Evolutionary stability

In this ESM we present a more extensive summary of the evidence in support of the evolutionary stability of the kind of cultural system we propose, where beliefs in powerful gods reduce people's willingness to shoulder the burden of punishing others for failing to cooperate. The stability of such a cultural system faces an obvious cultural evolutionary threat: invasion by free-riding defectors who are uncowed by the majority's fear of supernatural reprisal. This is an important concern, especially as religious defectors may seem difficult to detect. However, a growing body of evidence suggests that history's most successful religions have evolved to protect themselves from this possibility through three related mechanisms—indoctrination, credibility and selective enforcement of religious belief.

Most broadly, Atran and Henrich [S1] have argued that competition among religions over millennia has favored those cultural packages that most effectively exploit our cultural learning psychology so as to more deeply instill faith or belief. In other words, the religions that have survived and spread are those which effectively instill faith by exploiting our cognitive biases, using indoctrination practices built into rituals, devotions, sacrifices, etc. (see S2, S3), and learning cues such as conformity and prestige [S4-S6]. Relatedly, Alcorta and Sosis [S7] have argued that widespread initiation rites target adolescents at precisely the time when they are most susceptible to internalizing and committing to cultural beliefs.

Second, chief among the cues exploited by religions to ensure indoctrination are *Credibility Enhancing Displays* (CREDs, S5). CREDS are observable behaviors that non-believers are unlikely to imitate because they are costly. Religious martyrdom is the

quintessential example of a CRED; other examples include vows of poverty, fasts, chastity, costly sacrifices in the form of money or animals, and painful initiation rites, such as adolescent circumcision. Building on this, Henrich [S5] presents a formal cultural evolutionary model that shows how, if people rely on CREDs to identify who they should learn from (and some empirical evidence indicates they do), than culturally transmitted *beliefs* (e.g., God punishes wrong doers) can form mutually self-re-enforcing stable equilibria (ESSs) with costly behaviors. To invade this stable equilibrium, individuals would have to resist the use of CREDs as a cue to learning, a practice which is an otherwise highly adaptive feature of our evolved cultural learning psychology. Work on CREDs converges with work on the signaling of religious commitment [S3, S8]. Inspired by signaling models from biology [S9, S10], these approaches propose that seemingly costly behaviors can signal an individual's degree of commitment to the religious system. Such signals enable people to identify true believers [S3, S7, S11].

Third, in religions that rely on divine punishment, human beings may still be responsible for enforcing religion itself. People can sort believers from non-believers by observing costly signaling, and either specifically target non-believers with sanctions designed to compel them to participate in the practices that effectively instill faith in such things as the existence of a divine punisher, or exclude them from social interactions entirely. Taken together, these three lines of reasoning depict a system wherein people use earthy sanctions to enforce belief and punish non-belief, and forego the more difficult task of monitoring and punishing specific moral laxities [S12].

References

S1. Atran S, Henrich J. 2010. The evolution of religion: How cognitive by-products, adaptive learning heuristics, ritual displays, and group competition generate deep commitments to prosocial religions. *Biological Theory* 5: 18-30.

S2. Atkinson QD, Whitehouse H. 2011. The cultural morphospace of ritual form: Examining modes of religiosity cross-culturally. *Evol Hum Behav* 32: 50-62.

S3. Sosis R, Alcorta C. 2003. Signalling, solidarity, and the sacred: The evolution of religious behavior. *Evol Anthropol* 12: 264-274.

S4. Chudek M, Heller S, Birch S, Henrich J. 2012 Prestige-biased cultural learning: Bystander's differential attention to potential models influences children's learning. *Evol Hum Behav* 33: 46-56.

S5. Henrich J. 2009. The evolution of costly displays, cooperation, and religion: Credibility enhancing displays and their implications for cultural evolution. *Evol Hum Behav* 30: 244–260.

S6. Henrich J, Gil-White F. 2001. The evolution of prestige: Freely conferred status as a mechanism for enhancing the benefits of cultural transmission. *Evol Hum Behav* 22: 1-32.

S7. Alcorta CS, Sosis R. 2005. Ritual, emotion, and sacred symbols: The evolution of religion as an adaptive complex. *Hum Nature - Int Bios* 16: 323-359.

S8. Cronk L. 1994. Evolutionary theories of morality and the manipulative use of signals. *Zygon* 29: 81-101.

S9. Maynard Smith J, Harper D. 2003. *Animal Signals*. Oxford University Press.

S10. Zahavi A. 1977. The cost of honesty (Further remarks on the handicap principle). *J Theor Biol* 67: 603-605.

S11. Ruffle B J, Sosis R. 2007. Does it pay to pray? Costly ritual and cooperation. *BE Jour Econ Anal Poli* 7: 1629.

S12. Gervais WM, Shariff AF, & Norenzayan A. 2011. Do You Believe in Atheists? Trust and anti-atheist prejudice. *J Pers Soc Psychol* 101: 1189-1206.

Demographics

In this ESM, we present participant demographics for all studies.

		Study 1	Study 2	Study 3	Study 4a	Study 4b
		Median	Median	Median	Median	Median
Age	*M*	22.0	21.0	28.0	28.0	24.5
Gender	*n (%)*					
Female		11 (55)	27 (49)	33 (46)	17 (68)	42 (55)
Male		9 (45)	28 (51)	39 (54)	8 (32)	34 (45)
Ethnicity	*n (%)*					
African American		1 (5)	2 (4)	6 (8)	1 (4)	3 (4)
Arabic		1 (5)	-	-	-	-
Caucasian		8 (40)	7 (13)	60 (83)	19 (76)	63 (83)
E or SE Asian		9 (45)	26 (47)	3 (4)	1 (4)	6 (8)
Hispanic		-	-	2 (3)	-	1 (1)
S Asian			1 (5)	18 (33)	-	--
Other or N / A		-	2 (4)	1 (1)	4 (16)	3 (4)
Religion	*n (%)*					
Agnostic		-	4 (7)	8 (11)	3 (12)	7 (9)
Buddhist		1 (5)	3 (5)	2 (3)	-	3 (4)
Christian		10 (50)	15 (27)	34 (47)	14 (56)	41 (54)
Hindu		1 (5)	8 (15)	-	-	-
Jewish		-	-	1 (1)	1 (4)	-
Muslim		1 (5)	3 (6)	-	-	-
Sikh		-	4 (7)	-	-	-
Other religion		1 (5)	-	-	1 (4)	1 (1)
Atheist / no religion		6 (30)	18 (33)	26 (36)	7 (28)	24 (32)

Calculation of dependent measure

In this Electronic Supplementary Material, we describe in detail how we computed the dependent measure used in Studies 1, 2 and 4b: Altruistic punishment of non-cooperators.

The most straightforward measure of this conceptual variable was the total amount of money participants reported that they would spend punishing Player A for making selfish offers (i.e., offers of less than $10 to Player B). However, we noted that a substantial minority of participants (N = 6) sometimes punished non-selfish offers (i.e., offers of $10 or more to Player B) as well. Thus, we reasoned that the straightforward measure might reflect not only altruistic punishment of non-cooperators, but also other individual difference variables, such as unilateral aggression. We further reasoned that participants' punishment of non-selfish offers would reflect these individual difference variables only, not altruistic punishment. Therefore, because the individual difference variables were irrelevant to the present investigation, we computed an index of altruistic punishment of non-cooperators by taking the total amount of money each participant spent to punish selfish offers (< 50%) made by Player A, and partialling out the total amount of money that participant spent on non-selfish offers made by Player A. More specifically, we regressed participants' punishment of selfish offers on their punishment of non-selfish offers, and then used that regression equation as well as each participant's punishment of non-selfish offers to compute the predicted value for that participant's punishment of selfish offers. Then for each participant we subtracted the predicted value from the actual observed value, leaving with us with a number representing how much that participant punished selfish offers *above and beyond* punishment based on individual difference variables such as unilateral aggression.

Additional analyses

In this Electronic Supplementary Material, we present two additional sets of analyses. First, for Studies 1, 2 and 4b, we present results of analyses using as the dependent measure the highest offer that participants' said they would punish in the 3PPG. Although we believe that the index we computed better captures our conceptual variable of interest, the highest offer punished is the more typical output of the 3PPG. Second, given the relatively low internal consistencies of some of the measures used in Studies 4a and 4b, we present item-by-item analyses for these studies.

Study 1

We regressed participants' highest offer punished on their god beliefs and their religiosity (both centered around 0) simultaneously. This analysis showed that participants who believed more strongly in a powerful, intervening god reported less punishment of non-cooperators, β = -.59, $t(17) = 2.81$, $p = .01$; whereas more religious participants showed a trend towards reporting greater punishment, $\beta = .85$, $t(17) = 4.05$, $p = .001$. When we included it as a predictor in the regression, conservatism did not predict punishment of non-cooperators, β = -.11, $t(15) < 1$, $p = .59$, but the effect of religiosity, $\beta = .99$, $t(15) = 4.39$, $p = .001$, and the effect of god beliefs, β = -.51, $t(15) = 2.61$, $p = .02$, remained significant.

Study 2

We conducted a multiple regression, regressing participants' highest offer punished on condition (0 = not salient, 1 = salient), god beliefs and religiosity (both centered around 0), and the interactions between condition and god beliefs, and between condition and religiosity. The condition X god beliefs interaction did not quite reach significance, β = -.40, $t(49) = 1.50$, $p = .14$, but on the strength of our results with the alternative dependent measure we nonetheless broke it down by condition. When participants' god beliefs were salient, those who saw god as a powerful, intervening entity punished less than those who did not, β = -.78, $t(49) = 2.68$, $p = .01$. In contrast, when participants' god beliefs were not made salient before the 3PPG, they did not predict punishment, β = -.22, $t(49) < 1$, *ns*.

We found the reverse pattern when we probed the condition X religiosity interaction, $\beta = .67$, $t(49) = 2.47$, $p = .02$: when participants' religiosity was salient, more religious participants punished more than less religious participants, $\beta = .82$, $t(49) = 2.89$, $p = .01$. In contrast, when participants' religiosity was not salient, religiosity did not predict punishment, β = -.11 $t(49) < 1$, *ns*.

Study 4a

We repeated our main analyses examining each item separately for the scales assessing attributions of responsibility to god, perceptions of the appropriateness of punishment, and endorsement of free will. We regressed participants' god beliefs on their ratings of each of the nine items making up the three scales, all centered around 0. With regards to the items assessing attributions of responsibility, participants who believed more in a powerful, intervening god believed more that "punishing people's moral failings is up to our Maker, not other human beings," $\beta = .48$, $t(99) = 4.82$, $p < .001$, and marginally

more that "it is ok to let a bad deed go unpunished because there are forces outside of us that will ensure that wrongdoers are punished in the end," $\beta = .16$, $t(99) = 1.72$, $p = .09$. God beliefs were not, however, related to people's beliefs that "If we human beings don't enforce a moral order on each other, no one will," $\beta = .08$, $t(99) < 1$, *ns*. In other words, our predictions were confirmed on two out of three items. The only other significant effect indicated that those who believed more in a powerful, intervening god believed more that "Regardless of what external forces are at play, people are ultimately responsible for their own behavior," $\beta = .17$, $t(99) = 2.03$, $p = .04$. Although unexpected, this finding, if anything, casts further doubt on the idea that results from Studies 1 through 3 could have occurred because those who believe in a powerful, intervening god absolve wrongdoers of moral responsibility.

Study 4b

We first repeated our main analyses examining each item separately for the scales assessing attributions of responsibility to god, perceptions of the appropriateness of punishment, and endorsement of free will. We regressed punishment on condition (not salient = 0, salient = 1), each of the nine items making up the three scales, all centered around 0, and the terms reflecting the interactions between each of these items and condition. the nine items three psychological variables (perceptions of human responsibility for distributing punishment, perceptions of the appropriateness of punishment, or endorsement of free will, centered around 0), and the three interactions between condition and psychological variable to predict participants' altruistic punishment scores.

With regards to the items assessing attributions of responsibility, one significant interaction, $\beta = -.65$, $t(54) = 3.44$, $p = .001$, indicated that when participants' beliefs were salient, those who believed more that "punishing people's moral failings is up to our Maker, not other human beings," punished less, $\beta = -.49$, $t(54) = 2.94$, $p = .005$; a relationship whose direction flipped when beliefs were not salient, $\beta = .38$, $t(54) = 2.00$, $p = .05$. Another interaction approached significance, $\beta = -.32$, $t(54) = 1.49$, $p = .14$, indicated that when participants' beliefs were salient, those who believed more that "it is ok to let a bad deed go unpunished because there are forces outside of us that will ensure that wrongdoers are punished in the end," punished marginally less, $\beta = -.37$, $t(54) = 1.70$, $p = .09$; a relationship which did not emerge when beliefs were not salient, $\beta = .95$, $t(54) < 1$, $p = .80$. The interaction with the (reverse-scored) item "If we human beings don't enforce a moral order on each other, no one will," did not approach significance, $\beta = -.11$, $t(54) < 1$, $p = .56$. In other words, across both Studies 4a and 4b, the items that most supported our predictions directly reflected participants' beliefs that outside forces would punish wrongdoers. The item that consistently failed to support our predictions referred to the more vague "enforcement of moral order," and did not make explicit reference to specific external forces. Thus it was arguably the item the least suited to measure our construct of interest.

The only other significant interaction to emerge from the regression, $\beta = .90$, $t(54) = 2.36$, $p = .02$, indicated that when participants' beliefs were *not* salient, those who believed more that "There should most of the time be a punishment when someone does something wrong," actually punished less, $\beta = -.89$, $t(54) = 2.65$, $p = .01$; a relationship which was

eliminated when participants' beliefs were made salient, $\beta = .37$, $t(54) < 1$, $p = .38$. We find this interaction difficult to interpret, but in any case it does not support the alternative explanation that results from Studies 1 through 3 could have occurred because those who believe in a powerful, intervening god believe that punishment is an inappropriate consequence for wrongdoers.

We then repeated our original analyses, using the more traditional index of altruistic punishment calculated from the 3PPG. Using a regression analysis identical to the one described in the main article, the only significant effect we found was the interaction between condition and attributions of responsibility, $\beta = -.54$, $t(68) = 3.06$, $p = .003$. When participants' perceptions were salient, those who attributed more responsibility for punishing wrongdoers to god punished less than those who attributed less responsibility to god, $\beta = -.41$, $t(68) = 2.31$, $p = .02$. When participants' perceptions were not salient, however, this relationship flipped directions, $\beta = .32$, $t(68) = 2.05$, $p = .04$.