Because the Brain Agrees: The Impact of Neuroscientific Explanations for Belief

Dillon Plunkett (DillonPlunkett@Berkeley.Edu), Tania Lombrozo (Lombrozo@Berkeley.Edu)
Department of Psychology, 3210 Tolman Hall #1650
Berkeley, CA 94720-1650 USA

Lara Buchak (Buchak@Berkeley.Edu)
Department of Philosophy, 314 Moses Hall #2390
Berkeley, CA 94720-2390 USA

Abstract

Three experiments investigate whether neuroscientific explanations for belief in some proposition (e.g., that God exists) are judged to reinforce, undermine, or have no effect on confidence that the corresponding proposition is true. Participants learned that an individual's religious, moral, or scientific belief activated a (fictional) brain region and indicated how this information would and should influence the individual's confidence. When the region was associated with true or false beliefs (Experiment 1), the predicted and endorsed responses were an increase or decrease in confidence, respectively. However, we found that epistemically-neutral but "normal" neural function was taken to reinforce belief, and "abnormal" function to have no effect or to undermine it, whether the (ab)normality was explicitly stated (Experiment 2) or implied (Experiment 3), suggesting that proper functioning is treated as a proxy for epistemic reliability. These findings have implications for science communication, philosophy, and our understanding of belief revision and folk epistemology.

Keywords: Neuroscience explanations, intuitive epistemology, scientific communication, belief debunking

Introduction

Suppose you believe with some confidence that there exists an omnipotent God. How might your confidence in this belief be affected by learning that there is a strong correlation between having this belief and whether or not one's parents likewise believe? Alternatively, would it affect your conviction if you learned that this belief has been favored by natural selection? What if you learned that the belief is reliably correlated with activity in a particular brain region?

These questions, and their normative counterparts (e.g., how *should* your belief be affected by learning particular facts?), concern the actual and appropriate response to psychological, evolutionary, or neuroscientific explanations for beliefs. Such explanations figure prominently in the popular press, where there appears to be particular interest in neuroscientific explanations for religious belief. For example, suggestions that there could be a "God spot" in the brain, and attempts to induce religious experiences with a "God machine" that stimulates the brain, have recently been covered by CNN, NPR, the *Economist*, the BBC, ABC News, the *Telegraph*, and the *New York Times*, to name just a few. "A belief in God is deeply embedded in the human brain" begins one article from the *Independent* (Connor, 2009); A headline from an older piece in the *Observer* reads, "She thinks she believes in God. In fact, it's just a chemical reaction taking place in the neurons of her temporal lobes" (Hellmore, 1998). Such claims are at times presented as threatening to the corresponding beliefs, at times as irrelevant to them, and at times as supportive of them.

From a normative perspective, philosophers have debated whether evolutionary explanations for the origins of pragmatic, moral, and religious beliefs have implications for the truth of those beliefs (see Wilkins & Griffiths, in press), and similar questions arise for any psychological or neuroscientific explanations for belief. One reason to think that such explanations could have epistemic relevance is if they bear on the *reliability* of the mechanism generating the belief. For instance, learning that a belief results from an unreliable cognitive mechanism (as in the case of a perceptual illusion) could provide a *prima facie* challenge to the truth of the belief. Along the same lines, some recent work in experimental philosophy aims to challenge philosophical conclusions by showing them to be the product of mechanisms that do not reliably track what they purport to track (e.g., Alexander, Mallon, & Weinberg, 2010)

Here, we investigate whether brain-based explanations for beliefs are seen as having a reinforcing, undermining, or epistemically-neutral effect on confidence in those beliefs. Neuroscientific information is of particular interest in this domain not only because of the attention it garners in the popular press, but also because previous research has found that the inclusion of neuroscience information can impair the ability of non-experts to assess the quality of an explanation (Weisberg et al., 2008), and may also influence judgments of scientific rigor and moral responsibility (e.g., see Schweitzer, Baker, & Risko, 2013). Judgments about epistemic value could also be impacted by neuroscientific information.

One hypothesis is that all brain-based explanations will be seen as epistemically-threatening, as they ground the source of a belief in its proximate realization rather than its evidential base in the world. Alternatively, responses to brain-based explanations could differ depending on the epistemic relevance of the information provided. For example, a pattern of brain activity known to underwrite irrational beliefs might decrease confidence in the corresponding belief, while brain activity that is reliably

associated with accurate beliefs could increase confidence. On this view, simply identifying the neural basis for a belief should have no effect on confidence, unless that neural basis bears on whether the mechanism generating the belief is reliable in the sense that it is "truth-tracking."

In three experiments, we test these hypotheses by asking participants how the protagonist of a vignette will and should respond to receiving epistemically-neutral, supportive, or undermining neuroscientific information about his beliefs. In light of the findings from Experiment 1, we additionally test (and find support for) the hypothesis that implied normality in neural function is treated as a proxy for epistemically-relevant truth-tracking.

Experiment 1

In Experiment 1, participants read about an individual who received a brain-based explanation for one of his beliefs. They were then asked to indicate how the individual's confidence in that belief *would* and *should* change in response. The brain-based explanations varied in whether they were epistemically-undermining (i.e., linking the belief to a brain region associated with false beliefs), epistemically-supportive (i.e., linking the belief to a brain region associated with true beliefs), or epistemically-neutral. The beliefs varied in their domain (scientific, religious, and moral) and perceived prevalence (common versus controversial).

Method

Participants 229 adults (93 female, mean age 33) were recruited through the Amazon Mechanical Turk marketplace (MTurk) and participated in exchange for monetary compensation. Of these, 60 were excluded prior to analysis for failing to consent, failing to complete the experiment, having previously participated in a similar experiment, or failing a catch question designed to ensure close reading of the stimulus materials.

Materials and Procedure Different versions of the task involved one of two claims in each of three domains: science, religion, and morality (see Table 1). For each domain, one claim was "common," in that it is perceived as widely endorsed, and the other "controversial." For example, the common moral claim was "Killing an innocent person is morally wrong." The controversial moral claim was "Killing animals for human consumption is morally wrong." For each participant, one of the six claims was selected at random to be the target claim.

Participants were randomly assigned to one of the three information conditions (*positive*, *negative*, or *neutral*) and, after an initial page,[1] read a vignette. In each vignette, Michael, a participant in a psychology experiment, initially learns that a particular region in his brain—the "posterior striatum cortex"—was active when he considered his belief about a target claim. Michael subsequently learns additional information about the posterior striatum cortex. In the *positive* condition, Michael learns that the posterior striatum cortex is associated with accurate beliefs. In the *negative* condition, Michael learns that the posterior striatum cortex is associated with inaccurate beliefs. In the *neutral* condition, Michael learns only that the posterior striatum cortex is associated with beliefs of a certain kind (moral, religious, etc.). The vignette in the *positive* or *negative* condition was as follows (where text specific to this example, a common moral belief, is in bold):

> Michael decides to participate in a psychology experiment that involves having his brain scanned by a functional magnetic resonance imaging (fMRI) machine. During the scan, the researcher asks him a series of questions, including one about whether **killing an innocent person is wrong.**
>
> Michael believes the following claim, and tells the researcher this when he is asked.
>
> CLAIM: **Killing an innocent person is morally wrong.**
>
> After the experiment, the researcher tells Michael that there was activity in his posterior striatum cortex when he expressed his belief that **killing an innocent person is wrong.**
>
> Michael later reads in a reliable textbook that activity in the posterior striatum cortex is associated with [true/false] beliefs. When a person expresses a belief, and doing so is accompanied by activity in this brain region, the belief is usually [correct/incorrect] (even if the person expressing it has [low/high] confidence that it is true).

In the *neutral* condition, the last paragraph read:

> Michael later reads in a reliable textbook that activity in the posterior striatum cortex is associated with beliefs related to **morality**. For example, when a person expresses a belief that **killing an innocent person is wrong,** there is usually activity in the posterior striatum cortex.

Next, participants were asked several questions,[2] including the following two, labeled in italics:

> *Predictive* What effect do you think learning these facts will have on Michael's belief about whether **killing an innocent person is wrong**? Specifically, will it make him

[1] Participants in all three experiments began by reporting the extent to which they agreed with all six investigated claims, as well as six other claims matched for domain and prevalence (e.g., "Torturing an innocent person is morally wrong" as the matched common moral claim).

[2] In all three experiments, participants were also asked two questions about their own reaction if they imagined themselves in Michael's position, and several diagnostic and debriefing questions (including an instructional manipulation check, a free-response debriefing, and a question about the prevalence in the United States of the six investigated claims and six filler claims).

more confident that it is false that **killing an innocent person is wrong** or more confident that it is true that **killing an innocent person is wrong**?

Normative What effect do you think learning these facts should have on Michael's belief about whether killing an innocent person is wrong? Specifically, should it make him more confident that it is false that **killing an innocent person is wrong** or more confident that it is true that **killing an innocent person is wrong**?

Answers to these questions were selected from a seven-item scale ranging from "Much more confident that it is false" (recorded as -3) to "Much more confident that it is true" (recorded as 3).

Results and Discussion

The data for each of the questions above were analyzed with a 3 (information condition) x 3 (claim domain) x 2 (claim prevalence) ANOVA (see Figure 1). For the *predictive* response, this revealed a main effect of information condition, $F(2, 151) = 61.81$, $p < .001$, $\eta_p^2 = .450$, with no other significant effects. Participants in the *positive* and *neutral* conditions judged that Michael's confidence would increase, with a stronger effect in the former condition, while those in the *negative* condition predicted a decrease in confidence. All pairwise differences between conditions reached significance ($p < .001$).

For the *normative* response, there were main effects of information condition, $F(2, 151) = 27.64$, $p < .001$, $\eta_p^2 = .268$, and claim prevalence, $F(1,151) = 4.00$, $p = .047$, $\eta_p^2 = .026$, with no other significant effects. Participants in the *positive* condition were marginally more likely to judge that Michael *should* become more confident in the target claim than participants in the *neutral* condition ($p = .096$), who were in turn more likely to judge that Michael *should* become more confident in the target claim than participants in the *negative* condition ($p < .001$). Participants were also more likely to believe that Michael *should* become more confident in the target claim if the target claim was common, $t(167) = 2.26$, $p = .025$.

We compared mean responses against the scale midpoint to assess whether different information conditions had reliably reinforcing or undermining effects on belief. Participants in the *positive* and *neutral* conditions were significantly more likely than expected by chance to judge that Michael *would* (*positive*: $t(55) = 13.70$, $p < .001$; *neutral*: $t(55) = 7.16$, $p < .001$) and *should* (*positive*: $t(55) = 7.29$, $p < .001$; *neutral*: $t(55) = 4.41$, $p < .001$) become more confident in his belief in the target claim—that is, they found the information "belief reinforcing." In contrast, participants in the *negative* condition were significantly more likely than expected by chance to judge that Michael would, $t(56) = -3.45$, $p = .001$, and should become less confident in the target claim, $t(56) = -2.95$, $p = .005$—that is, they found the information "belief undermining."

In sum, participants' judgments about whether another person *would* and *should* adjust his confidence in a belief were appropriately responsive to information about the reliability of the mechanism generating the belief. When a belief was associated with a "truth-tracking" brain region, participants anticipated and endorsed belief reinforcement; when it was associated with a brain region linked to false belief, participants anticipated and endorsed belief undermining. Curiously, responses in the *neutral* condition followed the same qualitative pattern as those in the *positive* condition: information that was intended to be epistemically-neutral was taken to be belief reinforcing, a finding that we take up in Experiment 2. The same pattern of responses across information conditions was found for all three domains and for both common and controversial claims (although values for controversial claims were shifted towards lower confidence).

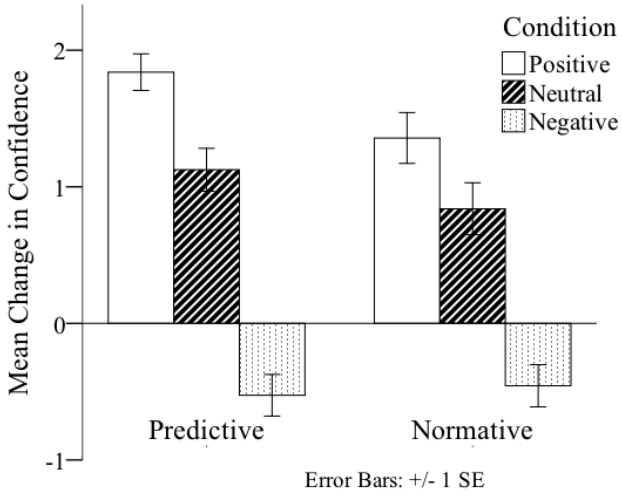

Figure 1: Belief change as a function of information condition in Experiment 1

Experiment 2

Experiment 2 examined why neuroscientific information presented in seemingly epistemically-neutral terms prompted participants to predict and advise belief reinforcement. Participants in the *neutral* condition from Experiment 1 were told that Michael had activity in his posterior striatum cortex when evaluating a claim from a particular domain, and that the posterior striatum cortex is associated with beliefs about that domain. We hypothesized that this information may have implied that Michael's posterior striatum cortex was functioning "normally" or as it should, and that this sense of reliable functioning may have been treated as a proxy for something like epistemic reliability or truth-tracking, leading to belief reinforcement.

To test this hypothesis, we presented participants with scenarios similar to the *neutral* condition of Experiment 1, but specified whether the relevant brain region was functioning "normally" or "abnormally." Our prediction was that participants would treat the former condition as belief reinforcing (like the *positive* and *neutral* conditions from

Table 1: Claims used in all three experiments

	Common	Controversial
Scientific	Some diseases are caused by microorganisms called 'germs' that can infect a host organism	Humans evolved via natural selection and share common ancestry with many other species
Religious	There is a God	Every person has a soulmate or life partner who has been preselected for him or her by God or some other spiritual force in the universe
Moral	Killing an innocent person is morally wrong	Killing animals for human consumption is morally wrong

Experiment 1), and the latter condition as belief undermining (like the *negative* condition).

Method

Participants 127 adults (42 female, mean age 32) were recruited through MTurk**.** Of these, 21 were excluded following the same exclusion criteria employed in Experiment 1.

Materials and Procedure The six claims from Experiment 1 were employed in Experiment 2. Participants were randomly assigned to either the *normal* condition or the *abnormal* condition and a target claim was selected at random. Participants were presented with the vignette below. (As before, the target belief in this example is the common moral belief, and text specific to it is in bold.)

> A new biotech company is studying a part of the brain called the posterior striatum cortex. The posterior striatum cortex is broadly associated with **moral** beliefs. When an individual expresses a **moral** belief, the posterior striatum cortex is active. However, there is no connection between the exact pattern of activity in the posterior striatum cortex and how confident an individual is in that belief. There is also no connection between activity in the posterior striatum cortex and whether the belief is actually true.
>
> Michael knows all of this information, and decides to volunteer for an experiment being performed by the biotech company. Michael has his brain scanned by a functional magnetic resonance imaging (fMRI) machine. During the scan, the researcher asks him a series of questions, including one about whether **killing an innocent person is wrong**.
>
> Michael believes the following claim, and tells the researcher this when he is asked.
>
> CLAIM: **Killing an innocent person is morally wrong.**
>
> After the experiment, the researcher tells Michael that there was activity in his posterior striatum cortex when he expressed his belief that **killing an innocent person is wrong**. The researcher also tells Michael that the specific pattern of brain activity observed in his brain suggests that his posterior striatum cortex is working [normally/abnormally].

As in Experiment 1, participants answered a *predictive* and a *normative* question. These questions, reproduced below, were answered on a seven-point scale ranging from "Much less confident that it is true" to "Much more confident that it is true" (again, recorded as -3 and 3, respectively).

> *Predictive* What effect do you think learning these pieces of information will have on Michael's belief that **killing an innocent person is wrong**?
>
> *Normative* What effect do you think learning these pieces of information should have on Michael's belief that **killing an innocent person is wrong**?

Results and Discussion

The data for each of the questions above were analyzed with a 2 (information condition) x 3 (claim domain) x 2 (claim prevalence) ANOVA (see Figure 2). For the *predictive* response, this revealed a main effect of information condition, $F(1, 94) = 14.59$, $p < .001$, $\eta_p^2 = .134$, with no other significant effects. Participants in the *normal* condition were significantly more likely than those in the *abnormal* condition to judge that Michael would become more confident in the target claim, $t(104) = 4.20$, $p < .001$.

For the *normative* response, analysis revealed a main effect of information condition, $F(1, 94) = 9.19$, $p = .003$, $\eta_p^2 = .089$, and a significant interaction between claim domain and claim prevalence, $F(2, 151) = 3.28$, $p = .042$, $\eta_p^2 = .065$, with no other significant effects. Participants in the *normal* condition were significantly more likely than those in the *abnormal* condition to judge that Michael *should* become more confident in the target claim, $t(104) = 3.53$, $p = .001$. The interaction reflected the fact that participants were more likely to judge that Michael should become more confident in global warming compared to germ theory (in the science domain), and were less likely to believe that Michael should become more confident in the existence of soul-mates compared to the existence of God (for religion). However, no pairwise differences between claims were significant.

On average, participants in the *normal* condition were significantly more likely than expected by chance to judge that Michael *would* become more confident in the target claim, $t(54) = 7.30$, $p < .001$, and to judge that he *should* become more confident in it, $t(54) = 4.27$, $p < .001$. By contrast, participants in the *abnormal* condition were no more likely than expected by chance to judge that Michael

would become more confident in the target claim, $t(50) = 0.85$, $p = .398$, or that he should become more confident in it, $t(50) = -0.46$, $p = .651$.

In sum, explicitly contradicting any normality implied in the *neutral* condition from Experiment 1 significantly reduced—and potentially eliminated—the belief reinforcement effect observed in the *neutral* condition, consistent with our hypothesis that "normal" and "abnormal" functioning were taken as cues to epistemic reliability. Surprisingly, however, we found that explicit abnormality did not reliably lead to belief undermining. On average, it neither reinforced nor undermined belief.

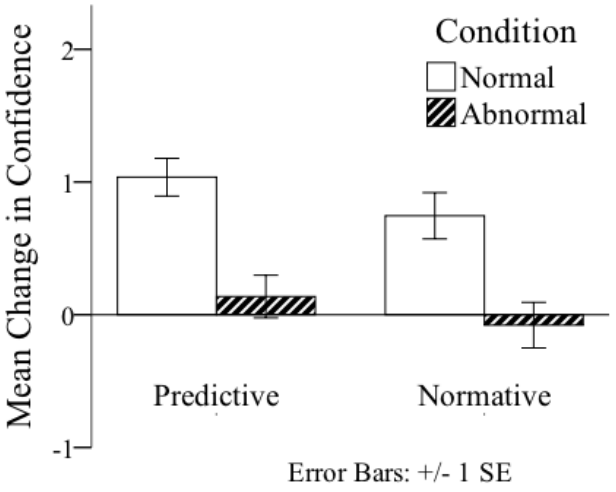

Figure 2: Belief change as a function of information condition in Experiment 2

Experiment 3

Experiment 3 was designed to replicate and extend the findings from Experiment 2 in a context that better matches popular discourse about science: one in which normality or abnormality must be inferred rather than being explicitly stated. Specifically, we investigated whether neuroscientific correlates of belief that imply significant abnormality are treated as epistemically-undermining. For the *implied abnormality* condition, "mini-seizures" were selected from a popular press article that discussed the connection between spirituality and temporal lobe epilepsy, or "faith [and] an electrical impulse that's gone awry" (Hagerty, 2009).

Method

Participants 185 adults (78 female, mean age 31) were recruited through MTurk. Of these, 29 were excluded following the criteria employed in Experiments 1 and 2.

Materials and Procedure Following the previous two experiments, one of the six investigated claims was selected to be the target claim for each participant. Vignettes also varied belief valence: whether Michael previously agreed with the claim (the *accept* condition) or disagreed with it (the *reject* condition). Participants were randomly assigned to either the *no abnormality* condition or the *implied abnormality* condition and to either the *accept* condition or the *reject* condition. They then read a vignette like the one below.[3] (As in the two previous experiments, the target belief in this example is the common moral belief and details specific to it appear in bold. The *accept* or *reject* condition manipulations appear in brackets.)

> Michael comes across the following claim on a website:
>
> CLAIM: **Killing an innocent person is morally wrong.**
>
> Michael has not given a lot of thought to whether **killing an innocent person is wrong.** But, if he were asked what he thinks about the claim he just read, he would say that he believes that it is [true/false].
>
> Michael next reads the following fact in a book:
>
> FACT: People are more likely to [believe/reject] this claim if they frequently have "mini-seizures" in the ventral striatum cortex in their brain.
>
> Michael trusts the book and believes the fact that he just read.

In the *no abnormality* condition, the "FACT" instead read:

> People are more likely to [accept/reject] this claim if they have "Type I neural activity" in the ventral striatum cortex in their brain.

Participants answered *normative* and *predictive* questions like those from Experiment 1. Responses were given on the same seven-item scale, but the data were coded relative to Michael's initial belief: responses of "Much more confident that it is false" were recorded as -3 for subjects in the *accept* condition and as 3 for subjects in the *reject* condition (and vice versa for "Much more confident that it is true").

Results and Discussion

The data for each of the questions above were analyzed with a 2 (information condition) x 2 (belief valence) x 3 (claim domain) x 2 (claim prevalence) ANOVA (see Figure 3). For both the *predictive* and the *normative* response, this revealed a main effect of information condition $F(1, 55) = 14.32$, $p < .001$, $\eta_p^2 = .207$, and $F(1, 55) = 7.32$, $p = .009$, $\eta_p^2 = .117$, respectively, with no other significant effects. Participants in the *no abnormality* condition were significantly more likely than those in the *implied abnormality* condition to judge that Michael would and should become more confident in his belief $t(77) = 3.28$, $p = .002$; $t(77) = 2.61$, $p = .011$.

In addition, participants in the *no abnormality* condition were significantly more likely than expected by chance to judge that Michael would become more confident in the target claim, $t(39) = 2.11$, $p = .042$, while participants in the *implied abnormality* condition were significantly more likely than expected by chance to judge that Michael would

[3] This experiment also included non-neurological explanations; here we report only data from the brain-based conditions.

become less confident in the target claim, $t(38) = -2.55$, $p = .015$. For the *normative* question, differences from chance were marginally significant in the same directions, $t(39) = 1.78$, $p = .08$, and $t(38) = -1.95$, $p = .058$.

In sum, Experiment 3 replicated the finding that stated or implied normality leads to belief reinforcement (as in the *neutral* condition in Experiment 1 and the *normal* condition from Experiment 2), and found that implied abnormality—in the form of mini-seizures—leads to belief undermining (as in the *negative* condition in Experiment 1).

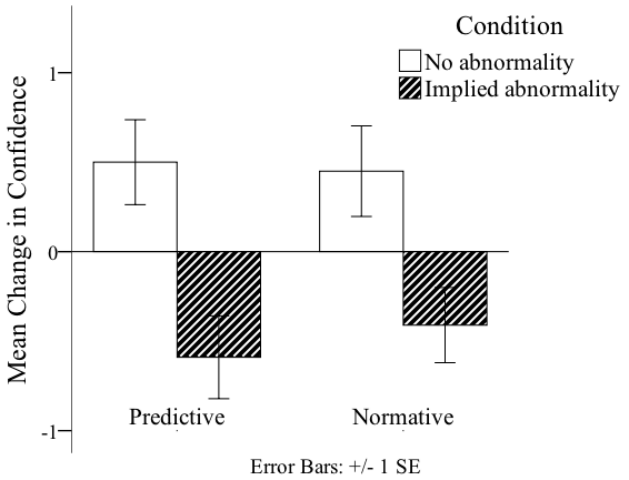

Figure 3: Belief change as a function of condition in Experiment 3

General Discussion

In three experiments, we find that neuroscientific explanations for beliefs impact perceived confidence in those beliefs. In Experiment 1, participants were appropriately responsive to information that was explicitly epistemically-relevant. However, information that was (arguably) epistemically-neutral was judged to *reinforce* belief. Experiment 2 replicated this effect and found that it could be counteracted by specifying that the brain activity was abnormal, even though it was stipulated to bear no connection to the belief's truth. Finally, Experiment 3 found that neuroscientific explanations that *imply* abnormality are taken to undermine the beliefs that they explain.

These findings help explain the many divergent responses to neuroscientific explanations for belief found in the popular press. Depending on whether findings are seen to imply normal or abnormal function, laypeople (and scientists) may take them to have different epistemic consequences, or no epistemic consequences at all. Our findings also offer some insights into "folk epistemology": they suggest that proximate (neuroscientific) explanations for belief are not taken to be universally undermining of belief, and, more surprisingly, that "normal functioning" may be treated as a cue to the epistemic reliability of a mechanism for belief formation. In other words, normality may be taken to imply truth-tracking, an implication that is only warranted under substantive assumptions about the evolution of the brain and how it develops. Finally, this folk epistemology bears a striking resemblance to philosophical theories about when we are entitled to a belief, most notably Plantinga's (1993) view that a belief is warranted insofar as it is produced by cognitive processes that are functioning properly; thus, our findings may also shed some light on philosophical debates.

Of course, many questions remain. These results are limited to third-person judgments, in which participants predicted and prescribed the actions of another. Though these judgments inform our understanding of folk epistemology and have implications for science communication, the equivalent first-person questions are also of interest. Additionally, in ongoing work, we are investigating effects of other kinds of explanations for belief. We have found evidence that psychological, developmental, and genetic explanations reinforce belief, and that this reinforcement is mitigated by indications of abnormality. These lines of research offer potentially new insights into people's understanding of science and of the relationship between the mind and the world, and shed light on an unexplored facet of folk epistemology.

Acknowledgments

This research was supported by a McDonnell Scholar Award and NSF grant DRL-1056712 to T. Lombrozo.

References

Alexander, J., Mallon, R., & Weinberg, J. (2010). Accentuate the negative. *Review of Philosophy and Psychology*, *1*, 297-314.

Connor, S. (2009, March 10). Belief and the brain's 'God spot'. *The Independent*. Retrieved from http://www.independent.co.uk/news/science/belief-and-the-brains-god-spot-1641022.html

Griffiths, P. E. & Wilkins J. S. (in press). When do evolutionary explanations of belief debunk belief? *Darwin in the 21st Century: Nature, Humanity, and God.* Notre Dame, IN: Notre Dame University Press.

Hellmore, E. (1998, May 3). She thinks she believes in God. In fact, it's just a chemical reaction taking place in the neurons of her temporal lobes; Science has gone in search of the soul. *The Observer*. p. 20.

Hagerty, B. B. (2009, May 19). Are spiritual encounters all in your head? *NPR*. Retrieved from http://www.npr.org/templates/story/story.php?storyId=104291534

Plantinga, A. (1993). *Warrant and Proper Function*. New York: Oxford University Press.

Schweitzer, N. J., Baker, D. A., & Risko, E. F. (2013). Fooled by the brain: Re-examining the influence of neuroimages. *Cognition, 129*, 501-511.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20,* 470-477.