When and why people think beliefs are "debunked" by scientific explanations of their origins

Dillon Plunkett, Lara Buchak, and Tania Lombrozo

Supplementary Materials

Supplementary Table 1: Mean responses to each test statement in Experiment 5 as a function of participant's belief, explained belief, and explanation type.

	Theists				Atheists				
	Own		Other		Own		Other		
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal	Significant Effects
Importance	0.73	0.52	0.45	0.40	1.13	0.76	1.10	1.66	*†
Science Class	0.14	-0.07	0.02	-0.08	0.89	-0.04	0.48	1.00	†§
Theology Class	0.34	0.13	0.15	-0.03	0.48	-0.01	0.46	0.72	
Acceptance	0.15	-0.02	-0.25	-0.11	0.52	-0.12	0.37	0.78	*†§
Government Funding	-0.56	-0.37	-0.80	-0.50	0.83	0.22	0.23	0.66	†
Transparency[a]	-1.69	-1.89	-1.71	-1.67	-1.63	-1.78	-1.71	-1.55	
Replication[a]	-1.68	-2.09	-2.29	-2.10	-2.13	-2.16	-2.35	-2.14	‡#§

Significant effects and interactions: * = explained belief, † = participant's belief, ‡ = mechanism type • explained belief, # = explained belief • participant's belief, § = mechanism type • explained belief • participant's belief. One significant two-way and three significant three-way interactions involving presence are not labeled.

[a] Transparency and Replication items are reverse coded.

Supplementary Experiment

Unlike the explanations used in Experiments 1-4, the kinds of scientific explanations for belief that laypeople encounter in the popular press and elsewhere generally do not explicitly describe a biological process as functioning "normally" or "abnormally." Normality or abnormality may need to be inferred. Accordingly, we ran an additional experiment to test whether explanations for belief that merely imply an abnormal mechanism undermine the beliefs that they explain. Specifically, we replicated Experiment 3, except that—for example—instead of specifying that activity in a brain region was either normal or abnormal, we appealed to either "Type I neural activity" or "mini-seizures." We took the latter (but not the former) to imply abnormality. "Mini-seizures" were selected based on their appearance in popular press article about the relationship between spirituality and temporal lobe epilepsy (Hagerty, 2009).

Method

Participants.

One-hundred-sixty adults (72 female, 88 male, mean age 31) were recruited through MTurk. An additional 24 participants were excluded for failing to complete the experiment (n = 4), reporting that they might have previously participated in a similar experiment (n = 15), or failing a catch question designed to ensure close reading of the stimulus materials (n = 5).

Materials and methods.

The procedure for this experiment was identical to that of Experiment 3 except for two differences in the explanations provided (see Supplementary Table 2). Instead of being presented with an explanation that explicitly appealed to an "abnormal" or "normal" process, participants were assigned to either an *implied abnormality* or a *neutral* condition (in which the provided explanation either implied or did not imply abnormal functioning). For example, in the *implied abnormality* condition, the *neuroscience* explanation read as follows. (As in Experiment 3, vignettes varied in belief valence: whether Michael initially agreed with or disagreed with the target claim. These manipulations appear in brackets.)

> People are more likely to [believe/reject] this claim if they frequently have "mini-seizures" in the ventral striatum cortex in their brain.

Additionally, explanations spanned only two scientific disciplines, *neuroscience* and *cognitive psychology*, rather than the four disciplines in Experiment 3.

Participants answered the same question as in Experiment 3 about how Michael's confidence in his belief should change, as well as the same question about how *they* would revise their beliefs if they learned the explanation were true.

Supplementary Table 2: Explanations used in in the Supplementary Experiment

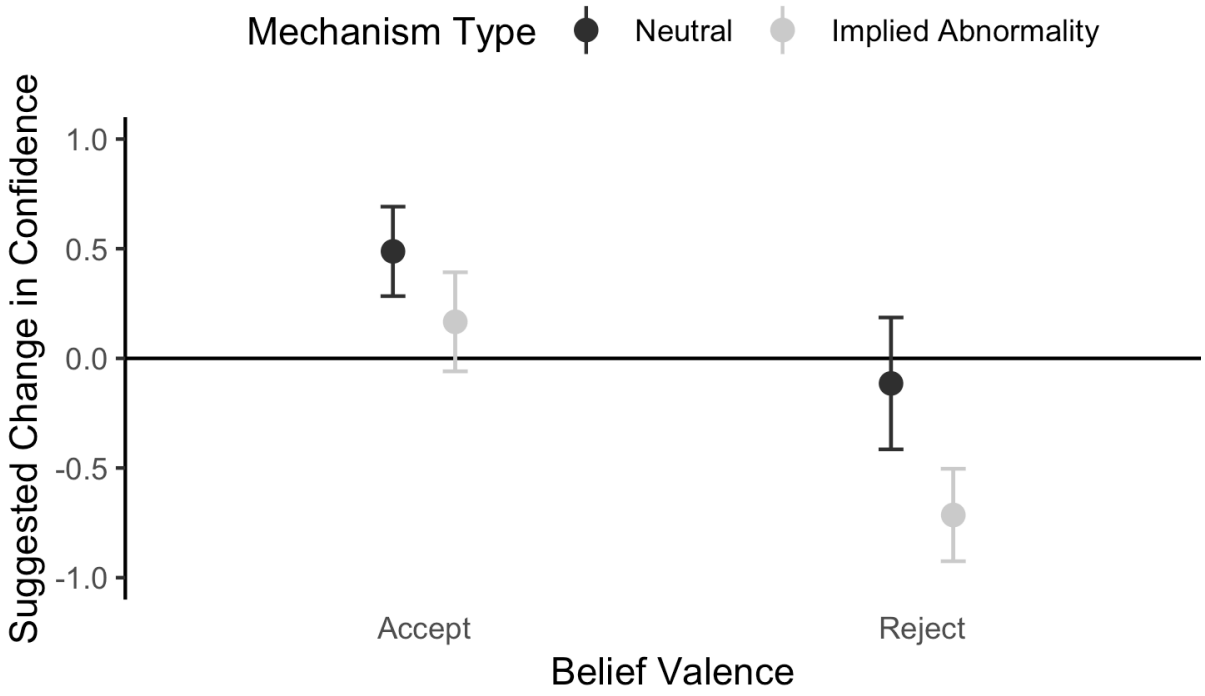
	Neutral	Implied Abnormality
Neuroscience	People are more likely to [believe/reject] this claim if they have "Type I neural activity" in the ventral striatum cortex in their brain.	People are more likely to [believe/reject] this claim if they frequently have "mini-seizures" in the ventral striatum cortex in their brain.
Cognitive Psychology	People are more likely to [believe/reject] this claim if they show a "cognitive pattern" of [embracing/rejecting] authority and finding comfort in the idea that good and bad outcomes [have causes that can potentially be identified and controlled/are beyond their control].	People are more likely to [believe/reject] this claim if they are subject to a particular "cognitive bias", namely a tendency to [embrace/reject] authority and find comfort in the idea that good and bad outcomes [have causes that can potentially be identified and controlled/are beyond their control].

Results

Effects of experimental conditions.

Responses were analyzed with an ANOVA with mechanism type (2: neutral, implied abnormality), belief valence (2: accept, reject), claim prevalence (2: common, controversial) and explanation discipline (2: neuroscience, cognitive psychology) as between-subjects factors (see Supplementary Fig. 1). As before, we collapsed across the three different domains of explained belief. Because exploratory visualization suggested a possible main effect of belief valance (absent in Experiment 3) and we had more participants per condition than in Experiment 3, we did *not* collapse across the valence of Michael's belief.

This analysis revealed marginal evidence for a main effect of mechanism type, $F(1, 144) = 3.20$, $p = .076$, $\eta_p^2 = .004$. Participants were somewhat more likely to judge that Michael should lose confident in his belief upon receiving an explanation for it if the explanation *implied*



Supplementary Figure 1: In the Supplementary Experiment, if a belief was associated with a process that was implied to be functioning abnormally, participants were somewhat more likely to suggest that a person decrease his confidence in that belief (error bars: 1 SEM).

abnormality, one-sided Welch's $t(154.22) = 2.01$, $p = .023$. As in Experiments 1-3, this effect was consistent across different belief domains (i.e., scientific, religious, and moral).

There was also a main effect of belief valence, $F(1, 144) = 10.00$, $p = .002$, $\eta_p^2 = .066$). Overall, participants in the *accept* condition advised more belief reinforcement (less undermining) than participants in the *reject* condition, but this effect did not interact with mechanism type, our primary manipulation of interest. This effect was also inconsistent across domains; it was only seen with scientific and moral beliefs. There were no other significant main effects or interactions.

Belief reinforcement or undermining.

Given the main effect of belief valence on responses (see Supplementary Fig. 1), we analyzed responses separately for each valence condition. In the *accept* condition, participants

judged that Michael should become more confident upon receiving a *neutral* explanation, M = 0.49, $t(40) = 2.39$, $p = .022$, $M = 0.49$, $t(40) = 2.39$, $p = .022$, but not an explanation that implied abnormality, $M = 0.17$, $t(41) = 0.74$, $p = .46$. In the *reject* condition, participants provided responses that were not significantly different from the scale midpoint when asked to judge how Michael should revise his confidence in the target claim when given a neutral explanation, M = -0.11, $t(34) = -0.38$, $p = .71$, but indicated that Michael should become less confident in his belief upon receiving explanations that implied abnormality, $M = -0.71$, $t(41) = -3.39$, $p = .002$.

First-person judgments.

Following the procedure used in Experiment 3, we again performed an ANOVA with first-person responses (what participants reported they would do) as the dependent variable. We performed a 2 (mechanism type) x 2 (belief valence) x 2 (explanation discipline) ANOVA with participants' belief (2: agreed with Michael, disagreed with Michael) as an additional between-subjects factor, pooling data across common and controversial claims. We found marginally significant evidence for the interaction between mechanism type and participant belief that we observed more clearly in Experiment 3, $F(1, 125) = 3.57$, $p = .061$, $\eta_p^2 = .023$, with participants who read explanations for their own beliefs indicating that their beliefs would be less reinforced (more undermined) by explanations that implied abnormal functioning, and participants who read explanations for the opposing belief saying that their beliefs would be *more* reinforced by explanations that implied abnormal functioning. This pattern of responses was seen across belief domains, except that participants did not report that abnormal explanations for opposing scientific views would be more likely to reinforce their scientific beliefs. There was also a significant interaction between explanation discipline and participants' belief, $F(1, 125) = 4.62$, $p = .034$, $\eta_p^2 = .037$, that did not interact with our primary manipulation. There were no significant main effects or other significant interactions.

Predictive Judgments

In Experiments 1-4, in addition to asking participants to make normative judgments (i.e., how a person *should* change their beliefs in response to receiving a scientific explanation of their belief), we also asked them to make predictive judgments (i.e., how a person *would* respond). In all cases, the predictive judgments were very similar to the normative judgments. Detailed analysis of the predictive judgments are reported below. (In instances where we asked participants to make first-person judgments about changes in their beliefs, predictive judgments were reported previously and the analogous normative judgments are reported here.)

Experiment 1

The key main effect of epistemic condition was also seen in predictive judgments, $F(1, 167) = 61.36$, $p < .001$, $\eta_p^2 = .42$, as was the main effect of claim prevalence, $F(1, 167) = 4.16$, $p = .043$, $\eta_p^2 = .025$. There were no new effects or interactions. As with normative judgments, responses in the *reliable* and *neutral* conditions were significantly above the scale midpoint and responses in the *unreliable* condition were significantly below the scale midpoint (*reliable*: $M = 1.36$, $t(55) = 7.29$, $p < .001$; *neutral*: $M = 0.83$, $t(58) = 4.54$, $p < .001$; *unreliable:* $M = -0.47$, $t(57) = -2.88$, $p = .006$).

Experiment 2

The same critical main effect of mechanism type was also seen for predictive judgments, $F(1, 103) = 15.15$, $p < .001$, $\eta_p^2 = .129$. Participants in the *abnormal* condition were significantly less likely than those in the *normal* condition to judge that Michael's confidence in the target claim would increase. There were no new effects or interactions, and predictive responses were also the same relative to the scale midpoint: Participants in the *normal* condition gave ratings significantly higher than the scale midpoint, $M = 1.04$, $t(54) = 7.30$, $p < .001$, but no different than the midpoint in the *abnormal* condition, $M = 0.19$, $t(51) = 1.18$, $p = .242$.

Experiment 3

Predictive judgments again showed the same main effect of mechanism type, $F(1, 242) = 33.72$, $p < .001$, $\eta_p^2 = .129$. As with normative judgments, participants in the *normal* condition provided responses significantly above the scale midpoint, $M = 0.68$, $t(131) = 5.83$, $p < .001$, and shifted toward belief-undermining in the *abnormal* condition. However, unlike the normative judgments, predictive judgments shifted far enough that they were significantly below the scale midpoint, rather than merely at it, $M = -0.37$, $t(125) = -2.80$, $p = .006$.

As reported in the paper, participants in Experiment 3 were asked to indicate how *they* would respond to receiving the explanation Michael received in addition to how they thought Michael should. We also asked participants to provide the predictive analogue of this first-person normative judgment (i.e., how they *should* respond in addition to how they *would*). We observed the same results in these judgments: an interaction between mechanism type and participant belief, $F(1, 226) = 6.25$, $p = .013$, $\eta_p^2 = .027$, and no main effects or other significant interactions. Participants indicated that "abnormal" explanations of beliefs contrary to their own should be more likely to reinforce their own beliefs than "normal" explanations of contrary beliefs, Welch's $t(131.47) = -2.36$, $p = .020$, whereas this effect was not seen for "abnormal" explanations of their own beliefs, Welch's $t(133.28) = -1.51$, $p = .132$ (although it was not significantly reversed for predictive judgments in the way that it was for normative judgments).

Experiment 4

As with the normative judgments, there was a main effect of functional normality on predictive judgments, $F(1, 192) = 28.19$, $p < .001$, $\eta_p^2 = .132$, no main effect of statistical normality, $F(1, 192) = 0.23$, $p = .63$, and no significant interaction between the two factors, $F(1, 192) = 0.20$, $p = .65$. Just as they indicated that he *should*, participants thought that Michael *would* become more confident in his belief in God if he learned that belief is associated with a

pattern of activity that indicates proper functioning in the brain (testing against the scale midpoint, $M = 1.09$, $t(108) = 8.35$, $p < .001$). In one of the only qualitative differences we observed between normative and predictive judgments, participants thought that Michael *should* become less confident in his belief in God if he learned that it was associated with improper functioning in the brain, but did not think that he *would* ($M = 0.13$, testing against the scale midpoint, $t(86) = 1.10$, $p = .27$). Finally, just as for the normative judgments, the main effect of proper functioning persisted even when accounting for differences in perceived plausibility of the explanations, $F(1, 191) = 26.80$, $p < .001$, and there was no significant relationship between plausibility and how people thought Michael would revise his belief, $F(1, 191) = 0.05$, $p = .82$.

Supplementary Experiment

Analysis of judgments about what Michael would do (in addition to what he should do) showed the same key main effect of mechanism type that we predicted, $F(1, 144) = 7.75$, $p = .006$, $\eta_p^2 = .053$: Participants thought Michael's belief was more likely to be undermined if the explanation he received for it implied that an abnormally functioning process was involved. Similarly, the main effect of valence seen in the normative judgments was also seen in the predictive judgments, $F(1, 144) = 7.73$, $p = .006$, $\eta_p^2 = .051$. The means of responses were also similar: In the *accept* condition, participants judged that Michael would become more confident upon receiving a *neutral* explanation, $M = 0.66$, $t(40) = 2.96$, $p = .005$, but not an explanation that implied abnormality, $M = 0.00$, $t(41) = 0$, $p = 1$. In the *reject* condition, participants provided responses that were not significantly different from the scale midpoint when asked to judge how Michael would revise his confidence in the target claim when given a neutral explanation, $M = 0.02$, $t(34) = 0.1$, $p = .92$, but indicated that Michael would become less confident in his belief upon receiving explanations that implied abnormality $M = -0.71$, $t(41) = -3.26$, $p = .002$.

First-person normative judgments were also consistent with first-person predictive judgments: We found the same marginally significant evidence of an interaction between mechanism type and participant belief, $F(1, 125) = 3.20$, $p = .076$, $\eta_p^2 = .022$, and a significant interaction between explanation discipline and participants’ belief that did not interact with our primary manipulation, $F(1, 125) = 8.48$, $p = .004$, $\eta_p^2 = .063$. There were no other main effects or interactions.