Brit. J. Phil. Sci. **0** (2015), 1–31

The Belief Illusion

J. Christopher Jenson

ABSTRACT

I offer a new argument for the elimination of 'beliefs' from cognitive science based on Wimsatt's ([1981]) concept of robustness and a related concept of fragility. Theoretical entities are robust if multiple independent means of measurement produce invariant results in detecting them. Theoretical entities are fragile when multiple independent means of detecting them produce highly variant results. I argue that sufficiently fragile theoretical entities do not exist. Recent studies in psychology show radical variance between what self-report and non-verbal behaviour indicate about participants' beliefs. This is evidence that 'belief' is fragile, and is thus a strong candidate for elimination.

1 Introduction

The received view among philosophers of psychology takes folk psychology to be a good starting point for theorizing in cognitive science. Many philosophers

doi:10.1093/bjps/axv006

(Botterill and Carruthers [1999]; Braddon-Mitchell and Jackson [2007]; Fodor [1987]; Sterelny [1990]) advocate taking folk psychological categories like 'beliefs' and 'desires' to pick out real mental states. Jerry Fodor, for example, has been quite explicit in advocating the scientific vindication of folk psychology: 'Holding onto the attitudes—vindicating common sense psychology—means showing how you could have [. . .] a respectable science whose ontology explicitly acknowledges states that exhibit the sorts of properties that common sense attributes to the attitudes' (Fodor [1987], p. 10). Fodor wrote this in the midst of a debate with eliminative materialists (Churchland [1981], [1986]; Ramsey *et al*. [1990]; Stich [1983]) who argued that folk psychology is a radically inadequate theory. However, even at the height of this debate, eliminative materialism was always a minority view and received its death knell when Stephen Stich himself argued that eliminativism depends on which theory of reference is true, but variance in reference intuitions make it difficult to see which theory of reference is, in fact, true (Stich [1996]).

There has been a surprising amount of resistance to the very idea of eliminativism even from philosophers who claim to have a naturalistic approach to philosophy. The thought seems to be that folk psychology is so close to home, so essential to everyday use, that it could not be wrong. Or that if it is, this would be an unmitigated disaster. Fodor writes, 'if commonsense intentional psychology really were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species' (Fodor [1987], p. xii). It should not be surprising that our intuitions about how the world works are often wrong. The history of science consists of a string of counter-intuitive results. For example, it is surprising to learn that the natural state for an object is movement at a constant speed instead of being stationary. In chemistry, phlogiston theory was intuitive in part because it is natural to think that when an object burns something leaves the object. It is surprising to learn that something enters the object, namely, oxygen. In biology, it is natural to think that offspring will have a blend of their parents' traits. For example, if one parent has blond hair and another has brown hair, it is intuitive to think that their offspring will have dishwater-blond hair or some other blend of blond and brown. It took Mendel's pea plant experiments and subsequent discoveries in genetics to find out that this is not how heredity works. Lewis Wolpert describes the nature of science as being 'unnatural' because it is so counter-intuitive: 'I would almost contend that if something fits in with common sense it almost certainly isn't science' (Wolpert [1992], p. 11). This is what makes genuine scientific discoveries at once so difficult and so exciting. If common sense were a good guide to reality, we would not need scientific investigation.

I propose that the eliminativists of the 80s and 90s ('old school eliminativists' hereafter) were on to something and that the issue deserves another look.

In this article, I offer a new argument for eliminating the mental category of belief from cognitive science. I am not arguing for the elimination of beliefs from everyday usage and I am not arguing for the elimination of all propositional attitudes as the old school eliminativists did. I restrict my argument here just to eliminating beliefs from cognitive science.

One objection to old school eliminativist arguments is that they wrongly treat folk psychology as a quasi-scientific theory used to explain and predict behaviour (Hannan [1993]; Wilkes [1993]).[1] According to this view, folk psychology does not treat beliefs and other propositional attitudes as discrete inner causes of behaviour, hence they are not candidates for elimination. One conclusion one might also draw from this is that it would be inappropriate to base cognitive science on the categories of folk psychology. While these arguments might undermine old school eliminativism, they are compatible with the sort of scientific eliminativism I am advocating. I argue that belief is not an appropriate category for cognitive science, but I have a very different set of reasons for this conclusion. An alternative response to arguments like Hannan's or Wilkes's is to argue that we sharpen folk psychology up into a scientific theory, or at least use some folk psychological categories as the basis for a scientific theory, even if folk psychology is not itself such a theory. This seems to be what Fodor has in mind when he argues for the vindication of 'commonsense psychology' (Fodor [1985], [1987]). My argument in this article is that this is still a bad idea in the case of belief. My argument avoids the problematic reductive and semantic strategies that were employed by old school eliminativists.[2] Instead, it is based on Wimsatt's ([1981]) concept of robustness and the related concept fragility. I start by reviewing the notions of robustness and fragility. Next, I review some assumptions about what constitutes a belief on the received view. Following that, I present a number of studies in psychology from the past forty years that provide evidence for the fragility of belief. These, I argue, make it a good candidate for elimination. Finally, I address a number of attempts to preserve belief as a scientific category.

2 Robustness and Fragility

From a common-sense perspective, how do you determine whether something is real? Imagine you are deep in the Sahara desert. You are out of water and

[1] This claim is distinct from the claim made by simulationists (Goldman [1989]; Gordon [1986]) that our ability to 'mind-read' or 'mentalize' is not explained by the existence of a tacit theory. In this article, I remain neutral about what the correct account of mind-reading might be.

[2] Lycan ([1988]) and Stich ([1996]) have pointed out that old school eliminativists arguments assume that a description theory of reference is correct and will not go through without that assumption.

would desperately like to get a drink. You see the shimmer of water in the distance. Is it an oasis where you can fill your canteen or is it just a mirage? In this scenario, where all you have are your senses, you know that you cannot rely on vision alone. To confirm if there really is water in the distance, you must go up to it and feel its cool wetness. Your strategy is to use multiple means of detection to determine whether or not the water is real. Suppose you want to determine if you have a glass of water in front of you or a glass of vodka. Since both water and vodka are clear liquids, you must smell the contents of the glass and take a sip. Again you use multiple means of detection to determine what is really the case.

Following on the work of Campbell ([1958]), Campbell and Fiske ([1959]), and Levins ([1966], [1968]), William Wimsatt ([1981]) developed a concept of robustness based on the idea of using multiple independent means of detection or measurement. According to Wimsatt ([1981], p. 126), robustness analysis involves the following procedures:

(1) Analyze a *variety* of *independent* derivation, identification, or measurement processes.

(2) Look for and analyze things that are *invariant* over or *identical* in the conclusions or results of these processes.

(3) Determine the scope of the processes across which they are invariant and the *conditions* on which their invariance depends.

(4) Analyze and explain any relevant *failures of invariance*.

Wimsatt intends his concept of robustness to apply to a number of different targets, including theories, concepts, laws, models, processes, and objects. According to Wimsatt, any target that is invariant under this analysis is robust.

Robustness analysis has come under some criticism (Justus [2012]; Odenbaugh and Alexandrova [2011]; Orzack and Sober [1993]; Plutynski [2006]; Woodward [2006]). While a lengthy discussion of the putative confirmatory effects of robustness analysis would fall outside of the scope of this article, it is worth noting one reason why the cases of robustness analysis I focus on hold the promise of being defensibly confirmatory. Woodward ([2006]) distinguishes between several types of robustness. I will focus on just two of them here. According to Woodward, inferential robustness involves applying multiple independent models, which make different and sometimes contrary idealizing assumptions, to a fixed data set in order to reach a conclusion about a hypothesis. Measurement robustness, on the other hand, consists of the agreement of multiple independent measurement or detection procedures. Most, if not all, of the criticisms of robustness analysis alluded to above are aimed at analyses of inferential robustness. My argument is about the ontological status

of entities and categories; hence, I rely on the much more defensible notion of measurement robustness. To the extent that an entity is robust, in this sense, we can have confidence that it is real.

2.1 A historical example of robustness

Wesley Salmon's ([1984]) discussion of Jean Perrin's argument for the existence of molecules is a helpful example of the relevant kind of robustness analysis. At the turn of the century, there was a great debate amongst scientists about the reality of molecules. Jean Perrin ([1913]) laid out a clear-cut argument in favour of their existence. The argument was based on the experimental determination of Avogadro's number, N—that is, the number of molecules in a mole of any substance. The thought was that if we could establish a way to actually count the number of molecules in a given quantity of substance, then we would, in effect, build a bridge between the microcosm and the macrocosm. Perrin used an ultramicroscope, which made it possible to observe particles with diameters as small as 5×10^{-3} microns. At this magnification the Brownian movement of colloidal particles could be viewed, and Perrin was able to base his determination of Avogadro's number on observations of the vertical distribution of these particles in suspension. Perrin would go on to develop two additional ways to determine Avogadro's number based on what was known about Brownian motion.

Subsequently, a number of independent experimental techniques were developed to determine Avogadro's number. Each of these techniques was significantly different from the others. Perrin counts thirteen different experimental techniques including those with a basis in Brownian movement, alpha decay, X-ray diffraction, blackbody radiation, and electrochemistry. All thirteen produced practically the same number. Perrin put particular emphasis on the number of independent means of detection that were used. Salmon ([1984]) quotes Perrin's account:

> Our wonder is aroused at the very remarkable agreement found between the values derived from the consideration of such widely different phenomena. Seeing that not only is the same magnitude obtained by each method when the conditions under which it is applied are as varied as much as possible, but that the numbers thus established also agree among themselves, without discrepancy, for all the methods employed, the real existence of the molecule is given a probability bordering on certainty. (Perrin [1913], pp. 215–6)

The basic argument is that it is incredibly unlikely that all of these independent means of measurement would produce such remarkable agreement about the number if it was not, in fact, the number of real molecules in the substances measured. The results of multiple independent means of experimental

detection of Avogadro's number were mostly invariant across a number of different substances. Thus, the evidence for Avogadro's number is robust. It was these arguments that convinced scientists who were sceptical of the existence of molecules that they were mistaken and that molecules were, in fact, real.

2.2 Fragility and elimination

Robustness analysis can also give us a useful eliminativist tool. There is a useful inverse of the concept of robustness: fragility. A theoretical entity is fragile if the results of multiple, independent, putatively reliable measures of that entity turn out to radically vary and this variation cannot be adequately explained away. If our theories posit an entity or process as real, then we should expect the detection of that entity, *ceteris paribus*, to be fairly robust. That is to say, one would expect the results of various methods of measurement or detection of the entity to produce invariant results. When the detection of an entity or category is robust or fragile, I will call it a robust or fragile entity. If an entity turns out to be sufficiently fragile, then we have reason to think the entity is not real.[3] If Perrin had gotten variant results across different experimental techniques, it would have been rational for his contemporaries to be much less confident in the existence of molecules. Furthermore, given that the existence of molecules was controversial at the time, it would have been reasonable to think that it is more likely that they do not exist than it would be to suppose that something had gone wrong with the experiments.

In his discussion of robustness, Wimsatt cites a series of papers on culture and personality theory as a good detailed example of a failure to meet the requirements of robustness (Shweder [1979a], [1979b], [1980]). Shweder presents evidence that global personality traits such as being introverted or extroverted are highly context dependent. The following passage is typical of the type of evidence he presents:

> Distinguishable qualities of character, e.g., 'autonomy' and 'ascendancy', typically show higher within method associations than parallel across method associations. For example, if 'autonomy' and 'ascendancy' are measured using two methods, a projective test (e.g., T.A.T.) and a clinical interview, 'autonomy' and 'ascendancy' will more positively correlate

[3] Leamer ([1983]) uses the term 'fragility' in a similar, but different way. Leamer's notion applies to inferential robustness, rather than measurement robustness. For Leamer, if multiple models lead to quite different conclusions, then inference to the conclusion is fragile and suspension of belief in that conclusion is recommended. I argue that measurement fragility gives us positive reason to think an entity does not exist. Measurement fragility is more like a rebutting defeater than an undercutting defeater in the sense of Pollock ([1986]).

> *within* the projective test data than 'autonomy' correlates with *itself* across the two methods. In general, features of personality measuring instruments (the clinical interview situation, the projective test stimulus and context) have been found to be more stable than the features of the people measured. (Shweder [1979a], p. 259)

So, instead of finding invariant results across different measures of 'autonomy', it is typical to find variable results. In this case, 'autonomy' turns out to be fragile, and so we do not have much confidence that there is such a personality trait, at least not as a global personality trait that generalizes across a number of contexts. This is essentially the argument Shweder makes. Most of the evidence he presents follows the same pattern. He takes the variability across different measures of personality traits to rebut their evidential support.

For example, on a number of hypothesized global trait dispositions like 'dependency', 'dominance', or 'friendliness', Shweder presents evidence of high variability across contexts. Shweder describes this variability as follows:

> The more assertive child at the breakfast table is not the more assertive child in the playroom. The child who seeks help more than others is not the one who is more inclined to seek physical nearness [. . .] The adult who is more hostile to a parent is not typically the adult who is more hostile to a boss, *nor* is he or she the one who is typically less hostile to a boss. Individual differences in one context do not predict individual differences in the other. Different situations, stimuli or domains seem to affect different people *differently*. (Shweder [1979a], p. 260)

When psychologists attempt to detect these global personality traits in a number of different circumstances, they get highly variable results. This is more evidence that these global personality traits are fragile. Shweder finds the lack of evidence for global traits surprising because they are a part of our common-sense understanding of psychology:

> The absence of impressive support for generalized or global traits of character is surprising. Most of us, social scientist and layperson alike, share certain intuitions or everyday personality theories which suggest that certain items or traits of behavior go together (e.g., 'smiles easily' and 'introduces himself to strangers'; 'gentle' and 'good-natured') or are opposed (e.g., 'aggressive' and 'friendly'; 'gregarious' and 'reserved') (see Brown [1965]; D'Andrade [1974]). Many personality psychologists and most laypersons interpret these 'everyday' personality theories or trait and type concepts as inductive generalizations. They are held to arise out of observational experience and accurately summarize or encode 'relative frequencies of joint occurrences of various personality attributes and behavioral dispositions in other persons' (Passini and Norman [1966]; also, see Brown [1965]). Recent evidence, however, challenges this view. (Shweder [1979a], p. 262)

The case that can be made for the fragility of global personality traits parallels the case I am trying to make for beliefs. Both beliefs and personality traits are taken for granted by the folk, but both turn out to be fragile when measured under experimental conditions. In fact, Campbell and Fiske ([1959]) argue that very few posits in the social sciences have 'convergent validity', a concept very much like Wimsatt's robustness. They argue that this is a major difference between the social and biological sciences. It is the source of many problems in the social sciences. Contra Shweder, we should expect that many folk psychological concepts would fail to have good evidential support.

The elimination of phlogiston can be understood in terms of fragility as well. According to phlogiston theory, combustible substances are combustible because they contain phlogiston and phlogiston is released when they burn. So, when a substance is burned, it should lose mass. This is precisely what was found when substances like wood were burned and their ashes weighed. However, when metals were burned the mass of the metal increased rather than decreased. These are radically variant results across two distinct attempts to detect phlogiston. One substance's mass increased and one substance's mass decreased. Phlogiston theory also predicts that, if a substance is burned in a bell jar, the volume of air in the jar should increase because the phlogiston is being released into the air. When wood is burned this is indeed what you find. However, when metal is burned, the volume of air in the jar decreases. Again, there is wild variation in results. This sort of fragility is what you would expect when trying to measure something that does not exist.

3 The Received View

Considering the central role folk psychology realists advocate for beliefs,[4] they have been surprisingly sparse in their account of what it is to believe something. On their view, beliefs can be individuated from other mental states by the causal role they play in the architecture of the mind.[5] One very clear and helpful description of the functional architecture of the mind in this tradition is one proposed by Shaun Nichols and Stephen Stich ([2003]). Figure 1 is a representation of this architecture.

[4] Following Botterill and Carruthers ([1999]), I mean to contrast those who argue that the categories of folk psychology are scientifically real with those such as Dennett ([1987]) who treat them as useful fictions for prediction. Fodor gives the following useful definition: 'I propose to say that someone is a *Realist* about propositional attitudes iff (i) he holds that there are mental states whose occurrences and interactions cause behaviour and do so, moreover, in ways that respect (at least to an approximation) the generalizations of common-sense belief/desire psychology; and (ii) he holds that these same causally efficacious mental states are also semantically evaluable' (Fodor [1985], p. 78).

[5] Ramsey ([2010]) has pointed out that the functional role of beliefs individuates them from desires and the other attitudes, but does not tell us their functional role qua representations.

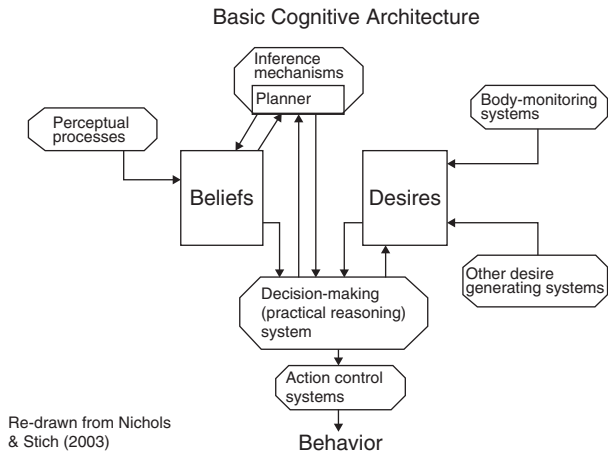

Figure 1. The basic architecture of the cognitive mind (Nichols and Stich [2003]) redrawn by Stephen Downes.

According to Nichols and Stich, to have a belief that *P* is to have a token representation with the content *P* in the 'belief box'. Some philosophers in this tradition require that these representations must be in a quasi-linguistic format (Botterill and Carruthers [1999]; Fodor [1975], [1987]; Sterelny [1990]). This is by no means universal: many prominent philosophers in this tradition, including Nichols and Stich, are not committed to quasi-linguistic representation and propose various alternatives (Braddon-Mitchell and Jackson [2007]; Dretske [1991]; Lewis [1994]; Nichols and Stich [2003]). The important feature to notice in Nichols and Stich's model, for my purposes, is that it has only one belief box. This means that if a mental state does not play the functional role indicated for the belief box, then it is not a belief. This severely restricts the number of sub-types of belief that are possible, if indeed folk psychology allows for any. There being just one belief box indicates an assumption that beliefs constitute a single, unified category of mental states. This is a nearly universal assumption amongst folk psychology realists. Keith Frankish ([2004]) has called this 'the unity of belief assumption'.

Another helpful account of folk psychology comes from Brian Loar ([1981]). Eric Schwitzgebel has helpfully summarized Loar's account into four important functional roles played by beliefs:

(1) Reflection on propositions (e.g., [*Q*] and [if *Q* then *P*]) from which *P* straightforwardly follows, if one believes those propositions, typically causes the belief that *P*.

(2) Directing perceptual attention to the perceptible properties of things, events, or states of affairs, in conditions favorable to accurate perception, typically causes the belief that those things, events, or states of affairs have those properties (e.g., visually attending to a red shirt in good viewing conditions will typically cause the belief that the shirt is red).

(3) Believing that performing action A would lead to event or state of affairs E, conjoined with a desire for E and no overriding contrary desire, will typically cause an intention to do A.

(4) Believing that *P*, in conditions favoring sincere expression of that belief, will typically lead to an assertion of *P*. (Schwitzgebel [2011]).

It is clear from this that there are two basic means of detecting what a person believes. Function (4) fits with the common-sense idea that if you want to know what someone believes, *ceteris paribus*, you should listen to what they say they believe. Function (3) suggests we can often use a person's non-verbal behaviour to determine what they believe. This is consistent with the common refrain that 'actions speak louder than words'. Of course, there are many different ways we might use these types of belief detection, but together they exhaust all possible means currently available for objectively detecting beliefs. I now turn to some research in social psychology that shows that if we treat beliefs as a unified category, then these methods of detecting beliefs produce wildly variable results.

4 Evidence for the Fragility of Belief

Research in psychology has consistently produced evidence that beliefs are fragile. That is to say, different means of detecting a person's beliefs vary to the point of contradiction. One classic example of this comes from Nisbett and Wilson ([1977]). They presented study participants with four pairs of identical nylon stockings laid out in a line. The participants, who were not told that the nylon stockings were identical, were asked which of the nylon stockings they liked best. The participants made their choices. When they were asked why they chose a particular pair, participants generally explained their choices in terms of some property or another such as colour, texture, or shape. The reasons given by the participants could not actually explain their choices because all of the nylon stockings were identical. Participants consistently chose the pair on the far right-hand side. There is no reason to think that the heuristic 'choose the one on the right' could make the nylon stockings on the right seem better in terms of its qualities. The heuristic was operating in the

absence of any distinguishing properties other than position relative to the other nylon stockings. It is likely that position determined the participants' choices and not features of the nylon stockings themselves.

This study naturally raises the question of whether study participants really believed the nylon stockings were different. According to Nisbett and Wilson, the subjects reported their reasons for making their selections sincerely. So, if we use what people verbally report as a good means of detecting their beliefs, as Loar's fourth functional individuating condition indicates it should be, we should think that they really did believe that the nylon stockings are different. However, their behaviour—picking the nylon stockings on the right—indicated that they did not believe the nylon stockings are different. Behaving as though *P* is, *ceteris paribus*, a good reason for thinking a person believes that *P* just as Loar's third functional individuating condition says it should be. If participants believed that the nylon stockings were all the same, it would have been natural to use a heuristic like, 'choose the one on the right'. In this case it is also worth noting that there is a tension in what belief the folk would attribute to the study participants based on their sensory perception on the one hand, and their non-verbal behaviour on the other. Since the nylon stockings are identical, one would think that they might be perceived this way. Following Loar's second condition, we would attribute the belief that the nylon stockings are the same to the study participants, but this is not what they verbally report. This is just one example of confabulating reasons in verbal reports for actions. Nisbett and Wilson detail a number of other examples. The same sort of confabulated reason-giving also occurs in other studies, such as Latane and Darley's ([1970]) work on the bystander effect. This sort of confabulation is not a rare anomaly. It appears to be quite common and widespread.

4.1 Contamination and fragility

More evidence for the fragility of belief comes from Paul Rozin's work on contamination. In a well-known study, participants were presented with a sweater, which they were told was owned by Adolf Hitler (Nemeroff and Rozin [1994]). Participants refused to wear the sweater, even after being told it had been thoroughly washed. This behaviour would indicate that the test participants believed that the sweater is tainted with evil and that they might be contaminated by it. If asked if this is what they really believe, at least some participants are likely to say no. In another study, participants are asked to eat fudge shaped to look like dog faeces, drink lemonade served in a sterilized bedpan, and to throw darts at a picture of a loved one (Rozin *et al.* [1986]). Not surprisingly, participants were reluctant to do so. Their behaviour would indicate that they believe that the fudge is dog faeces, the

lemonade is urine, and that they will hurt their loved one by throwing darts at her photo. However, when asked, these study participants will acknowledge that none of these actions will be harmful to them or their loved ones. In the latter experiment, the explanation might be that the dog faeces-shaped fudge, for example, just made study participants feel 'icky' and that is why they were unwilling to eat it. However, a natural explanation for why they feel icky is that they believe, at some level, that the fudge really is dog faeces. This sort of explanation has even more intuitive appeal for the study that uses the sweater. There is no inherent reason that a sweater should make someone feel icky, not in the same way that the appearance of dog faeces does. The idea that the study participants believe,[6] at some level, that the sweater is contaminated by Hitler's evil is a compelling explanation. In both cases, participants' behaviour seems to indicate one belief and their verbal reports indicate a contrary belief.

4.2 Implicit association tests and fragility

Over the last two decades, researchers in social psychology have been developing techniques to implicitly measure attitudes that respondents may not be willing to directly report or may not even be aware of having. One of the most interesting and fruitful techniques they use is the implicit association test or IAT (Lane *et al.* [2007]). The first use of the IAT revealed a disturbing amount of implicit racial bias amongst undergraduate students (Greenwald *et al.* [1998]). Using the IAT in conjunction with self-report methods, researchers have found that people who sincerely profess egalitarian views often display implicit biases (Lane *et al.* [2007]), and that these biases influence judgement and behaviour in a number of ways (Greenwald *et al.* [2009]).

Participants are first asked about their explicit attitudes under conditions of complete anonymity. For example they are asked, 'What best describes you?':

- I (strongly, moderately, or slightly) prefer European Americans to African Americans.
- I like European Americans and African Americans equally.
- I (strongly, moderately, or slightly) prefer African Americans to European Americans.

Next, participants complete the IAT. They are asked to classify words or images into one of two categories as quickly as they can while trying to remain as accurate as they can. In the first test they press a button if the image or word is 'European American or good' and another button if it is 'African American or bad'. The 'good' words are joy, love, peace, happy, and

[6] I address below the worry that the non-conscious cause of participants' aversion to wearing the sweater is something other than a belief (see especially Section 5.1).

so on; the 'bad' words are horrible, evil, hurt, failure, and so on. The images are of the faces of European Americans or African Americans. On the second test the pairings are reversed; 'good' and 'African American' are paired, while 'bad' and 'European American' are paired. Most participants take longer to respond on the second test than they do on the first test. This reveals an implicit racial bias in these participants.[7] However, most participants give egalitarian answers like 'I like European Americans and African Americans equally' on the explicit part of the test.

The studies discussed in this section show that what participants' self-reports indicate about their beliefs are systematically in conflict with what their non-verbal behaviour indicates about their beliefs. In a review of multiple IATs, Lane *et al*. wrote the following:

> The second clear message from the web data is that patterns of cognitions can vary widely across implicit and explicit measures. For example, although participants showed strong implicit preference for White over Black (average Cohen's *d* of two race attitude task = 0.80), the effect of their self-reported bias was much weaker (average Cohen's *d* of two race tasks = 0.31). (Lane *et al*. [2007], p. 66)

This data ranges over a number of IATs about not only race, but also about anything from gender to computer brands. This is not just an uncommon anomaly; it occurs across a wide range of socially salient categories.

I take it as uncontroversial that verbal self-report is a *ceteris paribus* reliable means of detecting a participant's beliefs. Remember, self-report was one of Brian Loar's four individuating functions for belief. In addition, there are other philosophical accounts of belief according to which it is partially constitutive of a belief that *P* to self-ascribe the belief that *P* (Heil [1988]; Shoemaker [1996]). If it is also reasonable to treat the IAT and other non-verbal behaviour as another means of detecting beliefs, then the data presented above constitutes good evidence that belief is a fragile category and a candidate for elimination. Rather than getting invariant results across independent methods of detection, as one would expect if belief were a robust category, researchers find a wide variance in results between self-report and implicit measures. In the case of race, test participants are consistently reporting egalitarian beliefs while their results on the IAT indicate that they do not have egalitarian beliefs. This is the widest possible variance across measures, and thus is evidence for the fragility of beliefs.

[7] For an assessment of the reliability of the IAT in detecting implicit attitudes, see (Lane *et al*. [2007]).

5 Attempts to Preserve the Belief Category for Cognitive Science

One attempt to preserve beliefs is to argue that the non-verbal behaviours in these studies do not really detect beliefs. However, the most natural candidate for detecting belief, other than a person's verbal behaviour (self-report either vocally, in writing, or answering a questionnaire), is non-verbal behaviour. The intuitive nature of this is captured by the common refrain that 'actions speak louder than words'. If we take the idea that folk psychological realists are basing the ontology of cognitive psychology on the common sense categories seriously, then it should be fair game to rely on folk intuitions about the boundaries of those categories. From this perspective, the idea that a person's non-verbal behaviour is a reliable guide to determining what they believe is completely reasonable. In addition, it is part of the received view that non-verbal behaviour is a good guide to a person's beliefs. We saw this in Brian Loar's third individuating function (see Section 3). Despite the fact that IAT participants, for example, seem unaware of the thought processes that lead them to their responses, they are behaving in a way that is consistent with having racist beliefs. Likewise, though they seem to be unaware of it, participants in Paul Rozin's study behave as though they believe evil can be transmitted through clothing, and participants in Nisbett and Wilson's study behave as though they believe that the nylon stockings on the right are preferable. At least some of the folk would judge that, deep down, these study participants really believe that the nylon stockings on the right are better, evil can be transmitted through clothing, or that African Americans are less intelligent or inferior in some way. In what follows, I focus on objections to the idea that the IAT detects beliefs. My responses to these objections can be applied *mutatis mutandis* to objections against the other two studies cited.

One objection to the idea that IATs are detecting beliefs is to argue that they are measuring preferences or attitudes, and these are not the same thing as beliefs. Baumeister and Bushman make the distinction between beliefs and attitudes in the following way:

> Attitudes differ from beliefs. Beliefs are pieces of information about something, facts or opinions. Attitudes are global evaluations toward some object or issue [. . .] Logically, attitudes are for *choosing*, whereas beliefs are for *explaining*. (Baumeister and Bushman [2008], p. 266, emphasis added)

This is a reasonable distinction, but it does not defeat the idea that IATs measure beliefs. First, this distinction does not track the way these terms are parsed in philosophy by those who advocate the received view. Usually, philosophers treat both states as propositional attitudes. Second, from a folk perspective, discovering what a person's attitudes are also tells you something about what they believe. It is natural to suppose that a person with racist

attitudes has them because of their beliefs about African Americans. Perhaps they wrongly believe that African Americans have lower IQs than Caucasian Americans or are more genetically predisposed towards violent crime. These beliefs would explain their racist attitudes. So, even if the attitudes measured by the IAT are pro-attitudes rather than belief attitudes, these attitudes serve as an indirect means of detecting beliefs via inference.[8]

5.1 Beliefs and aliefs

Another objection along these lines is to argue that the IAT does not detect beliefs; rather, it detects a new and distinct type of mental state. Tamar Gendler ([2008a], [2008b]) has proposed that we should posit two distinct types of mental states to explain the conflict between self-report and non-verbal behaviour. Her new theoretical posit is the 'alief', which she defines as follows: 'To have an alief is, to a reasonable approximation, to have an innate or habitual propensity to respond to an apparent stimulus in a particular way' (Gendler [2008a], p. 557). On Gendler's view, aliefs are most often not accessible to consciousness and are not responsive to reasons and evidence. By contrast, beliefs are accessible to consciousness and responsive to reasons and evidence. So, Gendler's account would explain what participants report on the IAT by saying that the participants really do believe what they report, namely, that European Americans and African Americans are equal. On her account, the participants' behaviour on the IAT is explained by their 'alief' that European Americans and African Americans are not equal.

Gendler is right to propose multiple mental states in an attempt to explain these results, rather than assuming a single unified category of belief as the received view does. However, Gendler has not made the case that the introspectively accessible states ought to be privileged with the label 'belief'. Gendler emphasizes the role of introspective awareness and responsiveness to reasons and evidence as hallmarks of belief. On her view, aliefs are not typically responsive to reasons and evidence in this way. I would argue that this distinction is overstated. There are instances when consciously accessible mental states are unresponsive to evidence and there are instances when implicit states are responsive to evidence.

There is substantial evidence from psychology that our beliefs are often not the result of a dispassionate, deliberative process. Rather, we have any number

[8] The difference between pro-attitudes like preference and belief is not as large as one might think. Sterelny ([2003]) has argued that we can trade talk of instrumental preferences into talk of beliefs that connect parts of a person's action to their ultimate preference. In other words, many preferences can just as easily be interpreted as beliefs. See his example of Alice, a thirsty woman in the Australian Bush (Sterelny [2003], p. 87).

of cognitive biases. Gilovich ([1993]) and Kahnemann ([2012]) provide nice summaries of these biases. Furthermore, there is the effect of belief perseverance. In one well-known study (Ross *et al*. [1975]), participants were given a number of real and fake suicide notes and asked to identify the real ones. By random choice, some participants were told that they correctly identified twenty-four out of twenty-five notes and other participants were told that they had correctly identified only ten out of twenty-five notes. At the end of the study both groups were told that the accuracy feedback they were given was completely bogus. Despite this, the participants who had received success feedback thought they were more accurate than others on the current test, and that they would be more accurate on future tests. This is clear evidence that what Gendler calls 'beliefs' are unresponsive to evidence. There is also good evidence that implicit biases can be changed in response to evidence. Stewart and Payne ([2008]) provide evidence that inculcating specific habits of thought can lower the effects of implicit biases and automatic stereotypes. This would indicate that some 'aliefs' are in fact responsive to evidence.[9]

Another problem with the distinction is that it undermines the usefulness of positing beliefs in cognitive science. Beliefs are meant to explain and predict behaviour. But beliefs fail to do this whenever there is a conflict with so-called 'aliefs'. If it is 'aliefs' and not beliefs that are explaining behaviour in these cases, it would seem that 'aliefs' are playing the explanatory role ordinarily assigned to beliefs by common sense. In fact, Fred Dretske might argue that, if anything, Gendler's 'aliefs' are the real beliefs in this case. He writes,

> If a structure's semantic character is unrelated to the job it does in shaping output, then this structure, though it may *be* a representation, is not a belief. A satisfactory model of belief should reveal the way in which *what we believe* helps determine *what we do*' (Dretske [1991], p. 79, emphasis original).

It is the 'aliefs' that determine what we do in these cases. The semantic content of the representation Gendler labels 'belief' does not. Gendler is right to posit multiple kinds of mental states, but she is wrong to hold onto the unity assumption about beliefs by privileging one kind of mental state with that label.

5.2 Contradictory beliefs

Folk psychology realists might accept the idea that the conflict between self-report and behaviour on the IAT could be explained by arguing that a person can simply hold contradictory beliefs. They might claim that the IAT

[9] I am using the term 'evidence' loosely here to mean something like 'responsive to information contrary to the attitude currently held'.

participants both believe that blacks are equal and that they are not equal. This would alleviate the tension we see between what study participants explicitly report and what their non-verbal behaviour indicates about their behaviour. If folk psychology allows for contradictory beliefs, then measuring or detecting contradictory beliefs would not count as a highly variant result because this is what a theory based on folk psychology would predict you should occasionally see. In this case, we would not have evidence that belief is a fragile category. However, this approach has a number of problems. First, folk psychology does not really seem to allow for contradictory beliefs. There is an assumption of rationality when we attribute beliefs to others. Second, maintaining that we can hold contradictory beliefs makes explanation and prediction more difficult. Finally, it is not clear how it would be possible for an individual to hold contradictory beliefs.

Daniel Dennett has plausibly argued that there is an assumption of rationality when we adopt the 'intentional stance':

> The presumption of rationality is so strongly entrenched in our inference habits that when our predictions prove false, we at first cast about for adjustments in the information-possession conditions (he must not have heard, he must not know English, he must not have seen x, been aware that y, etc.) or goal weightings, before questioning the rationality of the system as a whole. (Dennett [1978], pp. 9–10

Attributing contradictory beliefs is at best a sort of last resort for the folk. In fact, there is preliminary empirical evidence that we behave just as Dennett has suggested. Joanna Korman and Bertram Malle ([2012]) presented subjects with intentional actions paired with puzzling explanations for those actions. For example, the study participant sees a sentence like the following: 'The accountant got his books in order <action> because he wanted the ice cream truck to pass on the street <reason>'. Subjects are then asked to add information to make sense of these incongruent sentences. The following is an example of one participant's response: 'The accountant is very good with money <causal history>, and he wanted to know how much money he had in his account to spend on ice cream <desire>, because he realized he had purchased some big-ticket items at the mall the previous week <belief>'. Participants in this study were more likely to try to make sense of these sentences by adding belief–desire reason explanations than they were to add more straightforwardly causal explanations. Although participants were not presented with contradictions, this study seems to confirm that the folk work with a strong rationality assumption when giving explanations of actions.

Holding on to the idea of contradictory beliefs also creates confusion, rather than helping us generate good explanations. Imagine there is a social science professor, Sarah. Sarah has argued in print that races are not biologically real and has worked as a civil rights activist. Sarah also happens to be an

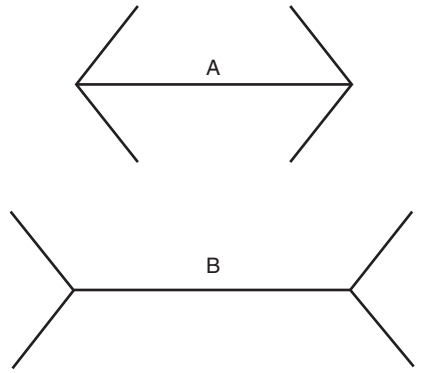

Figure 2. The Muller-Lyer illusion.

admissions officer at her university. After a routine check, it is discovered that Sarah has been disproportionately denying admissions to students with names that are more typical of African Americans, such as DeShawn or LaTonya. It would be confusing to explain this behaviour by saying that Sarah both believes African Americans are intellectually equal and that she believes that African Americans are intellectually inferior.

Even if the folk might allow for contradictory beliefs, we would immediately want to know how it is possible for a person to hold contradictory beliefs. The folk psychology realist might say that different mental modules contain contradictory representations. One example of this is the Müller-Lyer illusion (see Figure 2).

This illusion and many other optical illusions give rise to contradictory representations. Presumably people represent line A as being longer than line B in our visual cortex.[10] However, once a person becomes convinced about the fact that the lines are the same length, either by being told by a trusted source or using a ruler to measure them herself, another module represents them as being equal in length. Despite this fact, she still sees the lines as unequal because her visual cortex continues to represent A as longer than B no matter how many times she re-measures the lines.

Jerry Fodor ([1983]) argues that this is the case because our visual system is modular. An important feature of modules for Fodor is that they are informationally encapsulated, that is to say that modules are immune to the effects of information that is not directly associated with the domain that module operates in. Information encapsulation would explain how a person could simultaneously hold contradictory representations. However, as Eric

[10] Interestingly, this is apparently only true of people from WEIRD (western, educated, industrialized, rich, and democratic) societies (see Henrich *et al.* [2010]).

Schwitzgebel ([2010]) has pointed out, just having a representation is not sufficient to be a belief. A belief, on the received view, is a representation that plays a particular functional role. On Nichols and Stich's ([2003]) view, it must be a representation in the belief box. It seems that the mental state—call it an automatic representation—that explains the participant's behaviour on the IAT is functionally different than the mental state—call it the 'control representation'—that explains their self-reports. For example, the automatic representation often might skip the 'decision-making system' and directly affect the 'action-control system' in Nichols and Stich's model. The inputs that affect automatic representations would be different, at least some of the time, than the inputs that affect control representations. I do not yet have the details of this proposed architecture worked out. What seems clear, however, is that there is a need to posit more than one type of mental state to adequately explain these cases, and there is not a good reason to privilege one of these types of mental state as the 'real belief' and the other as something else as Gendler has attempted to do. The problem for the received view is that it tries to maintain the assumption that beliefs are a unified category of mental states.

5.3 In-between beliefs and the unity assumption

Despite the need for multiple kinds of mental states to explain the results of IATs, the folk want to attribute beliefs to the whole person rather than just a sub-system or module; Eric Schwitzgebel ([2010]), Robert Kurzban ([2010]), and Keith Frankish ([2004]) have all pointed this out. Beliefs, on the folk view, seem to be a property of a person, rather than of parts of the brain or functionally defined sub-systems, and beliefs are treated as a single unified category.

In the case of IATs, we are tempted to say that either the person 'really believes' that African Americans are not equal or that he/she 'really believes' that African Americans are equal. This is one of the intuitions that lie behind treating beliefs as a unified category. This is also the intuition that leads Schwitzgebel to propose these IAT cases are 'in-between beliefs'. That is to say that the IAT participants do not determinately believe that African Americans are equal nor do they determinately believe that they are not equal. The participants' beliefs are somewhere in-between. Schwitzgebel's view suffers from the same problem as supposing that we just have contradictory beliefs. Remember Sarah, the admissions officer. On Schwitzgebel's account, we would have to say that Sarah sort of believes that African Americans are intellectually equal and sort of believes that they are not. This would seem to add little to our understanding of why Sarah, despite being such a champion for civil rights, shows a bias in her admissions practices. It gives us no clue as to how Sarah might go about changing her behaviour and it does not seem to help us predict this kind of behaviour.

Schwitzgebel ([2010]) argues that we ought to tolerate this ambiguity in the concept in order to highlight the empirical fact that there is a gulf between our explicit avowals and our implicit attitudes. This is a laudable goal, but we can do this more effectively by letting go of the folk concept of belief altogether and positing more than one distinct type of mental state. I have suggested that we might posit 'automatic representations' and 'control representations'. The idea is to line mental state types up with something like Stanovich's ([1999]) system 1 and system 2, or possibly roughly like Fodor's ([1983]) central systems and input systems. We might explain Sarah the admission officer's behaviour by supposing that she has an automatic representation of African Americans as intellectually inferior and a control representation of African Americans as intellectually equal. We know that system 1 will tend to operate when a person is tired or cognitively depleted (Kahneman [2012]). This means that if Sarah does her admissions work late in the day or after a cognitively demanding task, her automatic representation of African Americans is more likely to determine her behaviour than her control representation that they are intellectually equal. Again, I am clearly not offering a fully worked out mental architecture here. The explanation sketch I offer above may not be accurate. The point is to show the explanatory advantages and theoretical possibilities that become available when beliefs are eliminated.

5.4 Belief sub-classes

One complaint might be that I have presented folk psychology in an unfair manner. According to this complaint, folk psychology is much richer, more flexible, and has the resources to explain the results of the experiments I have cited. One possibility along these lines is to argue that we can propose a number of belief sub-classes while keeping the over-arching category of belief. In this way beliefs are not eliminated and we get the diversity of categories we need to explain the relevant phenomena. For example, one might explain the IAT results discussed above by supposing that the study participants have a belief$_1$ that is egalitarian and a belief$_2$ that is biased. The reader is invited here to substitute his/her favourite belief sub-type labels, such as 'implicit belief' and 'explicit belief' in for belief$_1$ and belief$_2$. On this view, each of these mental states would count as a *bona fide* belief, but they are allowed to have distinct properties so long as there is still something that unifies them into one category. However, the folk conception of belief does not seem very tolerant of belief sub-classes. The folk response to such an explanation is likely something like, 'yes, but what do they really believe?'. As I noted above, one of the intuitions that led Schwitzgebel to propose his in-between beliefs view is that beliefs are properties of whole persons. The folk would want to know what the study participants in the IAT believe *simpliciter*. This is also what has

lead Gendler to argue that participants' verbal reports tell us what they 'really believe', while their non-verbal behaviour is explained by some other category of mental state such as 'aliefs'. If belief$_1$ and belief$_2$ are contradictory, there is no role for the over-arching category of belief to play in answering the folk demand for an attribution of belief to the whole person. In order for this view to be distinct from the contradictory beliefs view discussed above, there must be some useful explanatory or predictive role for the over-arching category of belief.

At first glance, the received view seems like it might give the over-arching category this role. Remember, on the received view, beliefs are individuated functionally. On Nichols and Stich's ([2003]) account, there is only one belief box. This means that anything that counts as a belief must play this functional role. However, there is a need for more than one type of mental state to explain the results of the empirical studies I detailed above. In order for the idea that there are belief sub-classes to work here, the sub-classes need to play functional roles that are distinct enough to provide the explanations needed while simultaneously sharing a common functional role. Beliefs as described in Nichols and Stich's model do not succeed in pulling off this trick. For example, mental states that are automatically activated and account for participants responses on the IAT, say belief$_1$, would bypass the decision making/practical reasoning box in their model altogether. Amending the model and making the path through the practical reasoning box optional could remedy this. However, the more amendments of this sort that are made just to preserve the unity of the category, the less useful it becomes in explanation and prediction.

In fact, given the diversity of properties that belief-like mental states have, it is unclear whether there are any explanatorily useful properties the over-arching belief category would retain. A list of properties that beliefs might be said to have include: conscious, non-conscious, pre-conscious, occurrent, non-occurrent, automatic, controlled, passively acquired, and actively required. If some of these properties tend to correlate with one another, we might want to form categories around these correlations. This is precisely what Kim Sterelny ([2003]) and Keith Stanovich ([2011]) do; I discuss their approach in the next section. We might treat these as sub-categories of belief, but there is no compelling reason to do this unless the over-arching category plays some useful explanatory role. As Frankish ([2004]) points out, most of the interesting questions about beliefs in cognitive science occur at the level of these supposed sub-types. While an over-arching category of belief might be useful for everyday purposes, it is hard to see what useful role it could play in cognitive science. All of the explanatory work in cognitive science is done at a finer grain of analysis. It is at this finer grain of analysis that philosophical debates about belief occur.

Edouard Machery has developed a helpful approach to determining whether an over-arching category with sub-classes that are natural kinds should count as a natural kind itself. An over-arching category, *K*, fails to pick out a natural kind when the following condition obtains:

> There are very few generalizations that are true of *K*s, besides the properties that are used to identify *K*s. At the same time, many generalizations are true of members K—$K_1, \ldots, K_n$. (Machery [2009], pp. 237–8)

Machery gives 'memory' as a good example of a term that fails to pick out a natural kind. He writes:

> This term has been *replaced* by several theoretical terms, such as 'working memory', 'long-term memory', 'declarative memory', 'procedural memory', 'episodic memory', 'implicit memory', and 'explicit memory'. While 'memory' is not believed anymore to pick out a natural kind, each of the replacing theoretical terms is fruitfully used to formulate psychological generalizations. (Machery [2009], p. 237, emphasis original)

Machery argues that theoretical terms that do not pick out natural kinds ought to be eliminated from the relevant science. We should not be fooled into thinking that there has been a smooth reduction of the folk concept of 'memory' to a useful scientific category just because the term 'memory' has been retained. According to Machery, there are not any generalizations that are true of 'memory' that are not already true of the kinds picked out by his list of theoretical terms. In other words, the over-arching folk concept of memory does no useful scientific work. Despite their names, the theoretical terms in Machery's list should not be seen as sub-types in a scientific context. They are distinct theoretical kinds. Because using the word 'memory' is misleading in this way, it was probably a mistake to retain use of it in the scientific context.

The most likely kinds of generalizations we can make about beliefs are the very same generalizations that characterize their functional roles. For example, '[b]elieving that performing action *A* would lead to event or state of affairs *E*, conjoined with a desire for *E* and no over-riding contrary desire, will typically cause an intention to do *A*' (Schwitgebel [2011]). Another example would be, 'believing that *P*, in conditions favoring sincere expression of that belief, will typically lead to an assertion of *P*' (Schwitgebel [2011]). Both of these generalizations are paraphrases of the functional roles Brian Loar ([1981]) says individuate beliefs from other mental states. It is difficult to see what other generalizations can be made that are general enough to cover an over-arching category of beliefs, rather than just putative belief sub-classes. By Machery's lights, the over-arching category of belief would not count as a natural kind. Because it is not a natural kind, he might argue that beliefs

ought to be eliminated from cognitive science just as he did for concepts. I do not advocate couching this argument in terms of natural kinds as Machery does. But I do think that his condition is a good guide to whether an over-arching category is scientifically helpful. Trying to preserve an over-arching category of belief is not scientifically helpful because it does not help us generate any useful generalizations that aren't already true of so-called sub-types, such as automatic representations and control representations, or the categories of mental state suggested by Stanovich and Sterelny below.

Remember, the claim here is not that belief should be eliminated from everyday use. The claim is that it should be eliminated from cognitive science. Consider the following analogy: Someone might argue that just because there are trees that have flowers and trees that do not have flowers does not mean that we should eliminate the over-arching category of trees. This is right from the perspective of the folk, but it is wrong from the perspective of science. There is no biological taxon that includes all and only trees. There are, however, gymnosperms and angiosperms. We could try to save the tree category by creating *ad hoc* sub-types like gymnosperm tree and angiosperm tree. However, there is no compelling scientific reason to do so. If we want to truly understand plant life, we do not desperately hold on to folk categories like trees. If we want to understand the mind, we should not try to preserve folk categories like belief at all costs.

5.5 Self-deception

Another possible resource the folk psychology realist might use to retain the category of beliefs for cognitive science is the idea of self-deception. It is a matter of deep philosophical controversy just how to characterize self-deception and there are even some philosophers, though in the minority, who argue that self-deception is not possible (Paluch [1967]; Haight [1980]). The claim would be that the cases I have presented from the psychological research cited above might be explained in terms of having false meta-cognitive beliefs. In what follows, I investigate this possibility using the race IAT as an example. I argue that if an intentionalist account of self-deception is correct, then the studies I cite in support of the fragility of belief are not cases of self-deception. If a non-intentionalist account is correct, then self-deception may occur in these cases; but this would only undermine my argument on an overly restrictive interpretation of what can be inferred from the study of participants' responses.

According to Deweese-Boyd ([2012]), traditional intentionalist accounts of self-deception require that a person self-deceives only when they know or believe that p and intentionally fool themselves into believing $\sim p$. In the case of the IAT, this would mean that participants would know or believe that

they do have a preference between African Americans and European Americans, but intentionally get themselves to believe that they do not have such a preference. However, the evidence indicates that IAT participants do not know about their bias and are not aware of any preferences they have regarding race. Participants regularly express surprise and dismay when their IAT results are revealed. If it is right to suppose that self-deception requires intent, then IATs are not cases of self-deception. The other studies I cite above in support of the fragility of belief also involve consciously inaccessible mental states. Hence, if self-deception requires intentional deception, then these could not be cases of self-deception either.

Non-intentionalist accounts of self-deception only require that someone believe falsely or be mistaken in believing. This may seem like nothing more than making an honest mistake. However, according to Deweese-Boyd, non-intentionalists argue that the false belief is not accidental because it is motivated by a desire, anxiety, or some emotion that is relevant to the belief. So, on this view, the IAT participants really have unconscious racist beliefs such as 'African Americans and European Americans are not equal'. When explicitly asked about their race preferences, IAT participants might be motivated by a desire to think of themselves as egalitarian or have anxiety about being labelled a racist. This causes them to report a belief like, 'I have no preference between African Americans and European Americans'. The belief indicated by their behaviour on the IAT falsifies, but does not contradict, the belief they report on the pre-IAT survey. This is because the IAT belief is about African Americans and European Americans, and the pre-IAT belief is about the participants' own mental states. If this is right, then there isn't the kind of radical variance between means of detecting beliefs that I have argued is evidence for the fragility of belief. However, it is important to note here that this argument only works because we are restricting ourselves to interpreting the answers participants give on the pre-IAT survey as being exclusively meta-cognitive beliefs (beliefs about their own mental states). I argue that this overly restrictive interpretation is implausible.

In order to see why the above interpretation is overly restrictive and implausible, it is worth spending some time discussing how the IAT measures participants' beliefs. Recall, on the pre-IAT survey, participants are asked about their race preferences. They are asked, 'What best describes you?':

- I (strongly, moderately, or slightly) prefer European Americans to African Americans.
- I like European Americans and African Americans equally.
- I (strongly, moderately, or slightly) prefer African Americans to European Americans.

A plurality report, 'I like European Americans and African Americans equally'. The folk are likely to make one or both of the following inferences about this answer. The folk might infer that these participants believe, 'I have no preference between European Americans and African Americans'. However, they are also likely to infer that these study participants believe, 'African Americans and European Americans are equal'. A person's preferences are indicative of the beliefs they have.

The first belief that the folk infer is a meta-cognitive belief. The second belief is not a meta-cognitive belief. It is not a belief about the participants' other mental states. It is a first-order belief about the properties of African Americans and European Americans. For the self-deception defence of belief to succeed, it would have to be the case that the folk would rarely if ever attribute the first-order belief (African Americans and European Americans are equal), and only attribute the meta-cognitive belief (I believe that I have no preference between African Americans and European Americans) to the IAT participants. This is because the participants' having the first-order belief is evidence of fragility. The first-order belief directly contradicts the belief indicated by the participants' non-verbal behaviour on the IAT.

I do not deny that it is reasonable to suppose that self-deception of this kind may be occurring some of the time on the race IAT, but it is implausible to suppose that it is always or nearly always the case. It is also important to note that the race IAT is not the only IAT that shows a tension between verbal and non-verbal behaviour. Lane *et al.* ([2007]) report similar results in IATs about attitudes toward commercial brands such as Coca-Cola and Pepsi, and Mac and PC. It is even more implausible in these cases to suppose that IAT participants are always or mostly engaged in self-deception. Sometimes people identify with commercial brands enough to engage in motivated reasoning about them, but this is surely the exception and not the rule. I conclude that defending belief in terms of self-deception is implausible and does not undermine the argument for the fragility of belief.

6 Alternative Mental States

It is natural to ask what we might replace beliefs with if they are eliminated. I take the fact that this question is raised to be a virtue of eliminating beliefs. We unshackle ourselves from folk intuitions and allow ourselves to hypothesize types of mental states that will play the explanatory and predictive job that we require of them. Above, I floated the idea of there being automatic representations and deliberative representations just as an example. More detailed accounts of alternative belief-like mental states have been developed by Kim Sterelny ([2003]) and Keith Stanovich ([2011]). Each of these thinkers

proposes three different types of belief-like states. To give you the flavour of this kind of theorizing, I will very briefly describe Sterelny's proposed mental states.

Sterelny proposes detection systems, robust tracking systems, and decoupled representations. At the very least, every organism needs 'mechanisms that mediate a specific adaptive response to some feature (or features) of their environment by registering a specific environmental signal that tells the organism of the presence of that feature' (Sterelny [2003], p. 14). Sterelny describes this as a sort of baseline for thinking about cognition. More specifically, Sterelny writes:

> I shall call single-cued discriminatory mechanisms of this kind *detection systems.* Some are built-in. Other, perhaps, are learned by simple associative mechanisms. ([2013], p. 14, emphasis added)

Examples of detection systems include bacteria responding to chemical gradients in their environment, plants registering the seasons by monitoring day length, and fireflies flashing a species-specific code to which females respond.

For detection systems to be effective, according to Sterelny, the environment needs to be 'informationally transparent', that is, these are environments in which 'the signal indicating the presence of a specific resource is reliable and if the agent can use its sensory apparatus to discriminate that signal from other stimuli' (Sterelny [2003], p. 20). Biological hostility can degrade an organism's informational environment in a number of ways. Hostile agents degrade an organism's information environment by concealment and disguise. Sterelny proposes to call the capacity to use multiple cues (either built-in or learned) a 'robust tracking system'. These systems function just like detection systems in that they tightly couple cues from the environment with specific behavioural responses, but they are sensitive to multiple cues rather than just one.

The last category in Sterelny's taxonomy is the decoupled representation. These are 'internal states that track aspects of our world, but which do not have the function of controlling particular behavior' (Sterelny [2003], p. 29). According to Sterelny, decoupled representations allow maximally flexible behaviour and were probably selected for in environments that are 'informationally opaque'. These states are not tied to any specific behaviour and are potentially relevant to any number of task domains. Sterelny is willing to label decoupled representations as 'belief' as long as 'belief' just means decoupled representation and nothing more. This is a temptation I think he should resist, and I have argued above that this would be a mistake. This would be the same type of mistake that Gendler makes. There is no reason to privilege decoupled representations as being the 'real beliefs'.

Sterelny's detection systems and robust tracking systems are not unlike what Gendler calls 'aliefs'. These are the kind of mental states that might

account for the non-verbal behaviour exhibited by the study participants in the psychological research described above. Meanwhile, his decoupled representations would nicely explain study participant's verbal reports. For example, the images of African American faces in the race IAT might act as cues that trigger a detection system or robust tracking system automatically producing an adverse response. Meanwhile, when participants are explicitly asked about their attitudes toward race, they may verbally express the contents of the relevant decoupled representation. One can see how Sterelny's proposed mental states do a much better job of explaining the psychological phenomena detailed above. They also have the advantage of being plausible as types of mental states that might really have evolved in humans. The forgoing has been an example of the type of theorizing that is possible once we let go of the requirement to force folk categories on cognitive science.

7 Conclusion

The received view in philosophy of psychology is that cognitive science ought to posit the basic categories of folk psychology such as beliefs and desires as real mental states. On this view, beliefs are a unified category that is defined by its causal role in our mental architecture. If beliefs were real we would expect them to be robust. However, several psychological studies have revealed wildly varying results from verbal reports on the one hand, and non-verbal behaviours on the other. Participants report egalitarian beliefs, but their IAT performances indicate they have racially biased beliefs. Participants in Nisbett and Wilson's ([1977]) study report various features such as colour or texture as the reason they selected one set of nylon stockings over the others, yet their behaviour indicates they are simply picking the nylon stockings on the right. Participants in Nemeroff and Rozin's ([1994]) study report that they do not believe evil can somehow be magically transferred to clothing, but they are not willing to wear Hitler's sweater. I have argued that there is just as much reason to treat these non-verbal behaviours as means of detecting beliefs from the folk perspective as there is to treat verbal reports as means of detecting beliefs. This pervasive variation between means of belief detection is evidence for the fragility of beliefs, and thus for their elimination from cognitive science. Again, I want to emphasize that I am not suggesting that beliefs be eliminated from everyday discourse. The folk probably ought to take an anti-realist position like Dennett's ([1987]). I have given ample reason above why this would not be a good strategy for cognitive scientists. Maintaining belief as a unified category only adds confusion and makes explanation and prediction more difficult and less accurate.

Acknowledgements

I would like to thank Steve Downes and Jonah Schupbach for their tireless efforts in helping me to prepare this manuscript. I would also like to thank Ron Mallon, Matt Haber, Jim Tabery, Jay Odenbaugh, Paul Griffiths, Karola Stotz, and Brett Sherman for their helpful comments and criticisms at various stages of this project.

Department of Philosophy
University of Utah
Salt Lake City, UT 84112, USA
chris.jenson@utah.edu

References

Baumeister, R. and Bushman, B. [2008]: *Social Psychology: Human Nature*, Belmont: Thomson Wadsworth.

Braddon-Mitchell, D. and Jackson, F. [2007]: *Philosophy of Mind and Cognition*, Malden: Blackwell.

Botterill, G. and Carruthers, P. [1999]: *The Philosophy of Psychology*, Cambridge: Cambridge University Press.

Brown, R. [1965]: *Social Psychology*, New York: Free Press.

Campbell, D. T. [1958]: 'Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities', *Behavioral Science*, **3**, pp. 14–25.

Campbell, D. T. and Fiske, D. W. [1959]: 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix', *Psychological Bulletin*, **56**, pp. 81–105.

Churchland, P. M. [1981]: 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy*, **78**, pp. 67–90.

Churchland, P. S. [1986]: *Neurophilosophy*, Cambridge, MA: MIT Press.

D'Andrade, R. [1974]: 'Memory and the Assessment of Behavior', in T. Blalock (*ed.*), *Measurement in the Social Sciences*, Chicago: Aldine-Atherton.

Dennett, D. [1978]: *Brainstorms*, Cambridge, MA: MIT Press.

Dennett, D. [1987]: *The Intentional Stance*, Cambridge, MA: MIT Press.

Deweese-Boyd, I. [2012]: 'Self-deception', in E. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/spr2012/entries/self-deception/>.

Dretske, F. [1991]: *Explaining Behavior*, Cambridge, MA: MIT.

Fodor, J. A. [1975]: *The Language of Thought*, New York: Thomas Crowell.

Fodor, J. A. [1983]: *Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J. A. [1985]: 'Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum', *Mind*, **94**, pp. 76–100.

Fodor, J. A. [1987]: *Psychosemantics*, Cambridge, MA: MIT Press.

Frankish, K. [2004]: *Mind and Supermind*, Cambridge: Cambridge University Press.

Gendler, T. [2008a]: 'Alief in Action (and Reaction)', *Mind and Language*, **23**, pp. 552–85.

Gendler, T. [2008b]: 'Alief and Belief', *The Journal of Philosophy*, **105**, pp. 634–63.

Gilovich, T. [1993]: *How We Know What Isn't So*, New York: The Free Press.

Goldman, A. [1989]: 'Interpretation Psychologized', *Mind and Language*, **7**, pp. 104–19.
Gordon, R. [1986]: 'Folk Psychology as Simulation', *Mind and Language*, **1**, pp. 158–71.
Greenwald, A., McGhee, D. and Schwartz, J. [1998]: 'Measuring Individual Differences in Implicit Cognition: Implicit Association Test', *Journal of Personality and Social Psychology*, **74**, pp. 1464–80.
Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L. and Banaji, M. R. [2009]: 'Understanding and Using the Implicit Association Test, III: Meta-analysis of Predictive Validity', *Journal of Personality and Social Psychology*, **97**, pp. 17–41.
Haight, R. M. [1980]: *A Study of Self-deception*, Sussex: Harvester Wheatsheaf.
Hannan, B. [1993]: 'Don't Stop Believing: The Case against Eliminative Materialism', *Mind and Language*, **8**, pp. 165–79.
Heil, J. [1988]: 'Privileged Access', *Mind*, **97**, pp. 238–51.
Henrich, J., Heine, S. J. and Norenzayan, A. [2010]: 'The Weirdest People in the World?', *Behavioral and Brain Sciences*, **33**, pp. 61–135.
Justus, J. [2012]: 'The Elusive Basis of Inferential Robustness', *Philosophy of Science*, **79**, pp. 795–807.
Kahneman, D. [2012]: *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.
Korman, J. and Malle, B. [2012]: *Practical Rationality in Action Explanation: A Crucial Role for Belief Reasons*, presented at the Society for Philosophy and Psychology 2012 Meeting, University of Colorado, Boulder, Colorado.
Kurzban, R. [2010]: *Why Everyone (Else) is a Hypocrite: Evolution and the Modular Mind*, Princeton: Princeton University Press.
Lane, K. A., Banaji, M. R., Nosek, B. A. and Greenwald, A. [2007]: 'Understanding and Using the Implicit Association Test, IV: Procedures and Validity', in B. Wittenbrink and N. Schwarz (*eds*), *Implicit Measures of Attitudes: Procedures and Controversies*, New York: Guilford Press, pp. 59–102.
Latane, B. and Darley, J. M. [1970]: *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century Crofts.
Leamer, E. [1983]: 'Let's Take the Con out of Econometrics', *American Economic Review*, **73**, pp. 31–44.
Levins, R. [1966]: 'The Strategy of Model Building in Population Biology', *American Scientist*, **54**, pp. 421–31.
Levins, R. [1968]: *Evolution in Changing Environments: Some Theoretical Explorations*, Princeton: Princeton University Press.
Lewis, D. [1994]: 'Reduction of Mind', in S. Guttenplan (*ed.*), *A Companion to the Philosophy of Mind*, Oxford: Blackwell, pp. 411–31.
Loar, B. [1981]: *Mind and Meaning*, Cambridge: Cambridge University Press.
Lycan, W. [1988]: *Judgement and Justification*, Cambridge: Cambridge University Press.
Machery, E. [2009]: *Doing without Concepts*, New York: Oxford University Press.
Nemeroff, C. and Rozin, P. [1994]: 'The Contagion Concept in Adult Thinking in the United States: Transmission of Germs and Interpersonal Influence', *Ethos*, **22**, pp. 158–86.

Nisbett, R. and Wilson, T. [1977]: 'Telling More than We Can Know: Verbal Reports on Mental Processes', *Psychological Review*, **84**, pp. 231–59.
Nichols, S. and Stich, S. [2003]: *Mindreading*, Oxford: Oxford University Press.
Odenbaugh, J. and Alexandrova, A. [2011]: 'Buyer Beware: Robustness Analysis in Economics and Biology', *Biology and Philosophy*, **26**, pp. 757–71.
Orzack, S. and Sober, E. [1993]: 'A Critical Examination of Richard Levins' "The Strategy of Model Building in Population Biology" ', *Quarterly Review of Biology*, **68**, pp. 533–46.
Passini, F. T. and Norman, W. T. [1966]: 'A Universal Conception of Personality Structure?', *Journal of Personality and Social Psychology*, **4**, pp. 44–9.
Paluch, S. [1967]: 'Self-deception', *Inquiry*, **10**, pp. 268–78.
Perrin, J. [1913]: *Les Atomes*, Paris: Alcan.
Plutynski, A. [2006]: 'Strategies of Model Building in Population Genetics', *Philosophy of Science*, **73**, pp. 755–64.
Pollock, J. [1986]: *Contemporary Theories of Knowledge*, Towota: Roman and Littlefield.
Ramsey, W. [2010]: *Representation Reconsidered*, Cambridge: Cambridge University Press.
Ramsey, W., Stich, S. and Garon, J. [1990]: 'Connectionism, Eliminativism, and the Future of Folk Psychology', *Studies in Cognitive Studies*, **3**, pp. 117–44.
Ross, L., Lepper, M. R. and Hubbard, M. [1975]: 'Perseverance in Self-perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm', *Journal of Personality and Social Psychology*, **32**, pp. 880–92.
Rozin, P., Millman, L. and Nemeroff, C. [1986]: 'Operation of the Laws of Sympathetic Magic in Disgust and Other Domains', *Journal of Personality and Social Psychology*, **50**, pp. 703–12.
Salmon, W. C. [1984]: *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.
Schwitzgebel, E. [2010]: 'Acting Contrary to Our Professed Beliefs, or the Gulf between Occurrent Judgment and Dispositional Belief', *Pacific Philosophical Quarterly*, **91**, pp. 531–53.
Schwitzgebel, E. [2011]: 'Belief', in E. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/win2011/entries/belief/>.
Shoemaker, S. [1996]: *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press.
Shweder, R. A. [1979a]: 'Rethinking Culture and Personality Theory, Part I: A Critical Examination of Two More Classical Postulates', *Ethos*, **7**, pp. 255–78.
Shweder, R. A. [1979b]: 'Rethinking Culture and Personality Theory, Part II: A Critical Examination of Two More Classical Postulates', *Ethos*, **7**, pp. 279–311.
Shweder, R. A. [1980]: 'Rethinking Culture and Personality, Part III: From Genesis and Typology to Hermeneutics and Dynamics', *Ethos*, **8**, pp. 60–94.
Stanovich, K. [1999]: *Who is Rational?*, London: Lawrence Erlbaum Associates.
Stanovich, K. [2011]: *Rationality and the Reflective Mind*, New York: Oxford University Press.
Sterelny, K. [1990]: *The Representational Theory of Mind*, Cambridge: Basil Blackwell.

Sterelny, K. [2003]: *Thought in a Hostile World: The Evolution of Human Cognition*, Malden: Blackwell Publishing.

Stewart, B. and Payne, K. [2008]: 'Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control', *Personality and Social Psychology Bulletin*, **34**, pp. 1332–45.

Stich, S. [1983]: *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press.

Stich, S. [1996]: *Deconstructing the Mind*, New York: Oxford University Press.

Wilkes, K. [1993]: 'The Relationship between Scientific and Common Sense Psychology', in S. Christensen and D. Turner (*eds*), *Folk Psychology and the Philosophy of Mind*, Hillsdale: Lawrence Erlbaum, pp. 144–87.

Wimsatt, W. C. [1981]: 'Robustness, Reliability, and Overdetermination', in M. Brewer and B. Collins (*eds*), *Scientific Inquiry and the Social Sciences*, San Francisco: Jossey-Bass, pp. 124–63.

Wolpert, L. [1992]: *The Unnatural Nature of Science*, Cambridge, MA: Harvard University Press.

Woodward, J. [2006]: 'Some Varieties of Robustness', *Journal of Economic Methodology*, **13**, pp. 219–40.