reduction. Yet exchange with others is frequently based, at least partially, on evidence of commitment to such abstractions. And such commitments are costly in the ways religious commitments are. A prediction of the theory outlined here is that altruistic exchange in the modern world is based at least partially on costly strategic commitments to ideals conceived of as relevant to individual fortune.

## Conclusion: Religion and Distortion

Religious cognition is based on a strategic distortion of reality as god infested. Given that much social bias follows from this distortion, it may seem useful to seek rectification, perhaps using psychological theory as a lens to correct blurred misapprehensions. But if the theory advanced here is correct, it predicts the systematic denial of disconfirming evidence, with moralistic overtones, if co-religionists perceive the stakes of religious defection to be high. It also predicts advocacy of the theory as a sign of defection, and therefore suspicion of any heretic advancing it. Like a Chinese finger trap, the harder one uses theory to pull religious understandings apart, the more firmly they will likely become entrenched. Even if it were possible to delete religious understandings from human thought, it is not clear this would be desirable. It would be surprising if an overall improvement to life would be secured by the labotimization of instincts that have been selected for altruism.

**Author's address**

*Joseph Bulbulia, Victoria University of Wellington, New Zealand.*
*Email: joseph.bulbulia@vuw.ac.nz*

## Notes

**1** Read "religious belief" as ontological commitment to the existence of *x* where *x* is a supernatural entity or force. A belief that a supernatural reality does not exist will not be considered a religious belief. I assume that there is a clear intuitive difference between "supernatural" and "natural" entities (BOYER/RAMBLE 2001).

**2** This convention departs from somewhat from ordinary language where "religion" generally signifies something institutional, not psychological, and "grace" "soul" and "num" and other such essences and powers, though supernatural, are not "gods" in the ordinary language sense.

**3** "Altruism" and "defection" here are descriptive terms, and should not be interpreted as denoting normatively good or bad actions. Normative inquiry involves a separate though related philosophical analysis (HARMAN 1998–1999). The terms describe the average effects of actions not intentions, which may conflict. Burning heretics at the stake may be intended for their benefit but nevertheless damage reproductive interests. "Inclusive fitness" henceforth "RS" denotes the number of surviving offspring plus the surviving offspring of relatives weighted by degree of relatedness (HAMILTON 1964).

**4** Throughout this paper, "cost" denotes "reproductive cost" in terms of an organism's inclusive fitness or –RS.

**5** FODOR argues we are "…devices built to find out what is true" (FODOR 1986, p18) and that "…a condition for the reliability of perception, at least for a fallible organism, is that it generally sees what is there, not what it wants or expects to be there. Organisms that don't do so become deceased" (FODOR 1983, p68). The idea is that selection tends to enhance *accuracy* in perception and forecasting ability.

**6** It may be that there are *other* non-altruistic benefits produced by the systems responsible for religious thought that we can add to the benefit side of the equation. For example, religion may instil hope, facilitates healing through placebo effects, morally instruct, silence the explanatory drive, give purpose to life, or bring some other reproductive advantage. All things considered, its benefits may exceed its costs. However, the problem remains. There are cases of optimism, healing, altruism, and global explanatory indifference outside of religious circles. Presumably, human beings could have evolved greater capacities in these domains without worshipping imaginary beings. It remains obscure why cognition does not bypass the substantial expense of religious practice, causing individuals to live out their days breeding and rearing before happily facing the grave.
Problem: +$RS value of religious [hope, healing, moral instruction, explanation, meaning…] < + RS value of non-religious [hope, healing, moral instruction, explanation, meaning…]
Standard adaptationist s require some further account for why selection did not produce systems that yield these reproductive benefits without the massive additional costs that religious belief and practice impose.

**7** For recent overviews of the literature, see (BARRETT 2000; ANDRESEN 2001).

**8** ATRAN (2002) notes that with respect to social groupings "To keep the morally corrosive temptations to deceive or defect under control, all concerned—whether beggar or king—must truly believe that the gods are always watching" (pp144–145). Furthermore "the successful communication of commitment through display implies that the displays themselves are critical to commitment…" (p145). The ZAHAVIAN approach to signalling I advance below explicates the evolutionary logic of such displays, revealing how their costs are intrinsically related to the message they encode. ATRAN himself situates sacrificial cost within a larger "evolutionary landscape" of the mind, and argues that: "Religion has no evolutionary function per se. It is rather that moral sentiments and existential anxieties constitute—by virtue of evolution—ineluctable elements of the human condition, and that the cognitive invention, cultural selection, and historical survival of religious beliefs owes, in part to success in accommodating these elements" (ATRAN 2002, pp279–280) This may be so. As with language, it is probable that religion is "so large and elaborated a system that any precise characterization of the total selection pressures acting on it over evolutionary time is beyond our present ability to analyse in detail" (TOOBY/COSMIDES 1990, p761). But adaptationist reasoning may nevertheless be brought to bear on specialised sub-systems that accomplish specific functional ends. I suggest that cost elements of the cognitive systems underlying religion are best viewed as a signalling system adapted for social co-ordination, and that many religious expressions—"existential, cultural, an historical"—are structured products of this exquisitely designed cognitive device, knowledge of which may be advance by reverse engineering the types of coordination problems it solves.

**9** Evolutionary game theorist Brian SKYRMS has modelled the evolution of signalling and detection systems that foster coordination in interactions of pure mutual gain. His conclusion is that the emergence of such systems is a "moral certainty" (SKYRMS 1996, p93); see also LEWIS (1969).

**10** Consider bargaining problems. Trundle and Ed can make $1,000 through some joint financial venture. The project's success depends on the collaboration: neither can make it work without the other. Suppose Ed is already rich whereas Trundle is broke. Because every penny counts for Trundle, Ed is in a stronger bargaining position. Ed can credibly demand (say) $900 or walk away. Trundle is rational to accept the offer because $100 beats nothing. Similarly consider problems of deterrence. Suppose Trundle is in a position to steal a sheep from Ed's pasture (Ed is elsewhere enjoying other of his flock). Ed may hunt down and attempt to punish Trundle but doing so generates expense both in material resources and opportunities lost. Catching Trundle also invites the risk that other of his sheep will be stolen or flee his loving pasture. Assume the cost of punishment exceeds the value of the sheep. The logic of self-interest yields the counterintuitive conclusion that Ed should let Trundle go as the lesser of two evils.

**11** Applied to interactions common in the evolutionary environment, in our species, the Pleistocene.

**12** Evolutionary game theoretic analysis must consider the effects of strategies played out over successive generations, where payoffs relate to RS. It is interesting that with respect to the prisoner's dilemma, the analysis fails to distinguish a single strategy as optimal. Assuming random pairing, if the initial number of defectors in a population is proportionately high, and co-operation imposes substantial costs, then co-operators will typically encounter defectors and will fair less well over time, eventually vanishing. Selection in this case does ratify defection as the single pure strategy. Co-operators gain by helping kin, but when the benefits of co-operation are high, and populations initially consist of related co-operators living in close proximity—in "viscous" communities—then co-operation emerges as one of two effective strategies. But defectors living on the periphery of co-operating communities also gain, enabling them to

maintain a presence in any population. Both defection and co-operation may be optimal, depending on where an organism is placed. More generally, when pairing is randomised ideal strategies often remain indeterminate. It is not possible to predict the proportion of defectors and co-operators within a population, or even whether there that proportion remains stable over time. Ratios depend critically on 1. the precise cost/payoff matrix for the relevant games; 2. the relationship of this payoffs to RS; 3. the frequency of interaction 4. the initial proportion of defectors and co-operators 5. their genetic relatedness and 6. factors unrelated to the game, a huge category. See extended discussion in (SKYRMS 1996, ch3). This result though inconclusive, is interesting because reveals that defection is not overdetermined in cases resembling the prisoner's dilemma and random pairing. Self-interest need not always come out on top, even where interaction is purely random.

**13** Policing could come indirectly in a species with the ability to make and enforce contracts. The capacity to institute binding agreements capable of altering the pay-off matrix could also favour strictly efficient solutions. Returning to the Joint-Investment game above (note 12), if Trundle could enter into a contract with a hit man (or a spouse) enjoined to kill Trundle if he allows himself to bargain for less than $500, then Trundle could induce a rational Ed into dividing the profits evenly. His hands, after all, are tied. Some theorists view morality as a collection of such binding contracts (HARMAN 2000).

**14** There are many instances in nature where distortion is beneficial. A defector who deceives herself into believing she is a co-operator may better deceive others, thus bringing strategic advantage (ALEXANDER 1987; TRIVERS 1991, 2001).

**15** Presumably other aspects of the psychological system that regulates altruism could punish the religiously mendacious. Those prone to false professions could, for example, acquire a reputation for lying.

**16** "The Handicap Principle states that the receiver of a signal has a stake in the signal's reliability, or accuracy, and will not pay attention to it unless it is reliable. Thus signals are not arbitrary; rather, each signal is the one best suited to reliably convey the specific message it carries" (ZAHAVI/ZAHAVI 1997, p229).

**17** Consider Trundle who has stolen Ed's sheep in the deterrence problem. Assuming this is a one-off crime, catching Trundle proves to be more expensive than the price Ed could fetch for the sheep, so a rational Ed should let Trundle go. But an irrational Ed, who can effectively signal his irrationality, would fair better. If Ed mistakenly believes that no price is too high to redress Trundle's injustice and effectively signals this to Trundle, then he can avoid the expense of the theft. It is interesting to think about the many ways in which Ed could demonstrate the relevant motivations. He could show his commitment by having frequently expressed explosive moral indignation in the past, developing a reputation for violence at being wronged. He could speak of his loaded gun and let gossip do the rest. He could tattoo his arms with regalia that Trundle will interpret as "scary." Notice that Ed's irrational displays manipulate Trundle's expected utilities and that an irrational Ed fairs better than a rational Ed does. From this it is easy to see how expression of many human emotions are based on strategic misunderstandings of the world.

**18** The intricate nature of emotional signalling is easy to miss until one tries acting feelings out:

- Imagine hearing sharp, grating sounds but displaying an expression of overwhelming rapture.
- Your aircraft has fire coming off its wings and as it descends nose first toward the ground. Imagine convening an expression of sexual coyness at this.
- Imagine thinking "Absolutely, yes!" while conveying scornful denial your face.

Even trained actors have a hard time reproducing emotional states, their techniques relying subtly on their ability to brainwash themselves into character (FRANK 1988).

**19** TRIVERS (2001) argues that the machinery of self-deception, the active distortion of reality, has arisen as an anti-detection technology. We deceive ourselves so as better to deceive others. It is worth pointing out, however, that if deception is thoroughgoing, the signalling organism for all intents and purposes *really will* possess the relevant motivations. Religion is, at any rate, a form of self-deception in TRIVERS's sense: an active distortion of information flow within an organism to advance its reproductive interests through the manipulation of an audience. Whether these beliefs attach to reality is a separate question from how an individual will act in the future, in light of those beliefs.

**20** This may be one reason that television and the internet—while connecting the entire world through an electronic medium—has proved an unpopular medium for religious ritual. In spite of the incredible convenience of the new media, judgements about the relevant emotional states of one's own particular exchange partners cannot be made.

**21** It is noteworthy that, excepting anthropologists, people do not generally adopt the religious practices of persons whose religious views strike them as incredible. And even anthropologists who report "going native" may unconsciously do so to maximise exchange.

**22** See OVERHOLT (1986, pp122–142) for a discussion of Wovoka and Paiute Ghost Dance, for an example of a fervent a millennial ritual at a moment acute cultural crisis.

**23** For example they may serve as systems of moral instructions that illustrate mores. They may also serve to test one's commitment to a regime of natural authority, that of the religious elites, whose power hinges on distinguishing genuine allies from potential challengers. Moreover, rituals may serve to illustrate the hegemonic power of a religious community. In our era of nation states, shows of force through military parades or the deployment of troops and weapons to borders, or terrifying "military exercises" can be viewed as evidence of strength and warnings to potential enemies. In like fashion, it may be that religious rituals put on display the natural power of a religious community, an awesome show to potential defectors of what they are up against.

**24** WHITEHOUSE has analysed the relationship between the frequency of a ritual and its staging, noting that frequently of a ritual is a good predictor of the "pageantry" in its staging (WHITEHOUSE 2000). Rites of passage such as inaugurations and weddings are far more vivid to the senses than rites that occur daily. MCCAULEY/LAWSON (2002) have recently offered a slightly different account in which the form of a ritual bears on the frequency of its repetition. A mature theory of ritual would need to incorporate both the cognitive constraints on ritual action of the kind these cognitive anthropologists explore with biologically motivated theory of religious signalling offered here.

**25** Among the Kalahari people, where the creator gods are usually imagined as indifferent, ancestor spirits and the magical healing substance num provide the basis of emotional display and ritual interaction (KATZ 1984, 1997).

**26** It used to be assumed that subjective probability departs significantly from Bayesian probability though this assumption has recently come under fire (COSMIDES/TOOBY 1996).

## References


**Alexander, R. (1987)** The biology of moral systems. Aldine De Gruyter: New York.

**Andresen, J. (ed) (2001)** Religion in mind. Cambridge University Press: Cambridge.

**Atran, S. (2002)** In gods we trust: The evolutionary landscape of religion. Oxford University Press: New York.

**Axelrod, R. (1997)** The complexity of cooperation. Princeton University Press: Princeton.

**Axelrod, R./Hamilton, W. (1981)** The Evolution of Cooperation. Science 211: 1390–1396.

**Barkow, J. H./Cosmides, L./Tooby, J. (eds) (1992)** The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press: New York.

**Barrett, J. L. (2000)** Exploring the natural foundations of religion. Trends in Congnitive Sciences 4: 29–34.

**Barrett, J. L. (2001)** Do children experience god as adults do? In: Andresen, J. (ed) Religion in mind: Cognitive perspectives on religious belief, ritual, and experience. Cambridge University Press: Cambridge, pp. 173–190.

**Barrett, J. L./Keil, F. C. (1996)** Conceptualizing a nonnatural entity. Cognitive Psychology 31: 219–247.

**Barrett, J. L./Keil, F. C. (1998)** Cognitive constraints on Hindu concepts of the divine. Journal for the Scientific Study of Religion 37: 608–619.

**Barrett, J. L./Nyhof, M. (2001)** Spreading nonnatural concepts. Journal of Cognition and Culture 1: 183–201.

**Boyer, P. (1992)** Explaining religious ideas: Elements of a cognitive approach. Numen XXXIX: 27–57.

**Boyer, P. (1994)** The naturalness of religious ideas: A cognitive theory of religion. Univ. of California Press: Berkeley CA.

**Boyer, P. (1999)** Cognitive aspects of religious ontologies: How brain processes constrain religious concepts. In: Ahlbäck, T. (ed) Approaching Religion. Scripta Instituti Donneriani Aboensis 17: 53–72.

**Boyer, P. (2001)** Religion explained: The evolutionary origins of religious thought. Basic Books: New York.

**Boyer, P./Ramble, C. (2001)** Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. Cognitive Science 25: 535–564.

**Boyer, P./Walker, S. J. (2000)** Intuitive ontology and cultural input in the acquisition of religious concepts. In: Rosengren, K./Johnson, C./Harris, P. (eds) Imagining the impossible: Magical, scientific and religious thinking in children. Cambridge University Press: New York, pp. 130–156.

**Brown, D. E. (1991)** Human universals. McGraw-Hill: New York.

**Chomsky, N. (1988)** Language and problems of knowledge. MIT Press: Cambridge MA.

**Chomsky, N. (2000)** New horizons in the study of language and mind. Cambridge University Press: New York.

**Cosmides, L./Tooby, J. (1996)** Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. Cognition 58: 1–73.

**Ellison, C. (1991)** Religious involvement and subjective well-being. Journal of Health and Social Behaviour 32: 80–89.

**FitzGibbon, C. D./Fanshawe, J. H. (1988)** Stotting in Thompson's Gazelle: An honest signal of condition. Behavioral Ecology and Sociobiology: 23: 69–74.

**Fodor, J. A. (1983)** The modularity of mind. MIT Press: Cambridge MA.

**Fodor, J. A. (1986)** Precis of the modularity of mind. Meaning and cognitive structure. Behavioral and Brain Sciences 8: 1–42.

**Frank, R. (1988)** Passions within reason: The strategic role of the emotions. Norton: New York.

**Gould, S. J./Lewontin, R. C. (1979)** The spandrels of San Marco and the panglossian program: A critique of the adaptationist programme. Proceedings of the Royal Society of London 205: 581–598.

**Grafen, A. (1990a)** Biological signals as handicaps. Journal of Theoretical Biology 144: 517–546.

**Grafen, A. (1990b)** Sexual selection unhandicapped by the Fisher process. Journal of Theoretical Biology 144: 473–516.

**Greenwald, A. (1980)** The totalitarian ego: Fabrication and revision of personal history. American Psychologist 35: 603–618.

**Guthrie, S. (1993)** Faces in the clouds: A new theory of religion. Oxford University Press: New York.

**Hamilton, W. (1964)** The evolution of altruistic behaviour. American Naturalist 97: 354–356.

**Hardin, R. (1995)** One for all: The logic of group conflict. Princeton University Press: Princeton NJ.

**Harman, G. (1998–1999)** Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. Proceedings of the Aristotelian Society 99: 315–331.

**Harman, G. (2000)** Justice and moral bargaining. Explaining value. Clarendon Press: Oxford.

**Hume, D. (1739)** A treatise of human nature. John Noon: London.

**Hume, D. (1993)** Of miracles. In: Flew, A. (ed) David Hume: Writings on Religion. Open Court: La Salle IL, pp. 63–88. Originally published in 1749.

**Katz, R. (1984)** Boiling energy: Community healing among the Kalahari Kung. Harvard Univ. Press: Cambridge MA.

**Katz, R. (1997)** Healing makes our hearts happy: Spirituality and cultural transformation among the Kalahari Ju/Hoansi. Inner Traditions International: New York.

**Lawson, E. T. (2000)** Towards a cognitive science of religion. Numen XLVII: 338–349.

**Lewis, D. (1969)** Convention. Harvard University Press: Cambridge.

**Lotem, A. (1993)** Secondary sexual ornaments as signals: The handicap approach and three potential problems. Etologia 3: 209–218.

**Maynard Smith, J. (1982)** Evolution and the theory of games. Cambridge University Press: New York.

**Maynard-Smith, J. (1993)** The theory of evolution. Cambridge University Press: New York.

**McCauley, R. N./Lawson, E. T. (2002)** Bringing ritual to mind. Cambridge University Press: New York.

**McClenon, J. (1997)** Shamanic healing, human evolution, and the origin of religion. Journal for the Scientific Study of Religion 36: 345–354.

**McClenon, J. (2002)** Wondrous healing: Shamanism, human evolution, and the origin of religion. Northern Illinois University Press: DeKalb IL.

**Nash, J. (1951)** Noncooperative games. Annals of Mathematics 54: 289–95.

**Overholt, T. (1986)** Prophecy in cross-cultural perspective. Scholars Press: Atlanta GA.

**Pearson, J. (2002)** Shamanism and the ancient mind: A cognitive approach to archaeology. Altamira Press: Walnut Creek CA.

**Pinker, S. (1997)** How the mind works. W. W. Norton: New York.

**Plotkin, H. (1998)** Evolution in mind: An introduction to evolutionary psychology. Harvard University Press: Cambridge MA.

**Ramachandran, V. S./Blakeslee, S. (1998)** Phantoms in the brain: Probing the mysteries of the human mind. Quill William Morrow: New York.

**Schelling, T. (1960)** The strategy of conflict. Oxford University Press: New York.

**Skyrms, B. (1996)** Evolution of the social contract. Cambridge University Press: New York.

**Sperber, D. (1990)** The epidemiology of beliefs. In: Fraser, C./Gaskell, G. (eds) Social psychological study of widespread beliefs. Clarendon Press: Oxford, pp. 25–44.

**Sperber, D. (1996)** Explaining culture: A naturalistic approach. Blackwell: Oxford.

**Tooby, J./Cosmides, L. (1990)** Toward an adaptationist psycholinguistics. Behavioral and Brain Sciences 13: 760–762.

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

**Trivers, R. (1972)** Parental investment and sexual selection. In: Campbell, B. (ed) Sexual selection and the descent of man 1871–1971. Heinemann: London, pp. 136–179.

**Trivers, R. (1991)** Deceit and self-deception: The relationship between communication and consciousness. In: Robinson, M./Tiger, L. (eds) Man and beast revisited. Smithsonian Institute: Washington DC, pp. 175–191.

**Trivers, R. (2001)** Self-deception in service of deceit. In: Trivers, R. Natural selection and social theory. Selected papers of Robert Trivers. Oxford University Press: New York, pp. 255–293.

**Whitehouse, H. (2000)** Arguments and icons. Oxford, Oxford University Press.

**Wilson, D. S. (2002)** Darwin's cathedral: Evolution, religion, and the nature of society. University of Chicago Press: Chicago.

**Wilson, E. O. (1998)** Consilience: The unity of knowledge. Alfred A. Knopf: New York.

**Zahavi, A. (1975)** Mate selection: A selection for a handicap. Journal of Theoretical Biology 67: 603–605.

**Zahavi, A. (1977)** The testing of the bond. Animal Behavior 25: 246–247.

**Zahavi, A. (1987)** The theory of signal selection and some of its implications. In: Delfino, V. P. (ed) Proceedings of the international symposium of biology and evolution. Adriatica Editrica: Bari, Italy, pp. 305–325.

**Zahavi, A. (1993)** The fallacy of conventional signalling. The Royal Society Philosophical Transaction B 340: 227–230.

**Zahavi, A./Zahavi, A. (1997)** The handicap principle: A missing piece of Darwin's puzzle. Oxford University Press: New York.