Reliabilism and Brains in Vats

Jon Altschul

Loyola University New Orleans

Published 2011 in *Acta Analytica*, Vol. 26, no. 3: 257-272

In the wake of Hillary Putnam's famous *Reason, Truth, and History*, philosophers have sought ways to connect content externalism with issues relating to epistemology. In this paper, I attempt to draw one such connection as it relates to the debate over internalism and externalism in epistemology. The internalism-externalism distinction can be drawn in a variety of ways. For the purposes of this paper, I will understand internalism as the claim that the justificational status of a subject's beliefs is determined by non-factive mental states, such as beliefs and experiences, and by relations between these states. Externalism is the denial of this claim. Externalists hold that certain external facts, such as facts about the world or the reliability of a belief-producing mechanism, affect a belief's justificational status.[1]

There has been an increasing amount of attention within the last twenty years or so devoted to Cartesian thought experiments in the context of the internalism-externalism debate in epistemology. The consensus among internalists is that considerations about evil demon victims and brains in vats provide excellent reason to reject externalist

[1] The version of internalism that is the focus of this paper is what Conee and Feldman (2001) call *Mentalism*. This is to be contrasted with *Accessibilism*, according to which the features that justify a belief are accessible to the subject. There are indeed important questions about the connection between accessibility and justification, but these cannot be addressed here.

theories of justification.[2] In particular, *the new evil demon problem* draws on an intuition that these deceived or envatted subjects are equally as justified in their beliefs as their non-deceived or non-envatted counterparts.[3] However, since externalists are unable to account for how subjects in skeptical situations are justified, internalism is said to win out over externalism.

I think these considerations do not at all help the internalist cause. I will argue that by appealing to the anti-individualistic nature of perception, externalists have the resources available to explain why the beliefs held by brains in vats are justified. Contrary to what internalists contend, skeptical scenarios offer us no reason to prefer internalism to externalism. Thus, by assuming that perceptual anti-individualism is a plausible view, I will show that externalists are able to circumvent the new evil demon problem.

§ I: THE INTERNALIST'S ARGUMENT AGAINST EXTERNALISM

What is the motivation behind this internalist strategy against externalism? We are asked to imagine that some skeptical hypothesis is true. Suppose that unbeknownst to her, a subject (call her Brainean) is nothing more than a brain in a vat of nutritious liquid that is hooked up to a super computer. It seems to Brainean that she has a body, that she goes about her daily life on Earth (drinking coffee in the morning, driving right around the speed limit in her Honda, etc.), yet all of the experiences she has is systematically caused

[2] Cohen (1984), Wedgwood (1999; 2002), Dretske (2000), Pryor (2001). See Majors & Sawyer (2005) for detailed overview of this internalist strategy.

[3] The new evil demon problem was introduced in Lehrer and Cohen (1983).

by the super computer. As it turns out, *none* of her experiences are accurate, and all of the beliefs she forms on the basis of those experiences are false.

If we highlight one historically popular externalist theory, Alvin Goldman's process reliabilism,[4] we can quickly see how the above story is meant to raise an objection. Goldman's theory states that a subject's belief is justified if and only if the belief was produced by a reliable belief-forming process, where a process is reliable just in case it has a sufficiently high truth-ratio. Since, as the present objection suggests, a systematically deluded brain in a vat like Brainean is a maximally unreliable believer, process reliabilism dictates that none of her beliefs are justified. However, there is a strong intuition that reliabilism delivers the wrong result here. Suppose, as the internalist assumes, Brainean currently enjoys a perceptual experience of a red cube in front of her (call this E1). Based on E1, Brainean comes to believe:

(B1) There is a red cube before me.

Surely B1 does not constitute knowledge, for knowledge requires truth and B1 is false. But many philosophers hold that Brainean is at least justified in holding B1. If this claim can be successfully defended, then there appears to be a reason to reject Goldman's reliabilism: B1 is justified, yet it was produced by an unreliable process—hence reliability seems not to be necessary for justification.

Moreover, the fact that B1 is justified would appear, on the face of it, to undermine *all* externalist theories of justification. Why is this? Recall that externalists hold that certain external facts are relevant to whether a belief is justified. In Brainean's situation, there seems to be nothing in her surrounding environment—such as facts about reliability

[4] Goldman (1979)

or the state of the world—that could figure in as a feature of what makes her justified. All that is epistemically relevant in the thought experiment, says the internalist, is what goes on in Brainean's mind, so the correct account for why she is justified would only mention internal states, as well as various relations between them. Thus, the internalist may want to argue for her position along the following lines:

BRAIN IN A VAT ARGUMENT AGAINST EXTERNALISM (BVAE)

P1. For all possible justified beliefs, brains in vats hold some of those beliefs.[5] (assumption)

P2. If a brain in a vat holds justified beliefs, then nothing in the external environment could play a role in constituting the justification for these beliefs. (assumption)

P3. So, nothing in the external environment could play a role in constituting the justification for a brain in a vat's beliefs. (P1, P2)

P4. Externalism is the thesis that the justificational status of a belief partially depends on facts about the external environment.

C5. Thus, externalism is false. (P3, P4)

The project of this paper is to offer an effective reply to this argument on behalf of the externalist, thus showing that BVAE does not successfully establish the falsity of externalism. Although it is not my aim to defend Goldman's theory, I hope to at least show that process reliabilism is *consistent* with the fact that brains in vats can hold justified beliefs. If I am successful, then I will have demonstrated that this argument fails to undermine reliabilism. Moreover, since reliabilism is an externalist theory of justification, epistemic externalism is not threatened here. Thus, even if we assume that P1 is true, I will show that P2 is false.

---

[5] As a qualification, it is unclear, and unlikely, that this argument could be applied to certain *a priori* beliefs (such as mathematical beliefs) we have as well as beliefs about one's own mental states. The argument specifically targets beliefs which are sourced in perception. It should therefore be read as focusing on our common, ordinary beliefs about the empirical world.

To arrive at my rejection of the above argument, we must first understand the internalist's primary motivation for P1. Why should one think that a brain in a vat can hold justified beliefs? In the aforementioned example, Brainean's belief that there is a red cube before her (B1) is based on her perceptual experience that represents a red cube (E1). Most epistemologists can agree that if B1 is a justified belief, E1 plays a crucial role in the explanation of why it is justified. E1 is a *ground* of Brainean's belief: it is that which makes it rational for her to believe that there is a red cube before her, as opposed to a blue cube or a red sphere or nothing at all. Part of the reason why E1 justifies B1 is that the representational content associated with the experience indicates the presence of a red cube. E1 does not represent a blue cube or a red sphere, so a belief that there is a blue cube ahead (or a red sphere) could not be justified on the basis of E1.

The content associated with E1 therefore must play an important epistemic role in the evaluation of B1. But why does the internalist think that E1 adequately justifies Brainean in holding B1, given that E1 is a non-veridical experience and B1 is produced (as we are assuming for the moment) by an unreliable process? Let us recall that the internalist holds that justification is determined solely by one's mental states. This implies that for any two subjects that have the exact same mental states, there should be no difference with respect to the justificational status of their beliefs. A helpful way to illustrate this point is to consider mentally identical twins, as several philosophers have done.[6] Imagine Actuan, a normal human in the actual world who causally interacts with objects in the external world and reliably forms true beliefs on the basis of her experiences. It is

[6] Cf. Cohen (1984), Sosa (1985), Wedgwood (1999), Dretske (2000), Conee and Feldman (2001)

reasonable to suppose that when Actuan has a visual experience of a red cube in front of her (call this E2) and on its basis believes:

(B2) There is a red cube before me,

her belief is justified. Now suppose that Actuan and Brainean are mental twins: all of their mental states are qualitatively and representationally type-identical, and their methods of reasoning and deliberation are indistinguishable. So, E1 is type-identical to E2 and B1 is type-identical to B2. Brainean and Actuan hold their beliefs *on the same grounds*. Assuming that Actuan is justified, the internalist claims that Brainean is justified because, given that the two are twins, whatever justifies Actuan in holding B2 must also justify Brainean in holding B1. As I see it, the internalist's defense of P1 therefore hinges on what I will call the *Same Grounds Principle*:

(SGP) For any two subjects S and S*, if S and S* both believe that *p*, and their token beliefs are each based on the same grounds, then S and S* do not differ with respect to the justificational status of their beliefs.

With the introduction of SGP, we can now see how the internalist will to want defend P1 of BVAE. She can argue for the claim that Brainean's belief that there is a red cube before her (B1) is justified as follows:

i) Actuan's belief (B2) is justified. (assumption)

ii) E1 is identical to E2 and B1 is identical to B2. (assumption)

iii) E1 and E2 are the grounds for B1 and B2 respectively. (assumption)

iv) Thus, Brainean's belief (B1) is justified. (i-iii, SGP)

Having now seen how the internalist attempts to reject externalism on the basis of skeptical thought experiments, let us pause to consider whether she is permitted to make the three assumptions given in this defense of P1. As for assumption (i), the overarching debate between internalists and externalists we have put forth centers on an intuition felt

by internalists that the beliefs held by brains in vats are no less justified than those held by their ordinary, real world counterparts. There would not be much point in having this debate if we could not assume (i). Thus, we ought to go along with the assumption that at least a large number of our everyday, common beliefs are justified. Assumption (iii) is unproblematic given the way we laid out the stories of Brainean and Actuan.

Assumption (ii), however, is more worrisome and deserves attention. As we have seen in this section, it is crucial for the internalist's argument against externalism that E1 is identical to E2 (that is, they can both be accurately described as representing something *as* a red cube).[7] For unless they are the same, the internalist cannot appeal to SGP and maintain that if Actuan's beliefs are justified, Brainean's beliefs are as well. In the next section, we will consider how reflections on the nature of perception place an important restriction on whether the internalist's skeptical scenario is consistent with the claim that Brainean and Actuan's experiences share the same type of representational content, and moreover whether assumption (ii) is true. These considerations will pave the way for my rejection of BVAE in Section III.

§ II: PERCEPTUAL ANTI-INDIVIDUALISM AND SAMENESS OF GROUNDS

[7] Note that we are investigating whether E1 and E2 have the same *representational* content. As we will see, perceptual anti-individualism places restrictions on when two token experiences can be representationally type-identical. However, it is a separate question to ask whether or not E1 and E2 are *phenomenally* or *qualitatively* identical. One can be anti-individualist about perceptual representational content, yet still hold that qualia are intrinsic, non-relational properties of the subject (though Cf. Dretske (1995; 1996) for an interesting argument for qualia anti-individualism). None of the present discussion hinges on whether or not Brainean and Actuan share the same qualia, so I will here assume individualism about phenomenal experience.

Is it the case that a brain in a vat like Brainean can share the same grounds for a belief with someone non-envatted like Actuan? Whether two subjects believe a proposition on the same grounds—such as a perceptual state—depends on the whether those grounds are of the same representational state-type, where two token representational states are of the same type just in case they share the same content. So the answer to this initial question hinges directly on our characterization of the individuation conditions for perceptual contents.

It has been proposed that perceptual states have anti-individualistic individuation conditions. One prominent advocate of perceptual anti-individualism (sometimes called perceptual content externalism) is Tyler Burge (2007b; 1986; 2003; 2005) and he describes the view as follows:

> Perceptual anti-individualism entails that the individuation and natures of perceptual states are necessarily associated with certain relations between the types of states that are part of the perceptual system of the individual, on one hand, and kinds of objects, properties, and relations in the environment, on the other. The relevant kinds of objects, properties, and relations are those that enter into causal relations that help set conditions under which perceptual states and standards of their correctness are individuated. (2005, p. 4)

Before describing this view in further detail, let me pause to note that while it is endorsed by many within the philosophical community, perceptual anti-individualism is by no means a universally accepted thesis. It not my intention to defend the view here; to do so would go well beyond the scope of this paper. Rather, my aim is simply to show that if one accepts that perceptual anti-individualism is correct, then the externalist has an effective way to respond to the new evil demon problem.

According to perceptual anti-individualism the possession of a given perceptual state-type depends on past regular and systematic interactions between the perceptual system and features in the environment. How the perceptual system of an individual will

represent a given object, or the way that the object is represented, is constrained by the specific history of that system, where this includes the kinds of environmental features with which the system came into causal contact. The representational content associated with a given type of perceptual state cannot be fully characterized without appeal to such a causal history.

What are the sorts of causal interactions that help establish what a subject represents in perception? Typically, as a matter of empirical fact, one's ability to represent, say, A as F is explained by one's past interactions with FA's. But the relevant interactions that contribute to the individuation of a perceptual state need not be of this sort. There are three alternative explanations.[8] First, the subject's representation can be a composite of percepts (e.g. A and F) the possessions of which are explained by her interactions with the referent of those percepts in isolation of the other (e.g. she has confronted A's that were not F, and F-ness that did not belong to A's). So it is consistent with perceptual anti-individualism that I perceptually represent a red cube even though I have never actually confronted cubes that were red, or that I misrepresent a large limestone boulder drenched in a dusk sunlight as a pink elephant, a creature I have never before confronted. Secondly, the possession of a state can be completely independent of the subject's personal history. For instance, the subject may have never interacted with FA's (nor with A's nor with F-ness), but past members of the species of which the subject is herself a member have. A harmless distal stimulus may trigger a (inaccurate) predatory alert representation in a creature, yet the relevant content-individuating predators posed threats only to her ancestors, and died out well before the creature was born. This ability would

[8] Burge (2007b, p. 202-203) and (2003, fn 7)

be a trait passed down through the generations within the species. Finally, it is consistent with perceptual anti-individualism that neither the subject nor her ancestors has ever confronted instances associated with a given representation-type at all. For example, I may (accurately or inaccurately) represent a shade of color with which I (or my ancestors) have never interacted. Similarly, I may represent a 26-sided polygon even if this marks the first confrontation I have had with this shape. This is possible, however, only on the background condition that successful interactions have occurred between my perceptual system (or that of my ancestors) and instances of other relevant and salient color and shape properties.

With respect to combating the new evil demon problem, the epistemic externalist should take interest in what the anti-individualist has to say about the minimal requirements for which two subjects are capable of being in the same type of perceptual state—such as Brainean and Actuan, whose experiences were both said to be representing a red cube. One of these requirements is that the histories of the perceptual systems associated with the two subjects include causal interactions with relevantly similar objects or properties in their environments. This suggests that from the anti-individualist's point of view, whether or not it can be assumed that Brainean and Actuan share the same mental lives depends on the temporal scope of Brainean's envatment. For, if Brainean has *always* been a brain in a vat, and hence has never had any interactions with red objects or with cubes, then—according to perceptual anti-individualism—E1 would not be a perceptual representation of a red cube (and moreover it would be incorrect to assume that Brainean is Actuan's envatted *mental twin*).

Representing red cubes requires some connection to features in the environment, and a permanently envatted brain does not satisfy this requirement.

Of course, if the internalist's proposed thought experiment was such that Brainean is instead a *recently* envatted brain in a vat, where all of her life up until very recently she was a regular person in the actual world, then the individuation of the contents of her perceptual experiences will have been established in relation to a causal history between genuine mind-independent objects and Brainean. Now once Brainean unknowingly becomes envatted and the super computer thereafter causes her to have the experience with phenomenal properties corresponding to what red cubes look like, it is consistent with perceptual anti-individualism that Brainean perceptually represents a red cube. Thus, there is a difference in representational content between the experiences of a permanently envatted Brainean and a recently envatted Brainean.[9] Only the latter character is capable of representing ordinary, mind-independent objects like red cubes.

In light of our discussion in this section, let us evaluate whether the internalist can maintain assumption (ii) in defending P1 of BVAE. Is it true that E1 and E2 have the same content? That depends on which version of the thought experiment the internalist presents. We have already seen that in accordance with perceptual anti-individualism, if Brainean is permanently a brain in a vat, E1 cannot be a representation of a red cube. Since E2 (Actuan's experience) most certainly *does* represent a red cube, E1 cannot be identical to E2 in the permanent-envatment version of the thought experiment, and thus

[9] Some have argued that considerations about a scenario in which a subject was envatted just yesterday severely weakens Putnam's (2000) own refutation of external world skepticism. See, for instance, Wright (1992).

SGP is not satisfied.[10] But as we also saw, a recently envatted Brainean *can* perceptually represent a red cube in experience. Hence, under—and I propose only under—the recent-envatment version of the thought-experiment, a computer-stimulated neuronal response in Brainean, and a genuine red cube-caused event in Actuan, can each give rise to the very same type of experience.

§ III: THE REJECTION OF THE INTERNALIST'S ARGUMENT AGAINST EXTERNALISM

BVAE is meant to establish that justification is entirely an internal matter. The original suggestion was that the possibility that someone like Brainean could be justified in holding a belief such as B1 was sufficient to reject the externalist position. I have claimed that on the condition that perceptual anti-individualism is a plausible view, the internalist can establish P1 of BVAE only if it is assumed that Brainean is a recently envatted brain. At this point, I shall argue that if the internalist proceeds on this assumption, P2 suddenly becomes a questionable premise. That is, internalism no longer seems to follow from the fact that Brainean's belief is justified. Why is this so? Recall from Section I that the internalist believes that reliability is not necessary for justification because she supposes that someone like Brainean can be justified while simultaneously remaining an unreliable believer. The considerations of the previous section now show this latter supposition to be false. Although Brainean is currently a brain in a vat and E1

[10] Presumably, B1 and B2 would also differ under this permanent-envatment version of the thought experiment (Cf. Putnam (2002) and Burge (2007a)), but this not our main concern here (though it will become relevant in Section IV.2). The more pressing issue is whether SGP is satisfied, so our attention should be on Brainean and Actuan's grounds (i.e. their experiences).

and B1 are both non-veridical, it is a mistake to think that she is unreliable. Indeed, Brainean is drastically mistaken about her surrounding environment, but many of her beliefs are nonetheless reliably produced.

In what sense are we to think of Brainean's beliefs as being reliable? A number of externalists in recent years have proposed emended theories, still in the reliabilist spirit, that attempt to explain how subjects like Brainean can be reliable believers. Goldman (1986, p. 107), for instance, argued that the appropriate domain of evaluation for whether a belief is reliably produced is relative to *normal worlds*, where he understands this to be worlds that are relevantly similar to the actual world. Evil demon victims and brains in vats live in abnormal environments, yet they can hold justified beliefs because their processes are reliable relative to normal worlds. According to Ernest Sosa's virtue theory (1991, p. 144), justification as well as reliability is relative to different environments. On Sosa's account, B1 is not justified relative to the vat environment, but it is justified relative to the actual, non-envatted environment. Finally, David Henderson and Terry Horgan have advocated what they call transglobal reliabilism, according to which "the relevant form of reliability is reliability relative to the set of *experientially possible global environments.*" (2007, p. 101) The vat environment that Brainean inhabits is but one of the many experientially possible environments in which she could have been a resident. Relative to the set of all these possible environments, the process that yields B1 is reliable, and hence B1 is justified.

My suggestion is that the externalist can get around the new evil demon problem without the need to rely on a relativized notion of reliability in the ways these and other

authors[11] have proposed. If we think of reliability as it was originally understood by the advocates of process reliabilism—according to which a belief is produced by a reliable process just in case that process has a sufficiently high truth-ratio, *regardless of the environment in which it operates*—it is evident that Brainean has satisfied even this reliability condition.

Let me explain this last point by considering what we know about Brainean in the recent-envatment version of our thought experiment. Up until very recently, she has been an ordinary person in the actual world. There had been many instances in her past in which she came into contact with red cubes. It was in virtue of these past interactions (as well as similar kinds of interactions among her ancestors) that later interactions with red cubes yielded token red cube perceptual representations. Some of those representations may have been false; there could have been some occasions in which she was confronted with an illusion or a hologram. But the *explanation* for why Brainean is capable of perceptually representing things as red cubes mentions a background of *reliable* connections between Brainean and what Brainean represented, namely red cubes. So, it is in the nature of Brainean's red cube experiences that she reliably perceives red cubes.[12] Beliefs based on those reliably produced perceptions are *ceteribus paribus* reliable as well. That is, in virtue of the fact that she lived most of her life not as a brain in a vat but as a regular embodied person, a significantly high number of Brainean's B1-like beliefs were caused by corresponding E1-like perceptual experiences *that accurately represented red cubes*. These past token beliefs were true. So even though Brainean's current belief

---

[11] Cf. Plantinga (1993), Majors and Sawyer (2005; 2007), Comesaña (2002)

[12] Many of these considerations are highlighted in Burge (2003).

that there is a red cube before her is false, past beliefs of that type have a historically successful record of being true. Thus, the process reliabilist can accommodate the intuition that Brainean's belief is justified according to this updated version of the thought experiment.

What I am suggesting is that by appealing to perceptual anti-individualism, the epistemic externalist can survive BVAE when the internalist appeals to SGP to establish P1. Even if we assume that SGP is true, internalism does not follow. Since Brainean and Actuan satisfy SGP only when Brainean is recently envatted, I have demonstrated that the reliabilist has an explanation available for why Brainean's belief is justified. This explanation is consistent with the truth of the antecedent of P2 and the falsity of consequent of P2, and hence the falsity of P2 altogether. The point is that considerations about brains in vats or evil demon victims offer us no reason to accept internalism over externalism.

§ IV: INTERNALIST RESPONSES TO THE REJECTION OF BVAE

I turn now to addressing three internalist objections to my rejection of BVAE.

IV.1: Is Brainean's belief-forming process still reliable?

My claim is that Brainean remains a reliable believer even after she enters into her envatted state. An internalist might object to what I have proposed by arguing that even though Brainean was perfectly reliable when she was a regular embodied person, once her brain is envatted, she comes to have an unreliable belief-forming process (either

immediately or a short time-period thereafter). That she is *now* an unreliable believer still does not undermine the claim that her beliefs can be justified. Hence once again, the objection goes, reliability seems not to be necessary for justification.

I find this objection unsatisfactory. Reliability theories of justification are motivated by the fact that belief-forming processes that are reliable yield justified beliefs because the inputs of such processes are *likely* to yield the truth of the contents believed in the outputs.[13] Perception and memory, for instance, are good sources of information precisely because they are truth-*conducive*. An information source may continue to present the world to the subject in a way that bears a likelihood that the world is indeed the way it is represented, even if that source misrepresents—and will continue to misrepresent—the world. Misrepresentations of this sort undermine knowledge, but they do not undermine reliability. The reliability of a process is not necessarily removed the moment one is placed in unfortunate epistemic circumstances.

The objection loses much of its force if the thought experiment was such that Brainean was envatted only for the brief moment in which she has E1 and forms B1, and she is thereafter immediately returned to her natural embodied life. At the solitary moment in which she is a brain and has E1 followed by B1, this proposed objection would suggest that she is an unreliable believer for this brief instant. Once she is placed back within her body, she becomes reliable again. However, this suggestion is implausible. If Actuan held B2 as a result of her observing a hologram of a red cube (and this was an isolated incident), we should not think that the process she utilized to acquire B2 was unreliable. It was simply erroneous; it produced a false belief on this one

---

[13] Cf. Goldman (1979)

occasion. However, these two situations are essentially analogous: in both cases, the reliability of Actuan and Brainean's belief-forming processes remains intact.

"Wait a minute," the internalist might say. "To support her own view, the reliabilist needs to deny that beliefs formed in strange, unfavorable environments are reliably produced; for otherwise she can't account for certain cases of true belief that intuitively are not knowledge." Consider Reba in fake barn country who is looking at what is in fact a real barn. Intuitively, Reba's corresponding belief that the structure before her is a barn fails to constitute knowledge, because there are numerous indistinguishable fake barns in the vicinity. To account for this intuition, as the internalist will argue, the reliabilist must admit that while Reba's perceptual processes are reliable in favorable environments, they are no longer reliable in her current environment. For, this would be the only way the reliabilist could generate the correct result that Reba's belief, although true, is not a piece of knowledge. But, if the reliabilist is forced into this position in Reba's situation, analogously shouldn't it also be the case that while Brainean's perceptual processes are reliable in her environment pre-envatment, they are unreliable in her environment after envatment?

In response to this objection, it is important for us to distinguish two kinds of reliabilism: knowledge reliabilism and justification reliabilism. The former view states, roughly, that knowledge is true belief produced by a reliable process. If this were the view I was supporting, the objection offered in the last paragraph would pose a severe problem for my rejection of BVAE. But in rejecting BVAE, I have appealed only to justification reliabilism. The justification reliabilist is free to admit that knowledge is something more than justified true belief. Indeed, one plausible assessment of Reba's

situation is that although she does not *know* that the structure before is a barn, she is nonetheless *justified* in thinking so; hence, the example describes a Gettier-style situation. If it is right that Reba's barn belief is justified yet not known, the justification reliabilist does not need to claim that it was unreliably produced. Similarly, then, this kind of reliabilist is not forced to admit that Brainean's beliefs formed in her envatted environment are unreliable.

IV.2: Permanently Envatted Brains

At this point, the internalist might concede that her position cannot be established on the basis of SGP; SGP offers us no reason to prefer epistemic internalism to externalism, since recently-envatted Brainean is a reliable believer. However, the internalist might attempt an alternative strategy for establishing her position by appeal to brains in vats. Perhaps she can argue that a *permanently* envatted Brainean is an unreliable believer, since the contents of her mental states were not individuated by reliable connections to features of the environment. Nonetheless, the internalist will argue, at least some of Brainean's beliefs are justified; and if this is the case, the fact that Brainean's (B1) belief is justified can only be explained by facts about Brainean's mind.

On this alternative version of the thought experiment, then, let us suppose that Brainean has always been a brain in a vat. Every experience she has ever enjoyed was produced by a super computer; with the exception of the super computer itself, she has had no causal interactions with external objects. Furthermore, suppose that Brainean is not a descendent of normal humans who have interacted with cubes and with instances of redness. The same goes for the programmers of the super computer. Perhaps Brainean

lives in a randomly generated universe that contains only her brain and the super computer—there were no computer programmers.[14]

Since we originally stipulated that Brainean's (E1) experience represented a red cube, and in this version of the thought experiment Brainean (as well as her ancestors, since she has none) never interacted with red things or with cubes, Brainean cannot have E1. Instead, permanently-envatted Brainean has E1*, an experience that is qualitatively identical to E1 though distinct in representational content. Of course, if Brainean has never had any of the appropriate kinds of interactions with red things or with cubes, then it is hard to see how she can represent red cubes in *thought* either.[15] So, the belief that Brainean forms on the basis of E1* cannot be B1, since B1 is the belief that there is a red cube before her. Instead, Brainean comes to believe B1*, the belief Brainean would express were she to utter in what Putnam (2000) calls vat-English: "There is a red cube before me." We must now ask two important questions: Is B1* a justified belief? And was it unreliably formed? An affirmative answer to both questions would yield the internalist's desired counterexample to reliabilism.

Given that we are now considering a situation in which the subject is, and has always been, completely shut off from a mind-independent world in the ordinary sense, it is not entirely obvious how to evaluate the scenario. Part of what complicates the issue is philosophers' disagreement over the correct characterization of the content believed in B1*. For instance, assume that an element of this belief is the concept that we will call *red cube**. We have already determined that *red cube** does not refer to red cubes

---

[14] The inspiration for this scenario arises from Putnam (2000). See also Peacocke (2004, Ch. 3).

[15] This point is demonstrated through Putnam (2002) and Burge's (2007a) famous Twin Earth thought-experiments.

(permanently-envatted Brainean is incapable of holding beliefs about red cubes), but does it refer to anything at all? If it refers, to what does it refer? If it does not refer, then how are we to answer whether or not it is justified?

In light of this complication, I think there are two possible ways that we can evaluate permanently-envatted Brainean's situation, and I will argue that each way fails to yield the internalist's intended counterexample. The first way assumes that *red cube** does refer (that is, when one has a thought that involves *red cube**, one's thought is *about* some object or entity). As for the question, "To what does this concept refer?" Putnam (2000) suggests three possible answers. Either the concept refers to 'red cubes-in-the-image', to the electrical impulses from the super computer that causally generates E1*-type experiences, or to the program feature of the super computer that is responsible for such experiences.[16] Regardless of which of these entities is the genuine referent for B1*, we can assume that all three are in operation when Brainean has E1* and forms the belief on its basis. In other words, when Brainean has E1* followed by B1*, there exists a red cube-in-the-image, and both the electrical impulses and the program feature are causally responsible for E1*. B1* accurately represents the way Brainean's world is, in that it is a representation of the relevant active features of the super computer. Hence, it is a true belief! Moreover, that E1* has the content that it has is explained by past *reliable* interactions between the red cube* program on the computer and Brainean. Therefore, if we suppose that permanently-envatted Brainean's beliefs refer, then even if beliefs like B1* are justified, Brainean is no less reliable than any non-envatted believer.

---

[16] Putnam (2000), p. 394

The second way to evaluate Brainean's situation assumes that *red cube** does not refer to anything at all. One worry for the internalist under this interpretation arises were we to construe B1* as purporting to be a demonstrative thought. That is, it may be that (from her own perspective at least) Brainean believes of some particular object that it is red (similar to a belief I would express were to assert, "That cube is red."). One of the problems with this approach is that, as Gareth Evans (1982) has argued, the conditions under which a subject entertains a thought with this sort of singular character are met only if there is, what he calls, an appropriate *information link* between the referent of the perceptual experience and the referent of the corresponding de re belief.[17] An appropriate information link is one in which the referents of the experience and belief match: that which the experience is a representation of is the same thing as what the belief is about.[18] But, as we are assuming for the moment, there is no referent associated with B1*. As Evans argues at great length, in this type of case, "…it is therefore true, in the strictest sense, that where there is no object, *there is no thought*." (Evans 1982, p. 136) We therefore cannot attribute any contentful thought to Brainean at all if B1* is construed as being singular in nature. It would hardly make any sense, then, to assess whether or not Brainean is *justified* in holding B1* if there is no belief to assess![19]

---

[17] Cf. Evans (1982) Ch's 5 and 6

[18] This does not imply that such demonstrative thoughts are necessarily true. The belief can inaccurately represent which properties the object possesses, but the subject cannot misidentify which object it is that she judges has that property. Demonstrative thoughts are immune to error through misidentification.

[19] Further difficulties arise once we attempt to determine what is expressed by Brainean's putative first-person concept in B1* (Cf. Evans (1980) Ch. 7.7). I will not address these problems here.

Whether Evans was right that a demonstrative thought of this nature depends for its existence on an associated referent is up for dispute.[20] Some would hold that demonstrative thought can occur in the absence of the relevant referent. But even if opponents of Evans' position are correct, and demonstrative thoughts need not be linked to a referent to occur, difficulty remains in trying to get a grip on the correct characterization of these thoughts. What exactly is it that subjects think, or think about, when they have such thoughts? Those who oppose Evans on this issue may at least agree with him that in these types of cases, "we would be extremely embarrassed if we had to provide an account of what it was that he thought." (Evans 1982, p. 134) The internalist, however, needs to provide such an answer if she hopes to establish that this sort of non-referring demonstrative thought is justified, yet unreliably produced. How can she assess the justificational status of this thought if she can offer no clear characterization of the thought's content? Without a proper characterization of such a thought, it is unclear how the internalist would be able to evaluate whether Brainean's doxastic behavior in forming this belief was epistemically appropriate. Thus, I suggest that this strategy cannot produce the internalist's desired counterexample either.

The foregoing discussion suggests that if the permanent-envatment version of the thought experiment can afford the internalist an example of a justified yet unreliably formed belief, B1* must be a *non-referring*, *purely descriptive* thought. Such a thought must also have a genuine content that is evaluable for truth or falsity, for otherwise questions of reliability cannot be assessed. What are the conditions under which such a belief would be true? It would be a mistake to maintain that B1* is true just in case there

[20] See Burge (2005, p. 50-53) for an objection to Evans' proposed object-dependency for demonstrative thoughts.

is a red cube before her, for we have already determined that B1* is not a belief *about* red cubes. Of course, as I have just argued, if B1* is true just in case Brainean is confronted with a red cube-in-the-image (or the red cube* computer program is activated), then the truth conditions are satisfied and Brainean is not an unreliable believer.

I myself am doubtful that such a non-referring, non-demonstrative, contentful thought can be had. It is the task of the internalist to show that I am wrong. But even if I am wrong, and there is a satisfactory account to which an internalist can appeal, it is clear that the internalist cannot simply *stipulate* that brains in vats have contentful, justified beliefs that are nonetheless unreliable.

IV.3: Durationally Envatted Brains

We have seen that the internalist faces problems when her proposed skeptical scenario involves either a recently envatted brain or a permanently envatted brain. What about a scenario that takes the middle ground? As one final objection to my argument, the internalist might suggest that she can obtain her counterexample to reliabilism by considering the skeptical scenario in which Brainean is envatted neither recently nor permanently so. Perhaps a subject who has lived a normal embodied life, but has now been envatted for a duration of several years, holds justified beliefs that are nonetheless unreliably produced?

How is this objection meant to yield the internalist's counterexample? Since the process reliabilist claims that justification is belief produced by a process that has a sufficiently high truth-ratio, the version of the thought experiment now proposed by the internalist would state that Brainean has lived long enough in her envatted state, and has

formed a large enough number of false beliefs about her environment, that her belief-producing process has fallen below the threshold for generating the number of true beliefs *sufficient* for justification. Where does this threshold for a sufficiently high truth-ratio lie? No satisfying answer has been put forth, but most reliabilists think this is a question that need not have a definitive answer. The common position is that whichever truth-ratio it is that is sufficient for justification, our trustworthy processes like perception, memory, good reasoning, and testimony possess it; while untrustworthy processes like wishful thinking and guessing lack it.[21] So how long must Brainean live in her newly envatted state until she becomes unreliable? Five years? Twenty? Thirty? The answer is perhaps indeterminate.

The fact that mental states are anti-individualistically individuated makes the situation even murkier. After some duration of having been envatted, the contents of Brainean's mental states will have shifted from making reference to mind-independent objects and properties in our normal environment to making reference to objects and properties in the vat environment (or to nothing at all). But at what point does this shift occur? Presumably, when Brainean has been envatted only for a couple of days, she still has experiences and thoughts about things like red cubes; after almost a lifetime in her envatted habitat, she now experiences and thinks about red cubes-in-the-image (or the electrical impulses sent from the super computer or the red cube* program or nothing at all). There is no clear answer as to *when* the individuation conditions of Brainean's mental states shift.

---

[21] Goldman (1979)

If, once Brainean has been envatted for a number of years, her mental states shift to being about the vat environment *before* her belief-forming processes fall below the threshold sufficient for justification, then the internalist cannot rightly maintain that Brainean's beliefs are justified yet unreliable. Before the shift occurs, Brainean would form false, yet reliably produced, beliefs about the normal, non-envatted environment (as we saw in Section III). After the shift occurs, and she starts to have "vat-environment"-type mental states, Brainean either comes to have thoughts about objects and properties in the vat environment—in which case they are reliably produced—or else her thoughts fail to refer—in which case it would be difficult to assess whether such beliefs are justified (as we saw in Section IV.2). Therefore, the internalist's best hope at providing her desired counterexample, according to this final objection, is to claim that the time at which Brainean's belief-forming processes fall below the threshold for having a sufficiently high truth-ratio *precedes* the time at which the contents of her mental state shift to making reference to features in the vat environment. If this claim is correct, then during this interlude (the period after her belief-forming processes becomes unreliable but before the contents of her mental states shift), Brainean will be capable of perceptually representing mind-independent objects, such as red cubes. And, on the basis of one such experience (E1), Brainean will come to hold the unreliably produced, yet justified, belief (B1) that there is a red cube before her.

Is such a claim correct? It is not at all obvious that there would be such a time interlude. This claim needs to be argued for; it cannot simply be assumed. The burden is once again on the internalist's shoulders to justify the claim that such an interlude is possible. But this required justification could be too far out of the internalist's reach,

given the indeterminacies surrounding the threshold below which a process becomes unreliable and also the time at which the individuation conditions for one's mental states shift.

In conclusion, the considerations of this paper have shown that Cartesian skeptical thought experiments provide no basis for rejecting reliabilist, and hence externalist, theories of justification. Contrary to what some internalists think, reliabilism is able to adequately explain why a brain in a vat holds justified beliefs.[22]

Bibliography

Burge, T. (1986). "Individualism and Psychology." *The Philosophical Review*, 95(1): 3-45.

———(2003). "Perceptual Entitlement." *Philosophy and Phenomenological Research*, 67(3): 503-548.

———(2005). "Disjunctivism and Perceptual Psychology." *Philosophical Topics*, 33(1): 1-78.

———(2007a). "Other Bodies." In T. Burge, Foundations of Mind. Oxford University Press.

———(2007b). "Cartesian Error and the Objectivity of Perception." In T. Burge, Foundations of Mind. Oxford University Press.

Comesaña, J. (2002). "The Diagonal and the Demon." *Philosophical Studies*, 100: 249-266.

Cohen, S. (1984). "Justification and Truth." *Philosophical Studies*, 46: 279-295.

Conee, E. and Feldman, R. (2001). "Internalism Defended." In H. Kornblith (ed.) Epistemology: Internalism and Externalism. Blackwell Publishers.

Dretske, F. (1995) Naturalizing the Mind. Cambridge, MA: MIT Press.

———(1996). "Phenomenal Externalism or If Meanings Ain't in the Head, Where are Qualia?" *Philosophical Issues, 7, Perception*: 143-158.

———(2000). "Entitlement: Epistemic Rights without Epistemic Duties?" *Philosophy and Phenomenological Research*, 60(3): 591-606.

Evans, G. (1982). The Varieties of Reference. Oxford: Oxford University Press.

[22] I would like to thank Tony Brueckner, Kevin Falvey, Michael Rescorla, and an anonymous referee at this journal for their helpful comments on earlier versions of this paper.

Goldman, A. (1979). "What is Justified Belief?" In G. Pappas (ed.) Justification and Knowledge. Reidel, Dordrecht.
———(1986). Epistemology and Cognition. Cambridge, MA: Harvard University Press.
Majors, B. and Sawyer, S. (2005). "The Epistemological Argument for Content Externalism." *Philosophical Studies*, 19: 257-280.
———(2007). "Entitlement, Opacity, and Connection." In S. Goldberg (ed.), Internalism and Externalism in Semantics and Epistemology. Oxford University Press.
Lehrer, K and Cohen, S. (1983). "Justification, Truth, and Coherence." *Synthese*, 55: 191-207.
Peacocke, C. (2004). The Realm of Reason. Oxford: Oxford University Press.
Plantinga, A. (1993). Warrant and Proper Function. Oxford: Oxford University Press.
Pryor, J. (2001). "Recent Highlights of Epistemology." *The British Journal for the Philosophy of Science*, 52: 95-124.
Putnam, H. (1981). Reason, Truth, and History. Cambridge: Cambridge University Press.
———(2000). "Brains in a Vat." In S. Bernecker and F. Dretske (eds.), Knowledge: Readings in Contemporary Epistemology. Oxford University Press.
———(2002). "The Meaning of 'Meaning'." In D. Chalmers (ed.) Philosophy of Mind: Classical and Contemporary Readings. Oxford University Press.
Sosa, E. (1985). "Knowledge and Intellectual Virtue." *Monist*, 68: 224-245.
———(1991). "Reliabilism and Intellectual Virtue." In Knowledge in Perspective: Selected Essays in Epistemology. Cambridge University Press.
Wedgwood, R. (1999). "The A Priori Rules of Rationality." *Philosophy and Phenomenological Research*, 59(1): 113-131.
———(2002). "Internalism Explained." *Philosophy and Phenomenological Research*, 65(2): 349-369.
Wright, C. (1992). "On Putnam's Proof That We Are Not Brains-in-a-Vat." *Proceedings of the Aristotelian Society*, 92: 67-94.