Aaron Zimmerman

The Nature of Belief

***Abstract**: Neo-Cartesian approaches to belief place greater evidential weight on a subject's introspective judgments than do neo-behaviorist accounts. As a result, the two views differ on whether our absent-minded and weak-willed actions are guided by belief. I argue that simulationist accounts of the concept of belief are committed to neo-Cartesianism, and, though the conceptual and empirical issues that arise are inextricably intertwined, I discuss experimental results that should point theory-theorists in that direction as well. Belief is even less closely connected to behaviour than most contemporary functionalists allow.*

1. The Concept of Belief

'Belief' cannot be defined. That is, barring trivial examples, the concept of belief is not equivalent to the content of any definite description. As with 'knows' we begin to doubt the possibility of a fruitful analysis by looking at the bevy of unsuccessful attempts. The steady stream of counter-examples casts doubt on our supplying anything substantive that is even extensionally equivalent to 'belief'; and as the analyses not subject to obvious refutation gain complexity, the claim of synonymy or even some looser form of cognitive equivalence grows ever more implausible.[1]

Can we nevertheless give an account of belief? Yes. We can say what belief *is* by identifying its properties and the relations in which it is involved. Some of these properties may turn out to be essential so

Correspondence:
Aaron Z. Zimmerman, Department of Philosophy, University of California, Santa Barbara, CA 93106, USA. *Email: azimmerman@philosophy.ucsb.edu*

[1] Jerry Fodor (1998) argues for an 'atomistic' conception of almost every concept expressed by a simple predicate. Cf. Timothy Williamson's (2000) case against analyses of the concept expressed by 'knows' when arguing for a 'knowledge first' account of epistemology.

that beliefs could not exist without having them. Some will be entirely accidental. Many will play an important epistemic role in our efforts to determine who believes what. All are potentially relevant to the task of better understanding belief given the actual state of things and minds.

Still, even if we abandon the aim of providing an analysis of the concept of belief, the nature of this concept must be grappled with before we can get a satisfactory grip on beliefs themselves, for it is at least possible that different theories of the concept will turn out to yield substantively different approaches to the metaphysics. I begin, therefore, by considering the *theory-theory* account of mental state attribution according to which grasp of the concept of belief consists in knowledge of a sufficient portion of common sense psychology. The theory-theory leaves open two substantially different metaphysical accounts of belief: these might with some justification be called *neo-Cartesian* and *neo-behaviorist* theories respectively. I describe experimental evidence that the theory-theorist must acknowledge as relevant to adjudicating between these two accounts, and suggest that the current state of inquiry provides tentative support for the neo-Cartesian approach. In the course of this discussion I address the chief rival to the theory-theoretical account of our concept of belief: *simulation theory*, and argue that it already commits us to neo-Cartesianism. Thus, at the end of the day, I think that the best current theories about our concept of belief and the best current theories about the biological realization of our beliefs favour a neo-Cartesian metaphysics of belief over its more behaviourist competitor. I close by considering how neo-Cartesianism can be best reconciled with the functionalist perspective on propositional attitudes that most theorists currently adopt.

2. Making Folk Psychology Explicit

Our concept of belief is not an overtly technical one in at least the following sense. History shows no record of 'belief' being introduced into the lexicon by a band of proto-psychologists in the process of elaborating a highly predicative and subsequently well-confirmed theory. If we had to learn a theory in order to adequately understand 'belief' this is not an event we can remember. If we instead acquired the concept via communication with someone who developed the theory (or someone who spoke to someone who spoke to someone . . . who developed the theory) this expert is no longer around for questioning.

There are nevertheless three possibilities left open for those who follow David Lewis (1972) in thinking of 'belief' as a theoretical term. (1) Our knowledge of folk psychology is the product of

evolution. Our grasp of its axioms is largely owed to structures that persist in the human genome because of the mechanisms of natural selection (see Carruthers, 1996). (2) Our knowledge of the theory is more the product of developmental processes. Children use general principles of reasoning to posit the principles of folk psychology in the course of explaining the behaviour of those animals (human and otherwise) they encounter during maturation (see Gopnik and Wellman, 1992, pp. 167–8). (3) Ancient humans (using general principles of reasoning) developed folk psychology at the time when the original ancestor of 'belief' was introduced into a public language. One acquires a second-hand familiarity with the theory when learning one's native tongue by deferring to the experts who introduced the expression.

In any event, if it exists, the psychological theory through which the concept of belief is introduced is a deeply tacit one. We must therefore look to common assumptions about belief reflected in our naïve use of 'belief' to achieve any measure of success in the theory's articulation.

Suppose I see S with her eyes open, facing a single, red apple in a well-lit room, and that I have no reason to think that she distrusts her senses. Surely I will assume that S believes that there is something red in front of her. The more general assumption in play is that people believe those relevant truths to which they have access. If the apple is taken away, I will assume that S no longer believes that something red is in front of her. Here the more general assumption is that our beliefs are responsive to changes in those facts that are in some sense 'available' to us. Let us put a name on this assumption (and dress it in somewhat precise language).

> *The Platitude(s) of Cause*: (1) S typically believes that p if she has available undefeated epistemic reasons for believing p or uncontroverted evidence that p — i.e. she has perceptual experiences or memories with the same or related content, or holds other beliefs relative to which p is probable. (2) S will typically lose the belief that p (or be prevented from forming that belief) if she has available defeating reasons or contrary evidence in regard to p — i.e. she has experiences with conflicting content or other beliefs relative to which p is improbable or that undercut purported evidence for p.

Now as S's case makes clear, we don't just assume that people believe those facts to which they have direct perceptual access. If I know that S believes that there is an apple in front of her, I will assume that she believes that there is an object in front of her, and (a bit more

precariously) that there is a fruit in front of her, something edible in that location, something plucked from a tree, and so on. Nor do we limit our assumptions to the implications of our observational beliefs. If I know that S remembers that Albany is the capital of New York, I will assume that she believes that the city is located within the borders of that state, that the legislature is housed there, etc. We typically assume that a person will believe all sorts of things that fairly obviously follow from those things we know she believes.

> *The Platitude of Inference*: If S believes that p she typically believes propositions that fairly obviously follow from p (where S has the conceptual capacities to entertain these implications).[2]

Still, to suppose that the influence of a given belief is limited to a believer's other beliefs would be to ignore the significant cognitive aspect that our emotional lives surely possess. Suppose I know that S believes that Slowpoke will not win the Kentucky Derby. I can then infer, with a great degree of reliability, that she will feel surprised if he does. Moreover, if I know that she hopes Slowpoke will win, I can reliably infer that she will be pleased if she comes to believe that Slowpoke will win and disappointed if she comes to believe that he will not. I here assume some kind of Platitude of Emotion.

> *The Platitude(s) of Emotion*: (1) If S believes that p (or, at least, believes that p with a fairly high degree of conviction) then she will feel surprised if she finds out that not p. (2) If S wants, hopes, or wishes that p with sufficient intensity, then she will feel good (e.g. pleased or happy) upon coming to believe that it will be the case that p, and bad (e.g. sad, anxious or disappointed) upon coming to believe that it will not be the case that p.

The Platitudes of Inference and Emotion don't just allow us to move from an initial assignment of belief, to attribute, on its basis, further mental states to the subject in question. Noticing someone's surprise or frustration at an outcome, I can infer the existence of the relevant belief as the best explanation of her reaction. But there is a further and even more widespread method of inference from behaviour: we see what someone says or does and while assuming that she is instrumentally rational, we at once explain her action or assertion as a manifestation of desire and belief. To indulge in a stock example: If I

[2] This principle should be understood in light of Kyburg's (1970) lottery paradox. It seems that I can believe (with a sufficiently high level of rational credence) of each ticket in a large, fair lottery that it will lose, without inferring that no ticket will win. Though Kyburg's case puts pressure on strong multi-premise closure principles, it is compatible with the truth of hedged generalizations like the Platitude of Inference. See Hawthorne (2004).

learn from S that she wants nothing more than to drink a beer and I see that she is making her way over to the refrigerator, I can assume that she believes that the best way to get a beer is by walking to the refrigerator. My version of this assumption is what I'll call the 'Platitude of Effect'.

> *The Platitude of Effect*: S typically believes that p if: (1) S wants q_1-q_n, (2) p represents (or constitutes) the information that w_1-w_n are available ways to satisfy S's desires for q_1-q_n, and (3) S is disposed to act in ways w_1-w_n.

There is yet a further route to initial attributions of belief. When I come to believe that p I typically do not base my higher-order belief on premises about how I have acted — as I would were I deploying the Platitude of Effect. Nor need I infer that I believe that p on the grounds that the evidence available to me indicates that p — as I would were I leaning on the Platitude of Cause. (This is especially clear when I attribute an unjustified belief to myself — a belief not in fact supported by good reasons.)[3] When we attribute beliefs to ourselves we typically proceed as though we don't need to do anything to figure out what we believe. Our practices of self-ascription therefore assume something like the following Platitude of Self-Knowledge:

> *The Platitude of Self-Knowledge*: If S believes that p, and entertains the proposition that she believes that p, she will typically come to believe that she believes that p without the aid of conscious inference.

These are, I think, the five most central platitudes that we can extract from our ordinary practices of belief attribution. If we think of them as theoretical principles developed in an attempt to explain and predict human cognition and behaviour we arrive at a picture of beliefs as states that: (1) typically arise from and are responsive to evidence, (2) play a certain inferential and (3) emotional role, (4) help guide and explain our actions, and (5) typically make themselves available to those that have them in a direct or non-inferential way. Again, to have something halfway precise (if disturbingly baroque) before us, we assume the following *belief paradigm*:

[3] Of course, in most (if not all) cases of this kind, the subject won't believe that the belief she self-ascribes is unjustified. But if she has no good evidence to believe what she does, she cannot gain access to her first-order belief by rehearsing the evidence. (Indeed, we might imagine that she is sufficiently rational that if she considered the evidence — instead of just reporting what she believes — she would abandon the unjustified belief that she knows herself to have.)

The Belief Paradigm: S believes that p if: (A) S is in some state R that represents that p; and (B1) S's occupying R is contingent upon the availability of undefeated epistemic reasons; and it is in virtue of S's being in R that: (B2) S is (subject to conceptual limitations) disposed to believe obvious implications of p, (B3) S is so disposed that she would feel surprise were she to discover that not p, (B4) S is disposed to act so as to satisfy those of her desires concerning which p represents an available way (or means) of satisfaction, and (again subject to conceptual limitations) (B5) S believes that she believes that p.

3. Looking for Belief's Essence

Still, while our implicit conception of belief has it playing many different roles, it seems that most of these roles are not token-essential. It must be admitted, for instance, that token beliefs do commonly fail to satisfy the Platitude of Cause. Perhaps one's beliefs as a whole must be regulated by evidence — so that a complete divorce of belief from epistemic reason is impossible — but we do have unjustified beliefs. Similarly, though familiar Quinean/Davidsonian considerations suggest that we cannot be accurately attributed beliefs singularly — or in isolation from other beliefs — we do sometime fail to believe even the most obvious implications of some of our beliefs. Finally, though we are typically surprised when confronted with the falsity of a firmly held belief, this is not always the case. At least in certain areas of our lives, we can develop a level of indifference that does not erode the strength of conviction.[4]

This leaves us with the Platitude of Effect and the Platitude of Self-Knowledge: behaviour and introspection. Are there any behavioural dispositions that are necessary for belief? Are there essential properties of belief that are revealed in introspection? Do individual beliefs even *have* essential properties? Certainly, our best efforts to use either platitude to discern what a person believes can come up short. If a person is a good enough actor she may know that she does not believe a proposition that the best third-person evaluation of her preferences and behaviour would suggest she believes. Similarly, Freud's influence has left us accustomed to allowing for cases of self-ignorance where third-person evaluation of a person's assertions and actions enables us to identify a disquieting belief that its bearer

[4] To cite an example of Ginet's (2001), an experienced card player might firmly believe that she has the winning hand, but experience defeat with the cool affect that helps maintain her poker face.

disavows. These epistemological limitations might lead us to conclude that beliefs have no essences, or that they are comparable to other natural kinds (such as water, gold, and heat) in having a hidden nature that can only be uncovered by advancing beyond folk theory to a more in-depth science of the mind. Perhaps introspective and behavioural evidence of various sorts must all be given some weight in our rough calculation of what a person believes, but no single epistemic route holds a privileged position over the others.

Nevertheless, even if this complex picture of belief attribution is an accurate one, there are certain cases where it is entirely unclear whether first-person introspection or third-person interpretation is to be given *greater* evidential weight when attributing beliefs to a subject. And, as we will see, examples of this sort highlight two substantially distinct views of how we should proceed in elaborating an account of belief itself — two different accounts of the relation between belief, attention, and the will.

4. Tacit Cognition

We do not have direct first-person access to the overwhelmingly vast majority of our cognitive states. This class obviously includes the sub-personal states of highly modular cognitive faculties — for example the 2.5 D sketch which David Marr (1982) claims is crucially involved in the processing of visual stimuli. But it also seems to extend further to include more 'informationally promiscuous' states of mind that are properly attributed to people (considered as whole organisms) and not just their cognitive parts. There are, for instance, various kinds of 'implicit memory' (Schacter, 1987). Long ago, Warrington and Weiskrantz (1968; 1982) showed that amnesiac patients retain memory traces of previously presented words despite an inability to explicitly recall having seen them. Wieskrantz (1986), in a now famous study, described patients with lesions of the striate cortex — so called 'blindsighters' — who sincerely report no conscious experience of relevant parts of their environment, but who perform better than chance when forced to guess about the shape and motion of stimuli in their visual fields. Prosopagnosic patients cannot explicitly recognize people they have met, but their emotional responses reveal memory-like representations with persisting cognitive effects (Bauer, 1984; Tranel and Damasio, 1985). Indeed, subsequent studies suggest that a great deal of person-level cognition may also be inaccessible.

Moreover, unconscious cognitive states are not limited to cases of cognitive pathology. For instance, social psychologists Greenwald and Banaji (1995) obtained tentative evidence of widespread unconscious race-, sex- and age-based prejudice. A representative experiment asks subjects to depress the 'e' key when good words — such as 'happy' and 'friendly' — appear on their computer screen's centre and to depress 'i' when bad words — such as 'miserable' and 'dangerous' — appear. Subjects are then asked to press one of these two keys when they see black faces on the screen's centre and another when they see white faces. Various permutations of these tasks follow. In one variation subjects are asked to press 'e' when white faces or bad words appear and to press 'i' when black faces or good words appear. Another task asks subjects to press 'e' when white faces or good words appear and to press 'i' when black faces or bad words appear. Most white subjects find it difficult to group good words with black faces and bad words with white faces, but find it easy to group good words with white faces and bad words with black faces. That is, white subjects make more mistakes and/or take longer when trying to pair good with black and bad with white than when pairing bad with black and good with white. These results hold independently of the attitudes toward race these subjects avow on a questionnaire.

Though experiments of this kind are still controversial, they are buttressed by reflection on ordinary cases of habituation, absent-mindedness and automaticity. Consider first a common example of absent-minded behaviour. Hope initially places a small trashcan in the cabinet beneath the sink in her kitchen. Eventually, she gets rid of the small, inconvenient can and places a large garbage bin next to the stove. Still, for months afterwards, when she is thinking about other things she absent-mindedly walks over to the sink with trash in hand. More often than not, when she grasps the handle to the cabinet door, she 'comes to' or realizes her mistake and turns to the new garbage bin in its place next to the stove.

What does Hope believe about the trashcan's location? Clearly, she quite often fails to be guided by the information that the trashcan is next to the stove. The somewhat complex behaviour of walking over to the sink, stooping, and opening the cabinet beneath is instead guided by an informational state that represents the area under the sink as the place for trash. Does the fact that Hope is not so disposed that she always (or even typically) acts as though she believes the trash is next to the stove imply that she does not really believe that the trash is next to the stove?

Note that if Hope's habit is persistent enough we must say that Hope *tends* to take the trash to the cabinet beneath the sink, and so does not tend to take the trash to the can next to the stove. It seems then that there is a straightforward statistical sense in which Hope is not disposed to take the trash to the can next to the stove even though this act would satisfy her desire to rid herself of trash.

For our second case, imagine that Rebecca is an architect well versed in structural engineering but that she is raised and educated on the plains and so lives a number of years without hiking up a mountain or visiting a skyscraper. Suppose too that she has a latent fear of heights that, because of her geographical isolation, has never been triggered. On her thirtieth birthday Rebecca goes to New York City to visit the Empire State building, but when she ascends to the top floor and approaches the enormous window to view the city displayed before her, she panics, retreats to the elevator, and returns to street-level as swiftly as possible. Does Rebecca believe that the Empire State Building is well built? Does she believe that ascending to its top story is perfectly safe? Did she at least believe these things when back on the plains? Her introspective judgments suggest that she has the belief, but Rebecca's latent fear insures that she does not have all the dispositions we commonly attribute to someone who believes that certain tall buildings are safe. There is a straightforward counterfactual sense in which she does not 'tend' to treat heights as safe even while living on the plains: were she exposed to heights of any substantial extent, she would act as though they were unsafe.

5. The Simulation Theory

There is a substantive difference between the evidence (grounds or reasons) we utilize when making self-ascriptions of belief and the evidence (grounds or reasons) that support our attributions of beliefs to others. Our examples highlight an important consequence of this epistemic asymmetry. Any explication of folk psychology that acknowledges the obvious difference between first-and third-person routes to belief will initially leave unsettled the classification of a representational state that fails to possess the full slate of properties seized upon in third-person belief attribution while fully displaying those phenomena seized upon in self-ascription. When we consider what our folk theory says about the typical behavioural effects of belief, we are led to conclude that Hope, Rebecca and the prejudiced liberal don't have the beliefs they claim to have; but when we consider the theory's claims about the customary introspective effects of belief

we are led into thinking otherwise. Theory-theorists must therefore look to scientific psychology, or theoretical considerations of a more general sort, for a determinate verdict of 'belief' or 'no belief' for the cases under discussion.[5]

But there are different ways of looking at our concept of belief. *Simulation theorists* argue that self-application of the concept is primary, so that ascription of beliefs to others piggybacks on our introspective processes. Different versions of the theory result from our marrying this idea to one or another account of self-ascription. For example, on Goldman's (1993) account, I will judge that I believe that p when it introspectively *seems* to me that I believe that p. Second-person attribution then works as follows: (a) I imagine myself in your position, (b) I note that (within this pretense) it seems to me that I believe that p, (c) I self-ascribe (within the pretense) the belief that p, and then (d) remove myself from the imagined situation and judge that (as I would believe that p were I in your position) you believe that p.[6]

Critics of the simulationist view have either argued that simulation theory ultimately collapses into some version of the theory-theory as one or more of the steps described above either consists in or requires for its successful completion mastery of the relevant theorems of folk psychology (Davies, 1994; Heal, 1994); or that existing evidence undermines the simulationist's explanation of the processes that actually underwrite third-person belief attribution (Gopnik and Wellman, 1992). Our purpose here, however, is not to adjudicate the debate over the truth or explanatory distinctness of simulationism, but to consider some of its theoretical commitments. If simulation theory were accurate, how would we apply the concept of belief in cases of absent-mindedness, irrational fear, and prejudice?

[5] Of course, our subjects won't lack *all* of the behavioural dispositions that typically accompany a given belief. For instance, as I've described the case Hope is disposed to assert that the trash can is next to the stove and is not disposed to assert that it is under the sink. I will assume, however, that being disposed to assert that p is not essential to believing p (as pre-linguistic children and non-human animals have beliefs). Moreover, though language using adults who believe an expressible p with sufficient conviction are typically so disposed that they will utter 'P' when they want to inform others that p, this fact cannot be used to decide between the neo-Cartesian and neo-behaviourist approaches to belief that I will describe below. There are no uncontroversial cases of absent-minded or weak-willed speech where assertion comes apart from self-ascription. Thus, when a subject asserts a proposition that she believes, the deliverances of first- and third-person routes to belief will perfectly coincide.

[6] Gordon's (1996) account differs in that self-ascription proceeds via *Evans' procedure*: a process modelled on Gareth Evans' (1982) observation that I typically answer the question of whether I believe that p by trying to figure out whether p. See Zimmerman (2004); (2005) for criticism of this kind of view and Zimmerman (2006) for criticism of Goldman's alternative account.

I ask you to put yourself in Hope's shoes. You've just removed the trashcan from under the sink and placed the new bin next to the stove. You can see the bin right there next to the stove. You have no idea that persisting habits will lead you to neglect this information in the future. Of course, you now think that you believe that the trashcan is next to the stove, and you are (seemingly) sincere when you say, 'The trashcan is next to the stove'. But your absent-minded behaviour indicates that you are not disposed to act in the right ways. If, as Robert Stalnaker argues, 'To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true' (1984, p. 15), then the fact that you lack the right dispositions implies that you do not believe that the trashcan is next to the stove, and you are mistaken in supposing that you do.[7] But if the simulationist is right about how we apply the concept of belief, we should all conclude, in contrast with Stalnaker, that Hope believes that the trash is next to the stove. This is the verdict that emerges when we imagine ourselves in Hope's position and — now removing ourselves from the pretence — attribute to her the belief we would attribute to ourselves were we in that position.

Next, imagine yourself in Rebecca's situation. You are a structural engineer living on the plains, studying the blueprints of the Empire State Building. You (seemingly) conclude that the building is well built, and that the people who enter into it, even those who ascend to its top floor, are quite safe. You have no idea that because of a latent fear of heights, you are disposed to panic. If this disposition indicates that you don't really believe that the Empire State Building is well built and don't really believe that heights (of the sort in question) are not dangerous, then you (i.e. Rebecca) are wrong about what you believe. But, again, this isn't what we should say about Rebecca's beliefs if our attribution results from our imaginatively projecting ourselves into her situation.

Finally, imagine that you are raised in a liberal household that stressed racial equality. You never needed an argument to convince you that people of African descent are neither evil nor bad — when you consider the proposition you (seemingly) think it obviously absurd. You are (seemingly) quite confident that a person's race has no bearing on her character. You then take the social psychology experiment described above and fail. Does that show that you never really believed in the moral irrelevance of race? If the simulationist is

[7] See Davidson (1984) and Dennett (1987) for neo-behaviorist views of belief crucially similar to Stalnaker's.

right, it shouldn't, and when we consider your situation in full detail, we shouldn't think that it does.

6. Theoretical Reactions

Bracketing our commitment to a particular model of the concept of belief, there are three distinct ways to respond to conflicts between our introspective beliefs and our affective and behavioural dispositions. First, we can salvage introspection. We can say that as Hope stands in front of the trashcan and notes its location she surely knows what she believes. If she knows that she believes that the trashcan is next to the stove then she has this belief, and if she has it, an adequate account of belief mustn't entail either that she does not have it or that it is indeterminate whether or not she does. If we adopt this tactic as a matter of policy, the beliefs that Hope, Rebecca and the liberal claim to have must instill behavioural dispositions that are more complex than those cited by Stalnaker when advancing the neo-behaviorist account of belief described above. Again, though adopting this strategy does not immediately require that we abandon physicalism (much less embrace substance dualism) the evidential privilege it grants to the first-person perspective somewhat justifies our labelling it *neo-Cartesian*. As we saw above, most simulation theorists are committed to the neo-Cartesian approach.

A second tactic would be to embrace the widespread fallibility of our best introspective judgments. Timothy Williamson, for instance, argues that a dispositional account of belief is incompatible with introspective 'transparency', and decides (for this and other reasons) that transparency must be rejected. As he says,

> [Transparency] fails for the state of believing p, for the difference between believing p and merely fancying p depends in part on one's dispositions to practical reasoning and action manifested only in counterfactual circumstances, and one is not always in a position to know what those dispositions are (Williamson, 2000, p. 24).

Similarly, Michael Smith argues for dispositional accounts of belief and desire on the grounds that competing 'phenomenological' conceptions are committed to problematic infallibility and luminosity theses (1994, pp. 104–25). Like Williamson, Smith claims that dispositional accounts of mentality are incompatible with a strong level of first person access; like Williamson, Smith decides that, of the pair, infallibility must be denied. Neither of these accounts goes so far as to identify beliefs with sets of behaviours or behavioural dispositions, and neither Williamson nor Smith directly addresses absent-

mindedness, fear and prejudice. But because they privilege evidence accessible from the third-person perspective, both reactions can be identified as *neo-behaviourist* in inspiration.

A third possible response to these cases would be to side with neither introspection nor behaviour by arguing that there is no fact of the matter as to what the imagined subjects believe. On this view, the content of 'belief' is fully captured by belief-ascribing platitudes like those we have uncovered. When platitudes conflict, the predicate with which they are associated neither determinately applies nor fails to apply in the given scenario. We are left with cases of what Eric Schwitzgebel (2001; 2002) has dubbed 'in-between' belief.[8]

Note that neo-behaviourists and deflationists cannot argue that simulationism collapses into the theory-theory as their reaction to our examples commits them to viewing the two theories as extensionally distinct. That is, simulationism is shown to be distinct from theory-theory by the fact that conducting the appropriate simulation leads one to the verdict 'outright belief' in our examples, whereas (on the neo-behaviorist's reckoning) correctly applying our folk psychological theory in these cases results in a verdict of 'no belief' or 'in-between belief'. Thus, neo-behaviorists must find some other reason for rejecting the simulationist approach. And reject the simulationist approach they must. Since simulationism vindicates neo-Cartesianism, the case for either neo-behaviourism or a Schwitzgebel-like deflationism must *begin* with arguments for a theory-theoretic treatment of the concept of belief.

Again, however, it is not my aim to adjudicate the debate between simulationists and theory-theorists here. Instead, I will try to draw out some of the theory-theory's methodological commitments and consider whether they might induce theory-theorists to join simulationists in embracing a neo-Cartesian metaphysics.

7. Naturalistic Evidence and Categorization

Someone who views 'belief' as a theoretical term cannot defend Schwitzgebel's deflationary answer on conceptual grounds alone. For suppose that 'belief' is neither determinately satisfied nor unsatisfied by the psychological sources of a certain class of actions. Then, so long as the theory through which 'belief' is introduced has no other label for the state of mind in question, it fails to explain and predict these actions: it has nothing substantive to say about them. If scientific

[8] See too Dennett (1995).

psychology contains terms that can be used to explain absent- minded, phobic and prejudiced behaviour and that are otherwise equivalent to 'belief' in their theoretical utility, these terms should come to replace 'belief' in our development of common sense psychology. (Folk theory would be supplanted on grounds of completeness or explanatory strength alone.) Moreover, once 'belief' is relegated to an outdated theory, eliminativism is hard to resist. Do the kinds of things denoted by the terms of abandoned theories really exist? Does phlogiston exist? Does the ether?

Of course, the deflationist might try to revise and improve upon folk psychology by giving a substantive *theory* of in-between belief — a theory that describes, among other things, how in-between beliefs combine with desires or preferences to produce behaviour; how they combine with beliefs (or other in-between beliefs) in the execution of inferences; how they respond to changes in the believer's evidence; and how they how they relate to cognitively conditioned emotions like surprise, anticipation, and disappointment. But there is a general (if somewhat abstract) worry about this move. If our characterization of in-between belief remains imprecise in nature in order to mirror the diversity of dispositions that ordinary thought attributes to outright belief, our account will allow for states that fall 'midway' between obvious cases of belief and obvious cases of in-between belief. If our deflationary intuitions tell us that cases falling midway between belief and non-belief are neither beliefs nor non-beliefs but instead in-between-beliefs, mustn't we say that states falling midway between belief and in-between-belief are states of in-between belief and in-between-belief belief? Surely this must stop somewhere — a manageable theory cannot posit an infinite number of distinct mental kinds. But if we resolve higher-order vagueness of this sort by positing a strict line between belief and in-between belief, why not reject the deflationist intuitions altogether and posit a strict division between belief and the lack thereof?

Notice that if the deflationist's account is to truly dissolve our puzzlement over the cases on hand, in-between belief cannot be advanced as a novel kind of belief. Nor can states of in-between belief be conceptualized as a kind of representational state other than belief. For if in-between belief is a kind of belief, Hope *et al*. do believe what they think they do despite failing to manifest important aspects of belief's paradigmatic behavioural profile. In contrast, if in-between belief is thought of as a non-doxastic mental state, the subjects under consideration are entirely mistaken about what they believe. In either event, our puzzle will not have been deflated. Instead, our 'deflationist' will

have chosen sides in the debate between the neo-Cartesian and the neo-behaviorist while introducing a novel term to cover his tracks.

Of course, we should all admit that folk psychology offers a somewhat rough-grained scheme for classifying states of mind. We might even go so far as to compare our common sense taxonomy to an underdeveloped scheme for classifying the colours — a vocabulary that distinguishes red from blue, but has no labels for many of their shades, and no label at all for purple. And, just as introducing 'purple' would allow us to improve our impoverished account by labeling a range of colours in-between the reds and the blues, and just as introducing 'crimson' would allow us to distinguish shades of red we could not previously name, one might think that introducing 'S in-between believes p' into our psychological theorizing will allow us to label states of mind we cannot currently discriminate with the dichotomous 'S believes p' and 'It is not the case that S believes p'.

But the analogy does not hold. We can surely improve the simple colour scheme we've imagined by introducing terms for shades of red and shades of blue, and we can also improve it by classifying colours, like purple, that are neither shades of red nor shades of blue. But we cannot truly *improve* our understanding of the colours by introducing the label 'in-between red' while insisting that the colour it denotes is neither a shade of red nor a colour distinct from red (or a shade of such). Gains in understanding cannot be bought with frivolous departures from classical logic; excluded middle should constrain the shape of our theory unless we are truly forced to abandon it by explanatory necessity. Thus, we are inevitably led to ask whether the frame of mind of a Hope or Rebecca is best categorized as kind of belief (a shade of red) or something distinct from belief (a colour other than red). If we refuse to make this choice, our theorizing will not result in an expansion or enrichment of folk psychology, but will instead risk supplanting our common sense view of the mind with something wholly other.

This more general worry that theoretical terms might come to supplant 'belief' in our best description of the mind is more than just a thought experiment, for there is a growing consensus within contemporary experimental psychology that there are (at least) three different systems of representations directly shaping human and non-human behaviour. One informational system is supposed to account for habits, skills, instincts and stimulus-response behaviour; another is supposed to underwrite our emotional reactions to previously experienced people and places; the third is thought to contain the kind of representations that inform speech and other robustly intentional

forms of action. Following Howard Eichenbaum and Neal Cohen's recent (2001) overview of the literature I will call these three systems respectively: (1) procedural memory, (2) emotional memory, and (3) declarative memory.[9] Now, though all three systems of memory recruit (and require) the cerebral cortex for their operation, Cohen and Eichenbaum argue that the procedural system involves the cerebellum and striatum, the emotional memory system the amygdala, and the declarative system the hippocampus. Evidence for this degree of anatomical distinctness comes primarily from functional dissociations accompanying neurological deficits. For example, an amnesiac with an impaired hippocampus but a healthy amygdala and striatum might report having no memory of a person, but have a positive or negative emotional response to a meeting that is commensurate with her past experience. Similar subjects can be taught to knit without being able to recall how they acquired the skill. When coupled with controlled experiments on rats and non-human primates, these observations look to dissociate declarative from emotional and procedural memory and link each system with a somewhat distinct neural correlate.

Though the evidence for Cohen and Eichenbaum's thesis is inconclusive, we can ask ourselves as metaphysicians of belief how we should react if the scientific community settles on its truth. Have scientists then discovered that beliefs are constituted by representations in the declarative memory system and so realized in (say) firing potentials spread over the cortex and hippocampus? Or should we conclude that there are three distinct types of belief corresponding to representations in the three different memory systems? Suppose — as would seem plausible if Eichenbaum and Cohen are vindicated — that Hope has a representation of the trashcan's being next to the stove instantiated in her hippocampal system but none in her cerebellum and striatum. Should we conclude that Hope believes that the can is next to the stove and has a non-belief-constituting procedural memory of its being located under the sink? Should we say that Hope both believes that the can is under the sink and believes that it is next to the stove though these are beliefs of two different neurological kinds? Or should we stick with the neo-behaviorist's metaphysics and conclude that unless Hope has the relevant representation in both her procedural and declarative systems, and so is disposed to both attentively and habitually act as though she believes that it is next to the stove, she

[9] The relevant body of work is too vast and the experiments too detailed to adequately discuss here. Some of the most important are Bechara *et al*. (1995), Knowlton *et al*. (1996), and McDonald and White (1993). Again, see Eichenbaum and Cohen (2001) for a book-length review with an extensive bibliography.

fails to believe what she takes herself to believe? How should we best integrate folk and scientific concepts of the mind?

Unfortunately, there are no obvious rules guiding the process. As Burge (1979) points out, lay people commonly allow the extensions of their proto-scientific concepts to be fixed by experts. But it is difficult to say exactly which considerations move scientists when they deploy folk concepts in reporting the results of their inquiries. For instance though jadeite and nephrite differ chemically, instances of both kinds answer to our shared concept of jade (Putnam, 1975, p. 241). In point of fact, jewellers must distinguish between the two stones because jadeite is more valuable than nephrite; but instead of securing this result by reserving 'jade' for jadeite, they appeal directly to our technical labels for the two kinds of jade. On the other hand, 'fool's gold' picks out pyrite not marcasite even though both minerals are chemically FeS_2. (The two substances are called 'polymorphs' because they are denoted by the same chemical formula despite their structural differences.) Of course, experts could say that pyrite and marcasite are both kinds of fool's gold, or that there is really only one kind of jade, but they typically don't.[10]

It seems that structurally identical theories of cognition will result so long as we employ *some* blanket term for all three systems of representation and *some* terms that exclusively apply to each of the members of its tripartite extension. And because considerations of theoretical utility leave more than one path for future expert use of 'belief' to take, we incur substantial risks by deferring to psychologists in fixing the contours of our concept. If a measure of brute historical accident decides the way that 'belief' will be used within the scientific community, our deference to experts will impart these contingencies to our thinking about belief. The outcome of the debate between neo-Cartesians and neo-behaviorists — and, perhaps somewhat more shockingly, an answer to the question of whether or not Hope believes that the trashcan is under the sink — will then have to await our discovery of linguistic patterns that are crucially underdetermined by rational constraints. Though we will not be forced to say, with Schwitzgebel, that there is no fact of the matter as to whether

[10] Analogously, metaphysicians could say that water has more than one possible chemical composition: H_2O and whatever constitutes water on Putnam's Twin-Earth. But we don't. (Perhaps this is because when we imagine the chemical theory that correctly describes XYZ and accounts for its superficial similarity to H_2O, the differences seem too great to treat 'water' like 'jade'.) We can recognize that our concept works in this way without possessing a fully adequate justification or rationale for its doing so. As use of 'jade' illustrates, there need be nothing of predictive or explanatory importance we can gain by conceptualizing XYZ as a kind of water.

inattentive, phobic and prejudiced people have the beliefs they claim to have, we will have to admit that the question is largely superficial. To use Wittgenstein's metaphor: we will find ourselves not riding down tracks laid in advance, but hastily supplying the needed rails as we progress into unmapped terrain.

8. Normative Constraints

There is, however, a way for the theory-theorist to resist the road to anti-realism. For it might turn out that even though neo-Cartesian and neo-behaviourist theories of belief mesh equally well with contemporary cognitive science, one theory makes better sense of belief's *normative* dimension than the other. Sometimes we believe what we shouldn't believe or fail to believe what we should. A belief can be criticized as false, hasty, unreasonable, or bizarre, while one's failure to believe something true or evident may be chastised as overly sceptical, narrow-minded, self-serving, or ignorant.[11] Furthermore, though radical expressivists are mistaken in seeking to identify these critical judgments with the kind of raw partisan allegiance one might adopt toward a sports team, our conceptually articulated criticisms are often accompanied by sentiments. For instance, I will often become embarrassed by a loved one's ignorance or grow proud that she has unflinchingly acknowledged a disturbing truth.

Of course, at the outset of formal inquiry we should allow the possibility that the most explanatory and predictive theory of human action will undercut rather than vindicate these important normative distinctions. Realists cannot insist from the get-go that a mature scientific psychology will make sense of our critical practices.[12] Nevertheless, when two different ways of integrating common sense and scientific psychology satisfy theoretical conditions equally well, it seems entirely appropriate to favour the one that makes better sense of our reactive attitudes, for these attitudes, no less than the truisms we assume when attributing states of mind to one another, lend shape to our pre-theoretic psychological concepts. If one of several empirically equivalent codifications of scientific psychology uses 'belief' in a way that is commensurate with our critical practices, it better respects the phenomena

[11] More formally, I can criticize a person's beliefs if they are baldly inconsistent or commit her to a disastrous series of bets (despite her well-ordered preferences). For a recent discussion see Christensen (2004).

[12] If, for instance, we were to discover that a person's morally reprehensible behaviour is caused by a genetically encoded neural structure that can only be eliminated with as yet unavailable surgery, this would place into doubt our warrant for continuing to blame her for it.

we were trying to explain. At any rate, this should be admitted by any variety of psychological realism worth defending.

There is, therefore, some point to asking whether our reactive attitudes are better vindicated by a neo-Cartesian or neo-behaviourist perspective. When we adopt positive and negative attitudes towards beliefs do we assume the truth of either theory? To answer this question in favour of the neo-Cartesian we need not defend a systematic account of doxastic norms and reactive attitudes. We need only recognize that common thought contains a distinctively doxasitic set of criticisms: cognitive evaluations that can be pried apart from those properly aimed at absentmindedness, phobia and prejudice. Surely one can criticize Hope as absentminded without meaning to belittle her as someone ignorant of the lay-out of her own kitchen; one can label Rebecca 'phobic' without intending to deride her as an architect ignorant of structural engineering; and one can insist that some of the liberal's reactions reveal a lingering prejudice without implying that he remains unconvinced of established accounts of the biological superficiality of race. Corresponding remedial dissociations are also easy to affect. Though it is proper to urge that Hope pay attention to what she is doing, it would be entirely improper to try to convince her that the trashcan is next to the stove. It is legitimate to ask Rebecca to engage in explicitly behavioural therapies — where, say, she is gradually introduced to heights in conjunction with calming stimuli — but it would be misguided to request that she learn more about building safety. (Similar points hold of the prejudiced liberal.)

Now suppose it turns out that these remedial dissociations correspond to the functional and anatomical dissociations for which Eichenbaum and Cohen argue. Suppose, that is, that representations realized in the hippocampus are sensitive to conversation, argumentation and school-book learning whereas representations in the amygdala and striatum only respond to habituation. In such a case, psychologists should report having discovered that belief is constituted by representations in the declarative memory system — representations instantiated in the cortex and hippocampus. Psychologists who fail to adopt a neo-Cartesian view of belief will have mistakenly characterized their own findings. They will have driven our train off rails laid down well in advance.

9. Conclusion

Because relevant empirical considerations have not yet been resolved, we should not pretend to possess a finished metaphysics of belief.

Still, we can now see the kinds of factors we must consult in developing a more nuanced and realistic account than those currently on offer. If simulationism can be shown to be correct, a neo-Cartesian metaphysics is insured. But even if some version of the theory-theory should prove better grounded, experimental evidence might turn out to vindicate the view. Indeed, though the needed empirical research is incomplete, the current state of things suggests that a neo-Cartesian metaphysics is more likely true than not.

Nevertheless, though the neo-Cartesian argues for looser ties between belief and behaviour than do functionalists of Stalnaker's ilk, the differences between the two camps should not be exaggerated. For even if we were to go so far as to identify token beliefs with representations in the declarative memory system, we could retain functionalism's central insights. We might maintain, for instance, that though the declarative system is realized in the cortex and hippocampus in humans, radically different physiological structures might serve the same cognitive function in other species. (Thus, the neo-Cartesian needn't overreact by asserting the type-identity, 'Believing that p just is having a representation that p in one's *hippocampal* system'.) Surely, neo-Cartesianism is compatible with the phenomenon of multiple realization that motivated many functionalists to turn away from straightforwardly biological approaches to the mind.

Nevertheless, the division between neo-Cartesians and neo-behaviourists is substantial. Neo-Cartesians insist that certain absent-minded, phobic and prejudiced behaviours are not truly indicative of belief, and they will further insist that these cases constitute more than just a string of unrelated counterexamples to neo-behaviourism. Hope's failure to act on the relevant piece of information is compatible with her believing it precisely because she wasn't paying *attention* to what she was doing. Rebecca's failure is compatible with her holding the belief in question because she wasn't in full *control* of her actions. Thus, according to the neo-Cartesian, a given belief will not instill a disposition to act that is simply conditional on certain desires and other beliefs. Instead, beliefs instill dispositions to act given certain desires, other beliefs, and a sufficient level of attention and self-control. Neo-Cartesianism thus raises *volition* and *conscious awareness* to a position within the study of mind that is at least as exalted as that enjoyed by belief, desire and other propositional attitudes.

Of course, whether the neo-Cartesian account also undermines the physicalistic perspective driving most functionalists away from dualism depends on whether attention and resolution can be rendered

physically respectable. At the very least, then, embracing neo-Cartesianism shifts the debate over physicalist theorists of belief into a quarrel over the nature of consciousness and self-control. We are left with metaphysical questions even more difficult than the one with which we started: a kind of progress perverse enough to be deemed philosophical.

Acknowledgement

I would like to thank Daniel Dennett, Tamar Szabo Gendler, Carl Ginet, Eric Schwitzgebel, Sydney Shoemaker, Robert Stalnaker and audiences at the University of Calgary, Johns Hopkins, the University of Memphis, and the University of Vermont for helpful discussions of earlier incarnations of this work.

References

Bauer, R.M. (1984), 'Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test', *Neuropsychologia*, **22**, pp. 457–69.

Bechara, A., Tranel, D., Damasio, H.D., Adolphs, R., Rockland, C. and Damasio, A.R. (1995), 'Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans', *Science*, **269**, pp. 1115–18.

Burge, T. (1979), 'Individualism and the mental', in P. French, T. Uehling and H. Wettstein (ed.), *Studies in Metaphysics* (Minnesota: MUP).

Carruthers, P. (1996), 'Simulation and self-knowledge', in P. Caruthers and P.K. Smith (ed.), *Theories of Theories of Mind* (Cambridge: CUP).

Christensen, D. (2004), *Putting Logic in Its Place: Formal Constraints on Rational Belief* (Oxford: Clarendon).

Davidson, D. (1984), *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press).

Davies, M. (1994), 'The mental simulation debate', in C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness: Proceedings of the British Academy 83*, (Oxford: OUP).

Dennett, D.C. (1987), *The Intentional Stance* (Cambridge, MA: MIT Press).

Dennett, D.C. (1995), 'Do animals have beliefs?' in H. L. Roitblat and J. Meyer (ed.), *Comparative Approaches to Cognitive Science* (Cambridge, MA: MIT Press).

Eichenbaum, H. and Cohen, N.J. (2001), *From Conditioning to Conscious Recollection: Memory Systems of the Brain* (Oxford: OUP).

Evans, G. (1982), *Varieties of Reference* (Oxford: Clarendon Press).

Fodor, J. (1998), *Concepts: Where Cognitive Science Went Wrong* (Oxford: Clarendon Press).

Ginet, C. (2001), 'Deciding to believe', in M. Steup (ed.), *Knowledge, Truth and Duty: Essays on Epistemic Justification, Responsibility and Virtue* (Oxford: OUP).

Goldman, A. (1993), 'The psychology of folk psychology', *Behavioral and Brain Sciences*, **16**, pp. 15–28.

Gopnik, A. and Wellman, H. (1992), 'Why the child's theory of mind really is a theory', *Mind and Language,* **7**, pp. 145–71.

Gordon, R. M. (1996), '"Radical" simulationism', in P. Caruthers and P. K. Smith (ed.), *Theories of Theories of Mind* (Cambridge: CUP).
Greenwald, A. G. and Banaji, M. R. (1995), 'Implicit social cognition: attitudes, self-esteem and stereotypes', *Psychological Review*, **102**, pp. 1–27.
Hawthorne, J. (2004), *Knowledge and Lotteries* (Oxford: OUP).
Heal, J. (1994), 'Simulation vs. theory theory: What is at issue?' in C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness: Proceedings of the British Academy 83* (Oxford: OUP).
Knowlton, B.J., Mangels, J.A. and Squire, L.R. (1996), 'A neostriatal habit learning system in humans', *Science*, **273**, pp. 1399–401.
Kyburg, H. (1970), 'Conjuctivitis', in M. Swain (ed.), *Induction, Acceptance and Rational Belief* (Dordrecht: Reidel).
Lewis, D. (1972), 'Psychophysical and theoretical identifications', *Australasian Journal of Philosophy*, **50**, pp. 249–58.
Marr, D. (1982), *Vision* (New York: W.H. Freeman).
McDonald, R.J. and White, N.M. (1993), 'A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum', *Behavioral Neuroscience*, **107**, pp. 3–22.
Putnam, H. (1975), 'The meaning of "meaning"', first published in K. Gunderson (ed.), *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science VII*. Minnesota: MUP, reprinted in Putnam, (1975), *Mind, Language and Reality: Philosophical Papers Vol. 2* (Cambridge: CUP).
Schacter, D.L. (1987), 'Implicit memory: history and current status' *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13** (3), pp. 501–18.
Schwitzgebel, E. (2001), 'In-Between believing', *Philosophical Quarterly*, **51**, pp. 76–82.
Schwitzgebel, E. (2002), 'A phenomenal dispositional account of belief', *Nous*, **36**, pp. 249–75.
Smith, M. (1994), *The Moral Problem* (Oxford: Blackwell).
Stalnaker, R. (1984), *Inquiry* (Cambridge, MA: MIT Press).
Tranel, D. and Damasio, A.R. (1985), 'Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics', *Science*, **228**, pp. 1453–4.
Warrington, W.E. and Weiskrantz, L. (1968), 'New method of testing long-term retention with special reference to amnesic patients', *Nature*, **217**, pp. 972–4.
Warrington, W.E. and Weiskrantz, L. (1982), 'Amnesia: A disconnection syndrome?' *Neuropsychologia*, **20**, pp. 233–48.
Wieskrantz, L. (1986), *Blindsight* (New York: OUP).
Williamson, T. (2000), *Knowledge and Its Limits* (Oxford: OUP).
Zimmerman, A. (2004), 'Unnatural access', *Philosophical Quarterly*, **54**, pp. 435–8.
Zimmerman, A (2005), 'Putting extrospection to rest', *Philosophical Quarterly*, **55**, pp. 658–61.
Zimmerman, A. (2006), 'Basic self-knowledge: Answering Peacocke's criticisms to constitutivism', *Philosophical Studies*, **128**, pp. 337–79.

Paper received January 2007