Having Hands, Even in the Vat:
What the Semantic Argument Really Shows about Skepticism

by
Samuel R Burns

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with Honors
Department of Philosophy
in The University of Michigan
2010

Advisors:

Professor Gordon Belot
Assistant Professor David Baker

”With relief, with humiliation, with terror, he understood that he also was an illusion, that someone else was dreaming him.”

Jorge Luis Borges, “The Circular Ruins”

“With your feet in the air and your head on the ground/Try this trick and spin it/ Your head will collapse/But there’s nothing in it/And you’ll ask yourself:

‘Where is my mind?’”

The Pixies

To Nami

Table of Contents

Acknowledgements

It is possible that this thesis was written in a vat (or in a computer simulation), but it was not written in a vacuum. Many people have contributed in many different ways to the finished product. I would like to take this opportunity to express my gratitude.

First, however, I should note that I am a notoriously stubborn person, and despite the help that the people listed here have provided, I have not always taken their advice. Any and all errors, omissions, confusions, and such are the result of this stubbornness, and are completely my responsibility.

When I came back to college after many years off, one of the first classes that I took was taught by Professor Corrine Painter at Washtenaw Community College. I had dabbled in philosophy since childhood, occasionally picking up Hume or Descartes, but under Professor Painter's encouragement I began to read philosophy in a much more systematic way. When I was trying (for the fourth time!) to transfer to the University of Michigan, Professor Painter put in a crucial call to the admissions department. For this I am eternally grateful.

Since beginning study at the University, several members of the philosophy department have contributed to the development of my ideas. Lina Jansson was the graduate student instructor for my course on metaphysics and epistemology. Her comments on my papers for the course were the beginnings of this project. Professor Laura Ruetsche also provided excellent teaching, guidance, and encouragement.

Professor Victor Caston taught the honors seminar in which I did the groundwork for this thesis. He held me to a much higher standard than I had expected. I thank him for the experience: This thesis is much stronger for it.

Professor Gordon Belot, my thesis advisor, provided crucial bits of advice along the way. Throughout the semester, when my brain began to wander aimlessly, lost somewhere between vats and computers and the skeptical void, Professor Belot kept gently pointing me back to earth. Professor David Baker graciously agreed to join this project at a very late stage, and his flexibility is greatly appreciated. Professor Eric Lormand read a draft of the thesis, and gave me many helpful comments. His influence is especially evident in section on environment independence.

The other honors students in the philosophy thesis seminar provided feedback that proved invaluable, and the members of the Wednesday evening philosophy happy hour group (which includes several of the aforementioned honors students) not only provided intellectual support, but also a welcome relief from hours of reading and typing. Many of the ideas in this thesis come from those sessions. Deserving special mention are Tyler Huismann, Ryan Murray, and Jason Szumanski.

I read an early version of this paper (by now nearly unrecognizable) at the 2009 Philosopher's Society conference at the University of Toronto Scarborough. Kevin Wang commented on the paper, and several of his insights found their way into the final paper. My discussion of computers and computer programs benefits from discussions with my coworker Kevin Martin. He kindly set me straight on several rather important misconceptions I had about computers.

While working on this thesis, I was also working full-time at the University of Michigan Museum of Art. I could not have completed this thesis if were not for my boss, Bruce Glazier, who accommodated my odd schedule and last minute requests for time

off. My other coworkers also put up with my zombie-like state at work and my habit of frequently launching into monologues about philosophy.

Finally, the person to whom this thesis is dedicated: My wife Nami. Her intellect is matched only by her warmth of character, and I am lucky to share this journey with her. I promise that after this, I won't bore her (much) with mumblings about brains in vats.

Samuel R Burns

April 19, 2010, Ann Arbor, MI

1. The Foundation

We think that we know things about the external world. We think that we know that we have hands, that there are trees in the park, and that tables are good for holding books. But some people have argued that we don’t really know these things. Many of their arguments start from skeptical hypotheses. Skeptical hypotheses are thought experiments that are supposed to show that in one way or another we are radically mistaken about the things we think we know about the external world. If we don’t know that we are not radically mistaken, the skeptic argues, then we don’t know that we have hands, or that there are trees in the park, or that tables are good for holding books. It might even be the case that there are no hands, no trees, no tables, and no books.

This conclusion is unsettling. It goes so strongly against our intuitions about how things are and about what kinds of knowledge we can have that it seems that there must be *something* wrong with these skeptical hypotheses. Philosophers have taken great pains to point out exactly what this something is. One influential anti-skeptical argument is derived from the work of Hilary Putnam.[1] This argument targets one specific type of total skeptical hypothesis (usually, but not always, in the form of

[1] Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981). The argument appears in the first chapter. I say that the anti-skeptical argument “is derived” from Putnam’s work because Putnam himself was not concerned with using the argument for anti-skeptical purposes.

familiar “brains in vats” scenarios), and claims that given a certain position in the philosophy of language known as *semantic externalism*, the hypothesis is self-refuting.

Putnam’s original statement of this argument caused something of a sensation when it was first published. Anthony Brueckner (among many others) adapted the argument to go against skepticism.[2] In this paper, I reexamine Brueckner’s resulting argument, and point out a major mistake in it. I will then show how semantic externalism still provides some relief from global skepticism about the external world, although in a different way than Brueckner supposes.

In the first chapter of this paper I develop some preliminary ideas that form the background of Putnam and Brueckner’s argument. First, I briefly outline the *causal theory of reference*. This theory was developed primarily by Saul Kripke in response to the traditional *descriptive theory of reference*. It essentially states that in order for an utterance of a word to refer to an object, the utterer must be part of a causal chain that leads back to an initial “baptism” of the word, which fixes the reference of the word to that object. Second, I show how the causal theory of reference leads to Putnam’s theory of semantic externalism. This theory states that the meanings and referents of words are not solely determined by internal mental states, but rather rely in certain ways on the community of language users and on features of the external world.

Chapter two is my reconstruction of Putnam and Brueckner’s semantic arguments. Putnam’s version of the argument uses semantic externalism to argue that if we were in certain total skeptical hypotheses, then we would not be able to say that we were. This trades on semantic externalism’s claim that the meanings and referents of our words

[2] Anthony Brueckner, "Brains in a Vat," *Journal of Philosophy* 83 (1986); ———, "If I Am a Brain in a Vat, Then I Am Not a Brain in a Vat," *Mind* 101, no. 401 (1992); ———, "Semantic Answers to Skepticism," in *Skepticism: A Contemporary Reader*, ed. Keith DeRose and Ted A. Warfield (Oxford: Oxford University Press, 1999).

are determined at least partially by features of our causal environment. If we are brains in vats, then, Putnam argues, our words ‘brain’ and ‘vat’ refer to things within the vat, not to what people outside the vat would call brains and vats. So our sentence ‘I am a brain in a vat’ would be false, not true.

Brueckner simplifies Putnam’s argument by introducing a disquotation principle. He argues that our sentences have disquotational truth conditions, but that brains in vats’ sentences do not. Thus, he argues, we are not brains in vats. Brueckner uses this to attack one of the premises of the skeptical argument: the premise that we do not know that we are not radically mistaken about the way things are.

In chapter three, I outline an alternative skeptical hypothesis that is immune to Brueckner’s argument. This hypothesis borrows some concepts and vocabulary from the work of Nick Bostrom and Brian Weatherson.[3] The hypothesis states that we might be nothing more than conscious simulations on a supercomputer. In the first section of the chapter, I briefly discuss computers and computer programs, and detail the structure of the hypothetical simulation. The topic of the second section is the language we would speak if we were in such a situation.

In the final chapter, I present a new skeptical argument from this hypothesis. I show how this skeptical argument is immune to Brueckner’s attack: we cannot use semantic externalism to argue that we know that we are not radically mistaken about the way things are. However, I conclude that we *can* use semantic externalism to attack another premise of the skeptical argument: the premise that if we are radically mistaken about the way things are, then we do not know that we have hands.

[3] Nick Bostrom, "Are We Living in a Computer Simulation?," *The Philosophical Quarterly* 53, no. 211 (2003); Brian Weatherson, "Are You a Sim?," *The Philosophical Quarterly* 53, no. 212 (2003).

1.1. The Causal Theory of Reference

The semantic argument against skepticism starts with a particular theory in the philosophy of language, the causal theory of reference. This theory was first widely argued by Saul Kripke.[4] (Earlier versions of the theory had been discussed by several other philosophers, especially Ruth Barcan Marcus.[5] But Kripke's version remains the most influential and widely-read account of the theory, and is the version that we will discuss here.) The theory was developed in response to the received view of the mid-twentieth century, which (like so much of the philosophy of the era) was based on the work of Gottlob Frege and Bertrand Russell.

1.1.1. The Received View: Names as Descriptions

Bertrand Russell famously claims that "common words, even proper names, are usually really descriptions. [...] The thought in the mind of a person using a proper name correctly can generally only be expressed explicitly if we replace the proper name by a description."[6] On this view, the name 'Albert Einstein' is a kind of shorthand for descriptions, such as 'the man who discovered the law of special relativity' or 'the original mad professor' or 'the winner of the 1929 Nobel Prize in physics', or some combination of these descriptions. Russell argues that the description for which a name stands can change over time, and will be different for different speakers. For instance, when a five-year old utters the name 'Albert Einstein' they mean 'the man with crazy hair whose picture is in my school book'. When that same person utters the name after years of studying the history of science, they mean 'the author of *Zur Elektrodynamik*

[4] Saul Kripke, *Naming and Necessity* (Cambridge: Harvard University Press, 1980).

[5] See Paul W. Humphries and James H. Fetzer, "Introduction," in *The New Theory of Reference: Kripke, Marcus, and Its Origins*, ed. Paul W. Humphries and James H. Fetzer, *Synthese Library* (Dordrecht: Kluwer Academic Publishers, 1999); Quentin Smith, "Marcus, Kripke, and the Origin of the New Theory of Reference," *Synthese* 104, no. 2 (1995).

[6] Bertrand Russell, "Knowledge by Acquaintance and Knowledge by Description," in *Mysticism and Logic: And Other Essays*, ed. Bertrand Russell (London: Longmans, Green and Company, 1919), 216.

bewegter Körper' or something similar. Yet all of these descriptions refer to the same person.

This view relies on the work of Gottlob Frege, who argues that a name has both a sense (*Sinn*) and a reference (*Bedeutung*).[7] (On one interpretation of Frege, what Frege calls the sense of a name is essentially a description. This seems to be the position that Kripke takes, but it is not uncontroversial.[8] However, since we are concerned here with Kripke's work, it is Kripke's reading of Frege that matters, whether or not that is the correct reading.) The names 'Phosphorus' and 'Hesperus' have different senses, but the same referent – the planet that we now call 'Venus'. This is how a sentence such as 'Phosphorus is Hesperus' can have a non-trivial meaning.

Of course, it is impossible to replace a name in every case with a single description. So on this view, a proper name substitutes for any member of a set of descriptions. Or as John Searle puts it, names function "as pegs on which to hang descriptions."[9]

1.1.2. The Theses of the Theory

On the way to refuting the descriptive theory of names, Kripke outlines what he saw as the main theses of the theory. As this outline is at once succinct and perspicacious, it is worth visiting.[10]

(1) To every name or designating expression '*X*', there corresponds a cluster of properties, namely the family of properties φ such that *A* [the utterer of the name or expression] believes 'φ*X*'.

(2) One of the properties, or some conjointly, are believed by *A* to pick out some individual uniquely.

[7] Gottlob Frege, "Sense and Reference," *The Philosophical Review* 57, no. 3 (1948).
[8] See Gareth Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982), 18-22.
[9] John R. Searle, "Proper Names," *Mind* 67, no. 266 (1958): 172.
[10] These theses are first laid out in Kripke, *Naming and Necessity*, 64-65.

(3) If most, or a weighted most, of the φ's are satisfied by one unique object *y*, then *y* is the referent of '*X*'.

(4) If the vote [of the relevant φ's] yields no unique object, then '*X*' does not refer.

(5) The statement 'If *X* exists, then *X* has most of the φ's' is known *a priori* by the speaker.

(6) The statement 'If X exists, then X has most of the φ's' expresses a necessary truth (in the idiolect of the speaker).

On this account, when I utter the name 'Barack Obama', the corresponding cluster of properties includes 'the 44th President of the United States', 'the first African-American president', 'the winner of the 2009 Nobel Peace Prize', and many other properties that I believe hold of Barack Obama. I also believe that Barack Obama is the only person who satisfies them. Thesis (3) simply states that if my belief is correct, then Barack Obama is the referent of my utterance of 'Barack Obama'. If I am mistaken, however, and there is no unique person who satisfies all of these properties, then my utterance of 'Barack Obama' fails to refer to anyone.

Let's say that I am a particularly uninformed person (perhaps someone in the distant future, who only has access to a few historical documents), and the only thing that I believe of Barack Obama is that he is the 44th President of the United States. Then the referent of my utterance of 'Barack Obama' is defined by this property, and on this theory if the person who I call 'Barack Obama' exists, then he must be the 44th President. If I understand the theory of names, then I know this *a priori*, just by thinking about the name 'Barack Obama'. Since my utterance of 'Barack Obama' means 'the 44th President of the United States', then there is no way that the person who I call Barack Obama could exist without being the 44th President of the United States.

1.1.3. The Problem

> "It really is a nice theory. The only defect I think it has is probably common to all philosophical theories. It's wrong."[11]

Let's consider the example of my utterance of the name 'Barack Obama'. According to (2), when I utter this name, I must believe that the properties that I associate with the name pick out one unique individual. However, this doesn't seem to be the way that we think about names. Let's suppose that I am even more uninformed than above, and the only property that I associate with Barack Obama is that at some time he was a president. I don't believe that he is the only person who has been a president, but it seems reasonable to think that my utterance of the name 'Barack Obama' still refers to Barack Obama. Thesis (3) is also problematic. Suppose that only thing that I believe of Barack Obama is that he is the 44th *person* to hold the office of president. (This is, of course, wrong, since Grover Cleveland is counted as both the 22nd and 24th president. So far, only 43 individual persons have been president.) So in this case, my utterance of 'Barack Obama' refers not to Barack Obama, but to whoever becomes the next president. (Perhaps then my utterance of the name 'George W. Bush' actually refers to Barack Obama, if the only thing that I believe about George W. Bush is that he was the 43rd person to hold the office of president.) But this doesn't seem to be right. Even if the only things I believe about George W. Bush and Barack Obama is that they are respectively the 43rd and 44th persons to be president, my utterances of 'George W. Bush' and 'Barack Obama' still refer to George W. Bush and Barack Obama.

It also doesn't seem that thesis (4) is right. Suppose that (through some colossal misunderstanding) I believe that Barack Obama is the first female president. Of course,

[11] Ibid., 64.

there is (at least currently) no person who matches that description. But that doesn't mean that my utterance of 'Barack Obama' fails to refer.

Even more difficulties arise for the descriptive theory of names when we consider counterfactual situations. Consider a sentence like this: 'If Barack Obama had not gone into politics, he would not have become the 44th president'. If the only thing that I believe of Barack Obama is that he is the 44th president, then on this theory my utterance is equivalent to 'If Barack Obama had not gone into politics, he would not have been Barack Obama'. But this is not how we think of counterfactual situations. We can easily imagine a situation where Barack Obama doesn't become president – his becoming president is a contingent fact.

1.1.4. Rigid Designators

Kripke calls names *rigid designators*. That is, a name refers to the same person across all possible worlds. A *non-rigid designator*, on the other hand, can refer to various persons in different possible worlds. For example, in a world where the Supreme Court decided differently, the phrase 'the 43rd President of the United States' would refer not the George W. Bush but to Al Gore. In other worlds, it might refer to Ron Paul, or Madonna, or no one at all. 'George W. Bush', however, refers to no one but George W. Bush, regardless of who won the election. This view contrasts with David Lewis's *counterpart theory*, which claims that a person is only in one possible world, but that there are similar persons (*counterparts*) in other possible worlds. When we say that there is a possible world in which George W. Bush did not win the election, we are actually saying that there is a possible world in which someone who closely resembles George W. Bush, one of his counterparts, did not win the election.[12] But Kripke argues

[12] See David K. Lewis, "Counterpart Theory and Quantified Modal Logic," *The Journal of Philosophy* 65, no. 5 (1968).

that counterpart theory is intuitively implausible. Contra Lewis, possible worlds are stipulated scenarios about *what might have been*, not real places. We can easily imagine that George W. Bush did not win the election, without having to imagine that he ceased to be George W. Bush.

1.1.5. The Baptism of Names

In response to these worries, Kripke presents a new theory of names. He begins be thinking about *how* names come to refer to objects. The name 'Barack Obama' refers to the specific person that it does because his mother named him that. Kripke calls this initial naming "baptism."[13] The baptism of a name fixes the reference of the name: Barack Obama's mother points at him and says "This baby is named 'Barack'." From then on, her utterances of the name 'Barack' refer to her son; not because the name 'Barack Obama' stands for any cluster of descriptions, but because she stipulated that the name refers to him. As she introduces her baby son to her friends and family, they learn that the name 'Barack Obama' refers to this specific individual. Years later, the name continues to spread.

There are some constraints on this spread. In order for the the reference of a name to stay fixed, it must be transmitted through appropriate causal chains. That is: "When a name is 'passed from link to link', the receiver of the name must [...] intend when he learns it to use it with the same reference as the man from whom he heard it."[14] This means that if someone hears the name 'Barack Obama' and decides to give that name to their dog, then that person's utterances of the name clearly do not refer to the 44th president. Indeed, that person can begin a new causal chain, passing the name 'Barack Obama' to others, with its reference fixed to a particular dog rather than to a particular

[13] Kripke, *Naming and Necessity*, 96.
[14] Ibid.

person. This happens with place names with some frequency. A quick search of the internet shows that within the United States alone, some forty cities, thirty counties, and one state are named some variation of 'Washington'. The original referent of that name is, of course, the first president of the United States, but now there are nearly a hundred separate causal chains of reference for this name.

Usually, the baptism of a name is "by ostension," but occasionally it is "by description."[15] Kripke cites the example of Neptune. Urbain Le Verrier fixed the reference of the name 'Neptune' before the planet itself was ever observed. Rather, Le Verrier stipulated that whatever object was causing the disturbances in the other planets' orbits should be called Neptune. This is a case of baptism by description. But future utterances of the name 'Neptune' need not have anything to do with Neptune's effects on its neighbors.

This account avoids many of the difficulties arising from the descriptive theory. A theory like this can support counterfactual uses of a name. Even if the only thing that I believe about Barack Obama is that he is the 44th president, I can still utter a meaningful sentence of the form 'If Barack Obama had not gone into politics, he would not have become the 44th president'. My utterance of 'Barack Obama' refers to Barack Obama because I am involved in a long chain of people who have borrowed the reference that Barack Obama's mother fixed, not because of any beliefs that I have about the properties attached to Barack Obama. So my utterance of 'Barack Obama' can refer to Barack Obama even if I am completely mistaken about which properties he has.

[15] Ibid.

Kripke extends this theory of reference to cover natural kinds.[16] Our word ‘tree’ refers to trees, not because we believe that there is some set of properties that trees have, but because when we were children, our parents pointed to trees and said, “There’s a tree!” Our parents acquired the word ‘tree’ in a similar way, and their parents as well. We are part of a long chain of people who use the word ‘tree’ to refer to this class of objects. Kripke argues that natural kind words are also rigid designators.[17] Once the reference of the word ‘tree’ has been fixed (whether by ostension or by description), the word refers to trees in every possible world.

1.2. Semantic Externalism

This theory of reference motivates the theory of *semantic externalism*. At the broadest level, semantic externalism is the idea that the meanings and references of our thoughts and sentences are determined by factors outside of our heads.

1.2.1. Twin Earth

In a famous paper, Hilary Putnam performs a thought experiment that he believes shows that “’meanings’ just ain’t in the head!”[18] In this thought experiment, we are asked to imagine a planet somewhere (called Twin Earth) that is almost exactly identical to Earth. There are people on Twin Earth who speak a language, Twinglish, which is grammatically and lexically identical to English. (Of course, since Twin Earth is so much like Earth, there are also Twin Earthlings who speak languages identical to

[16] Ibid., 127. Natural kinds are a controversial topic. Putnam defines natural kinds as a class of objects that shares some “essential nature.” Water is a natural kind, since water has an essential nature (H2O). See Hilary Putnam, "Is Semantics Possible?," in *Mind, Language and Reality*, ed. Hilary Putnam, *Philosophical Papers* (Cambridge: Cambridge University Press, 1979), 139-42.

[17] Kripke, *Naming and Necessity*, 134-35; see also Saul Kripke, "Identity and Necessity," in *Identity and Individuation*, ed. Milton K. Munitz (New York: New York University Press, 1972); Hilary Putnam, "It Ain't Necessarily So," *The Journal of Philosophy* 59, no. 22 (1962).

[18] ———, "The Meaning of 'Meaning'," in *Mind, Language and Reality*, ed. Hilary Putnam, *Philosophical Papers* (Cambridge: Cambridge University Press, 1979), 227.

Spanish and Japanese and Quechua). The only thing that differs between Earth and Twin Earth is that on the latter, the liquid substance that fills the oceans and rivers and swimming pools is not H_2O, as it is on Earth, but rather some substance with a different chemical structure, XYZ. Now the Twinglish speakers call this stuff 'water'. The Twin Earthlings' utterance of the water refers not to H_2O, but to XYZ. On Earth, the word 'water' refers to H_2O because some early English speaker pointed to H_2O and said, "This is 'water'." On Twin Earth, the word 'water' refers to XYZ by a similar baptism. (Of course, the way that words in a language come to refer to natural kinds is much more complicated than this, and there is rarely any single individual speaker who fixes the reference of a word. But the idea is clear enough.) If somehow we travelled to Twin Earth, then our word 'water' would still refer to H_2O, even if we mistakenly used it to refer to XYZ. If we point at a glass of XYZ and say, "There is water in the glass," then we are wrong.

Putnam points out that a few centuries ago, before modern chemistry was developed, people on Earth did not know that what they called 'water' was H_2O, and people on Twin Earth did not know that what they called 'water' was XYZ. So we can imagine two persons, one on Earth and one on Twin Earth, who have identical thoughts about water. But they each refer to something different by their utterances of the word 'water'. This shows, Putnam argues, that the reference of our words is fixed by something outside of our heads.

Putnam presents another example that proves the same point. He notes that he (like most non-experts) cannot tell the difference between an elm tree and a beech tree. Yet the extension of Putnam's word 'elm' is the set of elm trees. This is because those of us who are not experts on the the various species of trees rely on the experts to tell the difference between elms and beeches. This shows, Putnam argues, that there is a *linguistic division of labor*, which traditional semantics failed to consider. And once again, it shows that the meaning of our words relies (at least sometimes) on something external to ourselves.

1.2.2. The Meaning Vector

Putnam argues that the meaning of a word is most accurately represented by what he calls a *meaning vector*. This vector is a sequence of components that jointly determine the meaning of a word. Putnam lists four components (although he is open to the addition of others): (1) the word's syntactic markers, (2) the word's semantic markers, (3) a description of the word's *stereotype*, and (4) a description of the word's extension.

Syntactic markers include linguistic categories such as *noun* and *verb*, as well as categories such as *abstract* and *concrete*. Semantic markers are *category-indicators*, which are used to classify words into semantic classes, such as *vegetable*, *animal*, and *mineral.*

The stereotype of a natural kind word is similar to a Frege-Russell description. It consists of a set of properties that a community of speakers believes are true of paradigmatic examples of a word's referents. To borrow Putnam's example, the stereotype of 'tiger' includes that tigers are big cat-like animals, that tigers have stripes, that tigers are dangerous, and so on. We can determine whether or not a speaker has acquired a word in our idiolect by asking whether or not the speaker knows the stereotype of the word in question.

The stereotype is not analytically true of the referents of a word: albino tigers don't have stripes, but they are still tigers. Stereotypes do not determine the extension of a word, but rather are generalized descriptions of paradigmatic members of the extension.

The extension of the word consists of the referents of the word. This extension is different depending on the type of word being considered (or more accurately, on the type of object originally named by the word). If the word refers to a natural kind, then the extension consists of those objects that share the same hidden nature as the original object named by the word. If the word does not refer to a natural kind (that is, if there is no hidden nature), then the extension is determined by the superficial characteristics of the object first named. As Putnam states:

> "If there is a hidden structure, then generally it determines what it is to be a member of the natural kind, not only in the actual world, but in all possible worlds. Put another way, it determines what we can and cannot counterfactually suppose about the natural kind ('water could have all been vapor?' yes/ 'water could have been XYZ' no). But the local water, or whatever, may have two or more hidden structures - or so many that 'hidden structure' becomes irrelevant, and superficial characteristics become the decisive ones."

Putnam gives the example of jade. The word 'jade' actually refers to two different minerals (jadeite and nephrite). These two minerals have very different chemical makeups, but the same superficial characteristics. They are indistinguishable to a person who does not have a sophisticated knowledge of mineral chemistry. Yet the word 'jade' refers equally to both. The extension of the word 'jade' is all those objects that share the superficial characteristics of jadeite and nephrite, and jade is not a natural kind. Water, however, does have a hidden nature: H_2O. So water is a natural kind, and its extension is H_2O.

This extension does not need to be epistemically accessible to the users of a word for them to be considered competent speakers. For instance, speakers before the development of science successfully referred to H_2O when they uttered the word 'water', but they did not know that.

The first two of the components of a word's meaning vector are at least partially internal, but (3) is socially defined and (4) depends on features of environment. Putnam argues that this interpretation of meaning, now known as *semantic externalism*, corrects two long-standing mistakes in philosophy:

> The grotesquely mistaken views of language which are [...] current reflect two specific and very central philosophical tendencies: the tendency to treat cognition as a purely individual matter and the tendency to ignore the world[...]. Ignoring the division of linguistic labor is ignoring the social dimension of cognition; ignoring what we have called the indexicality of most words is ignoring the contribution of the environment. Traditional philosophy of language, like much traditional philosophy, leaves out other people and the world; a better philosophy and a better science of language must encompass both.[19]

These ideas – the casual theory of reference and semantic externalism – underpin the argument against skepticism that the rest of this paper will be concerned with. Internalists have presented arguments that aim to answer the worries raised by Kripke and Putnam, while preserving an internalist account of names and reference.[20] However, I will argue that even if we grant these two ideas, the anti-skeptical argument does not do what it is intended to do. So for the purposes of this paper, these fundamental ideas will go unchallenged.

[19] Ibid., 271.

[20] See, for example, John R. Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983), especially chapters 8-9.

2. The Semantic Argument

2.1. Putnam's Argument

From the position of semantic externalism discussed in the last chapter, Hilary Putnam presents an argument which he believes can rescue us from certain extreme versions of skeptical hypotheses.

Putnam outlines a special version of the standard "brains in vats" hypothesis. This hypothesis goes like this: Suppose that my brain is not in a physical body, but rather exists in a vat of nutritive fluid that keeps it alive. Some sort of machinery connected to my brain produces electronic impulses, which simulate all of my sensory experiences. I am not the only one in this situation: so are all other sentient beings. The only things that exist in this universe are brains, the vat which holds them, and the machinery which tends to the vat. The hallucination that this automatic machinery causes is collective, so that we brains can communicate with each other in a language that is lexically and grammatically identical to English (call this language *vat-English*).

This hypothesis, although intuitively implausible, seems to be completely possible. But Putnam claims that it is not. Or, he claims that if we were in this situation, we couldn't "*say* or *think* that we are."[21] It is, Putnam argues, a self-refuting utterance. Putnam outlines two ways in which a claim can be self-refuting. First, a claim is self-refuting if its truth implies its falsehood. A claim such as 'p and not-p' is self-refuting in

[21] Putnam, *Reason, Truth and History*, 7. (Italics in original).

this way. A second way in which a supposition can be self-refuting is if the mere consideration or utterance of the claim implies its falsehood. A simple example of such a claim is 'I am not thinking this sentence'. Another such claim is 'I do not exist'. As Descartes famously argued, if I am considering any claim (including this one), then I must exist. So if I think 'I do not exist', that thought must be false. Putnam claims that the supposition that we are brains in a vat is self-refuting in the second way. Our consideration or utterance of the claim implies its falsity. How Putnam reaches this conclusion is surprisingly simple.

Putnam notes that in the hypothesis, there is no appropriate causal connection between real trees and the word 'tree' in vat-English. That is, there is no way that the brains' word 'tree' could have come to refer to real trees. No brain in a vat could have ever pointed at a real tree and said, "That thing will be called a 'tree'." Neither could any brain acquire this use of the word 'tree' from another speaker. Rather, if the brains' word 'tree' has any referent, it must be something within the brains' experience (i.e., something a brain in a vat could point to and say, "That thing will be called a 'tree'."). As Putnam states:

> When a brain in a vat [...] thinks 'There is a tree in front of me', his thought does not refer to actual trees. On some theories we shall discuss it might refer to trees in the image, or to the electronic impulses that cause tree experiences, or to the features of the computer program that are responsible for those electronic impulses. [...] On these theories the brain is *right*, not *wrong* in thinking 'There is a tree in front of me.'[22]

In this passage, Putnam mentions three possible candidates for the entities that a brain in a vat's terms refer to. The first of candidate consists of objects "in the image."

[22] Ibid., 14.

Later in Putnam's discourse, this seems to be to candidate that he prefers. But this is problematic. If by "in the image" Putnam means (as Brueckner understands) "the stream of sense impressions had by the envatted subject,"[23] then this seems to mean that the truth conditions of a brain in a vats's utterances of sentences are merely that the contents of the sentences match up with the brain in a vat's phenomenal experiences. But then why can't we use the same truth conditions for our sentences? We seem to want our utterances of sentence 'There is a tree in front of me' to be true if and only if there actually is a tree in front of me, not if and only if I am having the sensory experience of there being a tree in front of me. If I am asleep in bed (in my treeless bedroom) dreaming of a tree, then my utterance of 'There is a tree in front of me' should be false. If we are going to reject a simple phenomenalist account of reference for our terms, then we should do the same for a brain in a vat's terms. So it seems that the brains in the vats should refer to something in the computer program or in the electrical impulses.

But in any case, brains in vats do not refer to real objects at all. Since the objects referred to by words in vat-English are not the same objects referred to by identical words in English, the truth conditions for identical sentences in vat-English and English are not the same. The sentence 'There is a tree in front of me' is true in English iff there is a actual tree in front of me; but the sentence 'There is a tree in front of me' is true in vat-English if and only if there is a vat-tree (let's call vat-trees *trees**) in front of me.

Putnam notes that in English, the truth condition of the sentence (S) 'I am a brain in a vat' is that I am a (real) brain in a (real) vat; but in vat-English the truth-condition of (S) is that I am a brain* in a vat*. If I am in the real world, speaking English, then I am not a brain in a vat, and (S) is false. If I am a brain in a vat, speaking vat-English, then by

[23] Brueckner, "Semantic Answers to Skepticism," 45.

the hypothesis I am not a brain* in a vat*, and S is false. If I am a brain in a vat, or if I am not a brain in a vat, my utterance of ‘I am a brain in a vat’ is false. Thus, Putnam claims, my utterance of ‘I am a brain in a vat’ is necessarily false. This argument can be schematized as follows:[24]

(a) Either I am a brain in a vat or I am not a brain in a vat.

(b) If I am a brain in a vat, then my utterances of ‘I am a brain in a vat’ are true iff I am a brain* in a vat*.

(c) If I am a brain in a vat, then I am not a brain* in a vat*.

(d) If I am a brain in a vat, then my utterances of ‘I am a brain in a vat’ are false. [(b), (c)]

(e) If I am not a brain in a vat, then my utterances of ‘I am a brain in a vat’ are true iff I am a brain in a vat.

(f) If I am not a brain in a vat, then my utterances of ‘I am a brain in a vat’ are false. [(e)]

(g) My utterances of ‘I am a brain in a vat’ are false. [(a), (d), (f)]

(h) My utterances of ‘I am not a brain in a vat’ are true. [(g)]

2.2. The Disquotation Principle

At one point in his essay, Putnam claims that his argument shows that “we are not brains in vats.”[25] This is not a straightforward conclusion of his argument, however. Recall the conclusion (h) of Putnam’s argument: “My utterances of ‘I am not a brain in a vat’ are true.” This, of course, is a metalinguistic claim. To go from this to the claim to the metaphysical claim that I am not a brain in a vat, we need to take an additional step.

[24] This schematization is borrowed from Ibid., 46-47.

[25] Putnam, *Reason, Truth and History*, 8.

To take this step, Anthony Brueckner employs the notion of disquotation.[26] (This principle is also used by Crispin Wright in his reconstruction of Putnam's argument,[27] a reconstruction that Putnam himself calls "the simplest form I know of the [...] argument."[28]) Brueckner defines disquotation as follows:

> In general, tokens of a sentence S uttered in a given object language L will have *disquotational truth conditions relative to a metalanguage L'* iff there is a true sentence of L' which consists of S surrounded by quotation marks, followed by an L'-translation of the phrase 'is true in L iff', followed by S itself.[29]

Using this notion, Brueckner introduces as a premise of the argument the following disquotation principle:

(DQP) My utterances of 'I am not a brain in a vat' are true iff I am not a brain in a vat.

With the addition of this principle, Brueckner can then make the move from the metalinguistic claim that *my utterances of 'I am not a brain in a vat' are true* to the metaphysical claim that *I am not a brain in a vat*. From here, Brueckner notes that using the disquotation principle allows us to formulate a stripped down version of Putnam's argument:

(I) If I am a brain in a vat, then my utterances of sentences have non-disquotational truth conditions and express non-disquotational contents.

[26] The disquotational principle was raised but rejected in Brueckner, "Brains in a Vat."; reconsidered in ———, "If I Am a Brain in a Vat, Then I Am Not a Brain in a Vat."; and strongly supported in ———, "Semantic Answers to Skepticism."

[27] Crispin Wright, "On Putnam's Proof That We Are Not Brains-in-a-Vat," *Proceedings of the Aristotelian Society* 92 (1992).

[28] Hilary Putnam, "Replies," *Philosophical Topics* 20, no. 1 (1992): 404 n. 29.

[29] Brueckner, "Semantic Answers to Skepticism," 47. (Italics in original).

(II) My utterances of sentences have disquotational truth conditions and express disquotational contents.

(III) I am not a brain in a vat. [(I), (II)]

This argument is clearly valid, but neither of its premises are immediately obvious. Thus, Brueckner takes care to defend each premise in turn.

2.2.1. Skepticism about the Knowledge of Content

How do I know that my utterances of sentences have disquotational truth values? An often raised criticism of semantic externalism is that it seems to engender *skepticism about knowledge of content.*[30] As Putnam outlines it, semantic externalism is the position that the meaning of our words are fixed at least in part by features of the environment external to our minds. If we are not in a position to know the features of this environment, then we it seems that we cannot know what the meanings of our words are. In other words, by denying that meaning is a purely internal matter, it seems that Putnam may have denied us privileged access to the contents of our minds. If this is the case, then if I do not yet know whether I am in a vat world or in a real world or in any other type of possible world, I don't know what kind of truth conditions my utterances of sentences have. But if I don't know what kind of truth conditions my utterances of sentences have, how can I claim to know that my utterances have disquotational truth conditions?[31]

[30] The literature on this topic is extensive, but see Sven Bernecker, "Knowing the World by Knowing One's Mind," *Synthese* 123, no. 1 (2000); Anthony Brueckner, "Scepticism About Knowledge of Content," *Mind* 99, no. 395 (1990); ———, "Trying to Get Outside Your Own Skin," *Philosophical Topics* 23, no. 1 (1995); Tyler Burge, "Individualism and Self-Knowledge," *The Journal of Philosophy* 85, no. 11 (1988); John Heil, "Privileged Access," *Mind* 97, no. 386 (1988); Ted A. Warfield, "A Priori Knowledge of the World: Knowing the World by Knowing Our Minds," in *Skepticism: A Contemporary Reader*, ed. Keith DeRose and Ted A. Warfield (Oxford: Oxford University Press, 1999).

[31] An argument towards this conclusion is found in Brueckner, "Brains in a Vat."

Brueckner argues, however, that skepticism about the knowledge of content does not actually threaten premise (II). Indeed, he argues, premise (II) is trivial.[32] To see why this is the case, we must go back to the definition of disquotation. Notice that if L and L' are the same language, then S is guaranteed to have disquotational truth conditions. That is, in the metalinguistic sentence ‘”S” is true in L iff S’, “‘S’” and ‘S’ are the sentence, making the metalinguistic sentence trivially true. Thus, as long as my metalanguage and my object language are the same, then my utterances of sentences have disquotational truth conditions relative to that language. It seems uncontroversial that this is the case: I do not suddenly switch languages in the middle of a sentence.

So, even if I cannot tell what language I am using (as the skeptic about the knowledge of content claims), I can still be confident that it is *my* language. Whatever my language is, I can be confident that my utterances of sentences have disquotational truth conditions relative to it.

2.2.2. Disquotation in the Vat

But now our attention must shift to premise (I). How do we know that a brain in a vat's utterances of sentences do not have disquotational truth conditions? Brueckner discusses this premise in some detail.[33] Brueckner asks us to consider the following sentence:

(**) My utterances of ‘I am a brain in a vat’ are true iff I am a brain in a vat.

If the mentioned sentence ‘I am a brain in a vat’ belongs to an object language which is fully contained in the metalanguage to which (**) belongs, then the sentence ‘I am a brain in a vat’ has disquotational truth conditions. But now consider the sentence

[32] The fullest exposition of this argument appears in ———, "Trying to Get Outside Your Own Skin," 99-101.

[33] This section closely follows ———, "Semantic Answers to Skepticism," 52-56.

(C) If I am a brain in a vat, then my utterances of ‘I am a brain in a vat’ are true iff I am a brain* in a vat*.

This sentence (C) is an instance of (I). If the semantic argument is sound, (I) must be true, and therefore (C) must be true. So our next task must be to evaluate (C) and see if it is, in fact, true. First, let’s break (C) down into its constituent parts. Let’s call the antecedent (P) and the consequent (Q):

(P) I am a brain in a vat.

(Q) My utterances of ‘I am a brain in a vat’ are true iff I am a brain* in a vat*.

We can easily see that (Q) and (**) cannot both be true if uttered in the same language. But as we have already seen, if the object language sentence ‘I am a brain in a vat’ is contained in the metalanguage to which (**) and (P) belong, then (**) is true. But is it possible for (Q) to correctly state the truth conditions of my envatted utterances of ‘I am a brain in a vat’? Only if the metalanguage to which (Q) belongs does not contain the object language to which the mentioned sentence belongs. So the proposition expressed by (Q) is true at a vat world only if my current metalanguage does not contain vat-English. That is, only if I am not currently at a vat world. Thus, Brueckner concludes, (C) is only true at a vat world if the actual world is not a vat world. So obviously, we cannot, without begging the question, use (C) to argue that the actual world is not a vat world.

But, Brueckner argues, if we accept semantic externalism, we must accept that (C) is true. He argues that the following strict conditional claim is a consequence of semantic externalism:

($) Necessarily, for all x, if x is a brain in a vat, then x’s utterances of ‘I am a brain in a vat’ are true iff x is a brain* in a vat*.

If ($) can be shown to follow from semantic externalism, then the proposition expressed by (C), an instantiation of ($), is a true strict conditional proposition. Thus, at all vat worlds, the proposition expressed by (Q) is true. But we have already seen that (Q) is true at a vat world only if the actual world is not a vat world. Thus, if ($) is true, then the actual world is not a vat world.

But does ($) follow from semantic externalism? Brueckner argues that on semantic externalism, ($) is exactly the type of strict conditional that determines the truth conditions of our utterances:

> Necessarily, if a being is in a treeless vat environment, then according to semantic externalism, its term refer to the entities playing the right causal role *vis-à-vis* its uses of terms. [...] Necessarily, if a being is in a normal environment, then its terms refer to the obvious candidates.[34]

Here, “vat environment” and “normal environment” are relative to my position. I may not be able to say what position that is, but whatever I am, I am not *what I call* a brain in *what I call* a vat. If I were what I call a brain in what I call a vat, then I would not call those things brains and vats.

2.3. The Semantic Argument and Skepticism

The semantic argument, as first presented by Putnam, was not meant to defeat skepticism. Rather, it was but one step in Putnam’s overall project, which is a rejection of traditional metaphysical realism.[35] Brueckner, however, is interested in the semantic argument as an anti-skeptical strategy. Indeed, he begins his discussion with a traditional *argument from skeptical hypothesis.*

[34] Ibid., 57.

[35] For a brief overview of Putnam's project, and the role that the semantic argument plays in that project, see Mark Sprevak and Christina McLeish, "Magic, Semantics, and Putnam's Vat Brains," *Studies in History and Philosophy of Science* 35 (2004).

This argument goes something like this:[36]

(i) If I know that I am standing up, then I know that I am not a brain in a vat.

(ii) I do not know that I am not a brain in a vat.

(iii) I do not know that I am standing up. [(i), (ii)]

This argument relies on the notion of *counter-possibilities*. Two propositions are counter-possibilities if the truth of one implies the falsity of the other. The two propositions *I am a brain in a vat* and *I am standing up* are considered counter-possibilities. If knowledge is closed under implication, then if I know that one or the other of these is true, then I know that the other is false. Conversely, if I do not know that one of the propositions is false, then I do not know that the other proposition is true. One strategy against this hypothesis is denying that knowledge is closed under implication.[37] However, denying closure is too counter-intuitive to be a plausible option. (We are concerned with defeating skepticism because it is so counter-intuitive. But if our anti-skeptical arguments end up being counter-intuitive as well, why not just accept skepticism and be done?)

Rather than denying closure, Brueckner’s semantic argument attacks (ii): from semantic externalism, I do know that I am not what I call a brain in what I call a vat. Even if Brueckner’s argument succeeds, it doesn’t follow that I *do* know that I am standing up. That positive argument is another task altogether. Rather, the argument tries to show that the brain in a vat hypothesis (as outlined here) doesn’t pose a threat to my knowledge that I am standing up. But some other hypothesis might. The semantic argument only has a chance of working against a limited subset of skeptical

[36] This argument appears in Brueckner, "Semantic Answers to Skepticism," 43-44.

[37] For two influential versions of this strategy, see Fred I. Dretske, "Epistemic Operators," *The Journal of Philosophy* 67, no. 24 (1970); Robert Nozick, *Philosophical Explanations* (Cambridge: Cambridge University Press, 1981).

hypotheses.[38] This subset consists of hypotheses where the subjective sensory experiences of the persons within the hypotheses are completely isolated from their external surroundings in such a way that their terms cannot ever come to refer to objects on the outside (let's call these *total skeptical hypotheses*). Thus, the argument has no chance against the hypothesis that I was kidnapped last week by a evil scientist who is keeping my brain in a vat, or against the hypothesis that I am in a coma but enjoying a lucid dream, or against the hypothesis that I am an inmate in a mental hospital suffering from acute schizophrenia. In all of these partial scenarios, my words *do* refer to objects outside of my current experience. In the first, I am *what I call* a brain in *what I call* a vat, and in all three I am not sitting at *what I call* a desk.

Brueckner recognizes this limitation, of course.[39] But he argues that, even given its limitations, the semantic argument can provide some significant relief against global skepticism. But it turns out that there is a total skeptical hypothesis that is immune to Putnam and Brueckner's semantic argument.

[38] This is an often cited problem with using Putnam's argument against skepticism; for example see Brueckner, "Brains in a Vat."; Peter Smith, "Could We Be Brains in a Vat?," *Canadian Journal of Philosophy* 14 (1984).

[39] Brueckner, "Semantic Answers to Skepticism," 59 n.10.

3. An Alternative Hypothesis

3.1. Sims and the Simulation

Suppose there is a super-computer, which is powerful enough to simulate neural networks that approximate human brains. Suppose further that the simulated network is detailed enough to support consciousness. (This hypothesis relies on a certain metaphysical position in the philosophy of mind. Nick Bostrom calls this position *substrate independence*. Substrate independence is the view that "mental states can supervene on any of a broad class of physical substrates."[40] On this position, it is at least possible that mental states could supervene on an appropriately constructed silicon chip. For the purposes of this paper, this possibility is the only thing that needs to be granted.) Say this computer simulates many conscious agents. Following Brian Weatherson,[41] let's call such a conscious simulated agent a Sim (with a capital 'S'). These Sims are causally isolated from the world outside of the computer, in such a way that the Sims' terms could not come to refer to outside objects. But the Sims have sensory experiences that are qualitatively identical to a regular human's sensory experiences. That is, they experience themselves climbing trees, sitting at tables, and talking about philosophy (Sims communicate with each other using a language,

[40] Bostrom, "Are We Living in a Computer Simulation?," 244.

[41] Weatherson, "Are You a Sim?."; Weatherson in turn borrows many ideas from Bostrom, "Are We Living in a Computer Simulation?." But I use the hypothesis in a very different way than Bostrom or Weatherson. Particularly, I am not concerned with the likelihood of this hypothesis being true, nor with how much credence we should give it.

Simglish, which is lexically and grammatically identical to English). From the inside, there is no discernible difference between the Sims' experiences and regular humans' experiences.

Although such a scenario seems unlikely to be realized, like the brains-in-vats scenario it seems to be at least physically possible. In many ways, this scenario is similar to Putnam's extreme brains-in-vats scenario: the Sims are completely isolated from the external world, and they have a "*collective* hallucination."[42] Like Putnam's scenario, this is a total skeptical hypothesis. Putnam's brains and these Sims seem to be completely deceived about the way the world is.

3.1.1. Computers and Computer Programs

Before we go through the details of the scenario, it might be helpful to think briefly about how computers and computer programs work. At the most basic level, a computer program is simply a sequence of binary data interacting with an appropriately constructed device. In a typical modern computer, this data is recorded on some sort of memory on the computer. The format of this recording can vary, since nearly any two-state medium can represent binary data. In early computers, for example, binary data was recorded on paper cards in the form of punched holes. Today, hard disk drives generally contain magnetic particles that are oriented in one of two directions, which correspond to the two binary states. This data is read by a head, which converts the magnetic binary data into electrical pulses, which in turn travel into a processor chip. This chip, which consists of a miniature electrical circuit built into a wafer of silicon, executes certain functions in response to certain sequences of binary data, and sends electrical outputs to various peripheral devices, such as a

[42] Putnam, *Reason, Truth and History*, 6. The main difference between this scenario and Putnam's is simplicity.

monitor or speakers. A single processor chip can process multiple streams of binary data at once. The processor can also send a stream of binary data back to the hard disk drive, where it is recorded as files that the processor can continue to manipulate. Modern computers can have multiple processors, each of which performs separate tasks. Of course, there are many different ways that our hypothetical super-computer could conceivably work, but let's assume that basically it works in a similar way to a modern computer. Even so, there are many different ways in which the computer might be set up. Let's briefly go through one particular scenario.

3.1.2. The Structure of the Simulation

In this scenario, a computer program is recorded onto a hard drive disk as a sequence of binary data. This program determines the total behavior of the simulated world. One central processor chip processes data from the program as well as from millions of independent inputs. These inputs come from additional processor chips (*Sim-chips*) that correspond to individual Sims. These Sim-chips simulate the organization of a human brain with a complicated pattern of simulated neurons and simulated synapses. This simulation is fine-grained enough to support consciousness. The Sim-chips also have an internal memory, which contains fragments of the data that passed through the chip.

Besides the program, the hard drive disk contains numerous data files, which correspond to inanimate objects in the simulated world. What appears to a Sim as a tree is actually a data file which has a certain pattern of data. This tree-file sends a stream of data to the central processor, where it interacts with other streams of data. If it interacts in a appropriate way with the stream of data coming from a particular Sim's processor, then that Sim has a sensory experience that is qualitatively identical to our experience of a tree. The computer, of course, is extraordinarily powerful, so the interactions between streams of data can be extraordinarily complex. Thus, the Sim can have experiences of seeing the tree, climbing the tree, and cutting the tree down. Everything in this simulated world operates according to the determined rules of program, which serve as the "laws of nature" for this world. If a Sim cuts down a tree, it falls to the ground at a predictable speed, and makes a sound that travels to other Sims at a predictable speed (although of course, we would say that all of this "falling" and "traveling" doesn't actually happen: it is all represented by patterns of binary data, and relationships between patterns).

This setup also allows for naturalistic explanations of Sim-dreams and Sim-hallucinations. When a Sim is asleep or is hallucinating, its Sim-chip can "remember" fragments of different sequences, and process them without any corresponding interaction with data files. Thus, a Sim can dream of cutting down a tree; but when she wakes up, the tree is still standing.

3.2. Simglish

3.2.1. Reference in the Simulation

What do Sims' terms refer to? Clearly, since the Sims have no appropriate causal contact with the world outside the computer, their terms cannot refer to physical objects outside of the computer. Rather, we must look for the referents of the Sims' terms within the simulation. The Sims' utterances of the word 'tree' refer most naturally to a tree-file. These files have the same causal relationship to the Sims' tree-experiences that real trees have to our tree-experiences. That is, when we are in an appropriate relationship to a real tree, we see or feel or hear a tree. When a Sim is in an appropriate relationship with a tree-file, the Sim sees or feels or hears a tree. The Simglish word 'tree' came to refer to tree-files by the same naming mechanism that caused our word ‘tree’ to refer to trees: A Sim points at the object that they see and says, “That will be called a tree.” This use of the word is then passed on from Sim to Sim. (Of course, this is simplified. Words in natural languages are rarely fixed in reference at a single time by a single person, and most words are passed from language to language. We can imagine that a Paleolithic proto-Sim pointed at a tree and uttered a sound that gradually turned into the Simglish word ‘tree’. But that is a story about etymology, not reference.) Similarly, the Sims’ utterances of the words ‘computer’, ‘brain’, ‘vat’, and so on refer to computer-files, brain-files, vat-files, and so on. This holds for any term that refers in English to physical objects.

Utterances of sentences in Simglish that are about objects in the simulated world are true or false in virtue of the state of the various data files that are referred to by the sentences. For example, a Sim’s utterance of the sentence ‘There is a hat on the table’ is true if and only if there is a hat-file and a table-file which have the appropriate relationship within the processor. The truth values of sentences like this do not rely on the contents of any Sim’s sensory experiences, but only on the state of affairs within the computer.

There is another class of words – indexicals – which do not have such a straightforward extension. Rather, the reference of an indexical is determined by an extension-function. Words in this class include such words as ‘I’, ‘here’, ‘now’ and so forth. For example, the word ‘I’ refers to whoever utters it. When uttered by a particular Sim, ‘I’ refers to that Sim. But what exactly is a Sim? It turns out that this is question is rather difficult to answer.

3.2.2. What is a Sim?

Like us, the Sims experience themselves as physical beings. They wave their hands in front of their faces and count their fingers, and they feel pain and heat and cold in physical bodies. Since in this simulation physical objects correspond to data files, we can imagine that each Sim has a data file that corresponds to it. When a Sim intends to raise its hand, a stream of data flows out of its Sim-chip and into the processor, which modifies the Sim's body-file accordingly. The body-file sends data back to the Sim-chip, so that the Sim can feel and see her hand changing position.

Some Sims study these bodies closely, and they tell the other Sims that their bodies are made out of carbon. (But of course, the Sims' utterances of the word 'carbon' refer to carbon-files.) Other Sims study the brains that are found in these bodies, and tell the other Sims that the brains are a network of neurons and synapses. (But of course, the Sims' utterances of the words 'neuron' and 'synapse' refer to neuron-files and synapse-files.)

In English, non-philosophers generally think that the word 'I' refers to the speaker's physical body. Similarly, Sims generally think that when they utter 'I', they refer to their physical bodies (what we would call their body-files). But at least in the case of the Sims, this is not exactly right. On the setup outlined above, the Sims are conscious because of processes that take place within the Sim-chips, not because of processes that take place within the body-files on the hard drive. In fact, on this set-up, the Sim's consciousness is entirely independent of their body-files. The body-files help determine the *content* of the Sims' conscious thoughts, but not the fact that the Sims are having conscious thoughts. So it seems that the files on the hard drive could be wiped clean (which on a hard drive actually means being overwritten with random data) without the Sims losing consciousness. Their 'bodies' would be gone, and their sensory experiences would be terribly chaotic and random, but as long as the program and the silicon chips remain intact they would still be conscious. Or an individual Sim's body-file could be accidentally overwritten with another file: the Sim might wake up and find herself to be a bat, or a goldfish, or a table. As disconcerting as this might be for the poor Sim, it seems that at least she would still be conscious.

It seems very strange to say that in these cases the Sims' utterances of 'I' would change reference. So the Sims' utterances of 'I' cannot refer to their body-files, but rather to something associated with their Sim-chip. (This is analogous to the problem in the philosophy of mind about what matters for personal survival. Some hold that a human's utterance of 'I' *does* refer to the human's physical body.[43] Perhaps it is so for humans, but for Sims, given the way the scenario is set up, it clearly is not the case that the Sims are referring to their body-files by uttering 'I'.)

They do not refer directly to the Sim-chip, however. The Sims' consciousness arises from a detailed simulation of a human mind. The Sim-chip, however, is merely a wafer of silicon with electrical circuits built into it. Without an electrical current, it does nothing. It merely sits inertly with the computer. However, when an electrical current begins to travel through the chip, the chip responds. If the current is a chaotic sequence of pulses, nothing extraordinary happens. It is only when the sequence of pulses takes a certain order that consciousness arises. Without the chip, the sequence of pulses is just a pattern. Without the pulses, the chip is an inert wafer of silicon. The Sims' consciousness arises from the combination of the two.[44] So it seems that the Sims' utterances of the word 'I' refer to this instantiation of a specific sequence of data on what we would call a silicon chip.

[43] For example, see Eric T. Olson, "An Argument for Animalism," in *Personal Identity*, ed. Raymond Martin and John Barresi, *Blackwell Readings in Philosophy* (Oxford: Blackwell, 2003).

[44] This story is similar to the emergent property dualist account of human consciousness, which sees consciousness as an emergent phenomena that arises from brain organization. It is by no means uncontroversial, but in this context only needs to be recognized as a possibility.

3.2.3. Environment Independence

So far, we have looked at words whose referents are different in English and in Simglish. But I believe that there are other words whose referents are stable across languages. These are words that refer to a specific class of thing. For the lack of a better term, let's call these things *environment independent*. A complete account of this class of things is beyond the scope of this paper, but we can at least begin to point some examples.

First, what does it mean for something to be environment independent? In this context, I will call something environment independent if someone can refer to it regardless of which world the person inhabits. On semantic externalism, then, trees and brains and vats all fall outside of this class. By uttering the word 'tree', a person in the actual world refers to trees, but a person in a vat world refers to vat-trees, and a person in a Sim-world refers to tree-files. But consider the number three. It seems that a person in the actual world, a person in a vat-world, and a person in a Sim-world all refer to the same thing be uttering the word 'three'. (What exactly this thing is goes well beyond the confines of this paper.) Similarly, it seems that some properties, such as 'red', should also fall into this category.

What is the difference between trees and brains on the one hand, and three and red on the other? This is a difficult question, but our inability to answer it should not be taken as evidence that there is no difference. Perhaps the answer has to do with the way we come to refer to different things. We come to refer to physical objects like trees and brains by being in proper causal relationship to them; but we come to refer to properties like three-ness by being in a proper causal relationship to objects that instantiate the property. For instance, our word 'tree' refers to trees because someone pointed to a tree and said, "That is a tree." Our word three, however, refers to the number three because someone pointed to a set of three objects and said, “These objects are three in number,” or, “This set of objects instantiates three-ness.” Perhaps I was introduced to three-ness by my mother showing me three trees, and my friend could have been introduced to three-ness by her mother showing her three apples, but nobody would say that we had been introduced to *different* three-nesses! Rather, we both have been introduced to the same three-ness; and thus my utterance of the word ‘three’ and my friend’s utterance of the word ‘three’ refer to the same thing. So why can’t we say the same for people in the vat, or in the Simulation? They may have been introduced to three-ness via three trees* or three tree-files, but the three-ness is the same.

It seems that anything that can be instantiated by various objects, and whose instantiation by an object doesn’t rely on any facts that are unique to certain worlds, is environment independent. Many mathematical terms refer to things that fall into this category. Importantly for this paper, sequences of data fall into this category as well. Remember, computer programs are sequences of binary data that are instantiated on appropriately constructed devices. For example, consider the sequence of data that is Apple Pages when instantiated on my computer. If this sequence is instantiated on a piece of paper, it would not be a functional computer program. It only when it is instantiated on the proper kind of device that it becomes a functional program. But the sequence of data itself is the same, whether it is instantiated on a computer, a piece of paper, or a paper-file in the Simulation. The sequence itself is environment independent.

3.2.4. Putnam's Question in Simglish

We are now in a position to ask the question that Putnam poses: if we were in this scenario, "could we [...] *say* or *think* that we were?"[45] It turns out that the answer to this question depends on exactly what kind of statements we try make about our situation. First, consider the following sentence:

(A) 'I am a Sim'.

As we saw above, a Sim's utterance of 'I' refers to the instantiation of a sequence of data on what those of us outside of the computer would call a silicon chip. But a Sim's utterance of the word 'Sim' (where a Sim is understood as a conscious computer simulation) does not refer to this. For the extension of the Simglish word 'silicon chip' is not the set of silicon chips, but a set of data-files that exist on the hard drive and that cause the Sims' experiences of silicon chips. So the truth condition of a Sim's utterance of the sentence 'I am a Sim' is that the Sim is a *Sim-in-the-simulation*. This truth condition is not met, and the sentence is false. So we can conclude that whatever I am, I am not what I call a Sim.

But that is not the end of the story. Consider a Sim's utterance of the next sentence:

(B) 'I am a conscious mass of carbon'.

Unlike (A), this is a sentence that the Sims would expect to come out true. But it doesn't. Let's examine the truth conditions of this sentence. The Sims' utterances of the phrase 'conscious mass of carbon' refers to body-files on the hard-drive. But again, as we saw above, the Sims' utterances of the word 'I' refer not to the Sims' body-files, but to an instantiation of data on a silicon Sim-chip. So the truth conditions for this

[45] Putnam, *Reason, Truth and History*, 7.

sentence are that the instantiation of data on the Sim-chip is identical to the body-file. But this is not the case, and (B) turns out false.

So far we have looked at two sentences in Simglish, each of which turned out to be false. Now let us consider a sentence that surprisingly comes out true if uttered in Simglish. Let's imagine a particular Sim, who happens to be interested in philosophy. This Sim is troubled that she might be mistaken about the world around her. She considers many different scenarios, including the scenario outlined above. After careful consideration, she realizes that whatever she is, she is not what she would call a Sim. But this is not enough. Considering the scenario outlined above, the Sim realizes that if she were in such a scenario, her phrase 'sequence of data' would refer to the same thing it would if she weren't in that scenario (that is, if she were a regular embodied human). But she realizes that if she were in this scenario, she wouldn't be able to refer to the silicon chip (at least not by uttering the words 'silicon chip'). So she thinks, "What if I am a conscious sequence of data instantiated on something beyond my experience, something I can't even think about? Would I be able to say that I was?"

Letting 'Data' with a capital 'D' stand for the unwieldy phrase 'a conscious sequence of data instantiated on some unnamed substance', the sentence that this Sim is considering is:

(C) 'I am Data'.

If uttered by a Sim, this sentence is true. Its truth condition is exactly the condition in which we have described the Sims existing. This sentence is immune to Brueckner's semantic argument, since the Simglish word 'Data' refers to the same thing as the identical English word (and of course the identical vat-English word, and the identical twin-English word).

Recall Brueckner's semantic argument:

(I) If I am a BIV, then my utterances of sentences have non-disquotational truth conditions and express non-disquotational contents.

(II) My utterances of sentences have disquotational truth conditions and express disquotational contents.

(III) I am not a BIV. [(I), (II)]

The argument relies on the words 'brain' and 'vat' referring to different things depending on whether they are uttered by me or are uttered by persons that I would call brains-in-vats. But 'Data' refers to the same thing whether uttered by me or uttered by what I would call 'Data'. Another way to look at this is to recall Brueckner's argument for (I). Brueckner argues that (I) follows from this general principle:

($) Necessarily, for all x, if x is a BIV, then x's utterances of 'I am a BIV' are true iff x is a BIV*.

Now replace 'BIV' with Data. But notice that Data and Data* are equivalent. So now we have:

($*) Necessarily, for all x, if x is Data, then x's utterances of 'I am Data' are true iff x is Data.

Unlike ($), this principle cannot possibly be used to support the claim that if I am Data, then my utterances have non-disquotational truth conditions. So the semantic argument fails.

4. The New Skeptical Hypothesis

Recall why Brueckner began investigating the semantic argument: he started with a skeptical argument, and began looking for a way to defeat it. Since we have now presented a scenario that avoids the semantic argument, we might now wish to present a new argument from a skeptical hypothesis. The hypothesis that this argument starts with is sentence (C) from the previous section. Recall Brueckner's argument:

(i) If I know that I am standing up, then I know that I am not a brain in a vat.

(ii) I do not know that I am not a brain in a vat.

(iii) I do not know that I am standing up. [(i), (ii)]

(i) is the claim that the two propositions 'I am standing up' and 'I am a brain in a vat' are counter-possibilities. The semantic argument tries to show that (ii) is false: I do know that I am not a brain in a vat, because my terms 'brain' and 'vat' refer to things that are not the same as what my term 'I' refers to. But now replace the phrase 'brain in a vat' with the phrase 'Data'. As we saw in the previous chapter, (B) turns out true when uttered by a Sim. So if we replace 'brain in a vat' here with 'Data', the semantic argument doesn't work. So we have stepped right back into the skeptical doubt that Brueckner wants to defeat.

Brueckner's mistake here is a simple one. His argument against the brains in vats hypothesis fixes on the fact that brains and vats are particular objects. This feature of the brains in vats hypothesis, however, is not essential. The point of the brains in vats hypothesis is not that we might really be brains in vats, but rather that we might be radically mistaken about our position in the world. Semantic considerations seem to have little chance at defeating the latter possibility. But the story is not yet over. Let's look at the resulting argument:

(a) If I know that I am standing up, then I know that I am not Data.

(b) I do not know that I am not Data.

(c) I don't know that I am standing up. [(a), (b)]

This argument is immune to Brueckner's semantic argument. But now we must look again at (a). Are the two propositions expressed here true counter-possibilities? Maybe not! Consider the truth conditions of a Sim's utterance of the sentence 'I am standing up'. If by this sentence the Sim means that her mind is 'standing up', she is of course wrong. Her mind is, as we saw, a sequence of data on a silicon chip. But the verb 'to stand up' isn't meant to apply to minds, but to bodies. The truth or falsehood of the sentence 'I am standing up' has nothing to do with the utterer's mental state, but rather with a state of affairs in the world that the utterer regularly inhabits. So a Sim's utterance of the sentence is true if and only if the Sim's corresponding body-file is in an appropriate state. A regular human's utterance of the sentence is true if and only if the regular human's body is in a certain state. So it seems that I could be what I call Data, and still be what I call standing up. (a) fails to express a true conditional.

But the skeptical hypothesis still has some force. Although a Sim may be able to say something true when she says that she is standing up, she cannot say something true when she utters (B) 'I am a conscious mass of carbon'. A Sim's mind is not what a Sim would call a conscious mass of carbon. The skeptical hypothesis shows that we could always be mistaken in our claims about what our mind is. Something like this is, I think, what Thomas Nagel is getting at in his criticism of semantic arguments against skepticism:

> Although the argument doesn't work it wouldn't refute skepticism if it did. If I accept the argument, I must conclude that a brain in a vat can't think truly that it is a brain in a vat, even though others can think this about it. What follows? Only that I can't express my skepticism by saying, "Perhaps I'm a brain in a vat." Instead I must say, "Perhaps I can't even *think* the truth about what I am, because I lack the necessary concepts and my circumstances make it impossible for me to acquire them!" If this doesn't qualify as skepticism, I don't know what does. [...] The traditional skeptical possibilities we can imagine stand for limitless possibilities that we can't imagine. In recognizing them we recognize that our ideas of the world, however sophisticated, are the product of one piece of the world interacting with part of the rest of it in ways that we do not understand very well. So anything we come to believe must remain suspended in a great cavern of skeptical darkness.[46]

Semantic considerations, however, do seem to protect our claims about what we perceive as the physical world. The possibility that I *am* massively deceived about what my mind is is not a counter-possibility to the claim that I am standing up, or that I have hands, or that I am drinking water. We could be in all sorts of weird skeptical hypotheses, and we might even be able to successfully say that we are such scenarios; but if the scenarios are total, our words and sentences can still refer to the things that cause our sensations. I might be Data, but who cares? I still have what I call hands.

[46] Thomas Nagel, *The View from Nowhere* (Oxford: Oxford University Press, 1986), 73.

Bibliography

Bernecker, Sven. “Knowing the World by Knowing One’s Mind.” *Synthese* 123, no. 1 (2000): 1-34.

Bostrom, Nick. “Are We Living in a Computer Simulation?” *The Philosophical Quarterly* 53, no. 211 (2003): 243-55.

Brueckner, Anthony. “Brains in a Vat.” *Journal of Philosophy* 83, (1986): 148-67.

———. “If I Am a Brain in a Vat, Then I Am Not a Brain in a Vat.” *Mind* 101, no. 401 (1992): 123-28.

———. “Scepticism About Knowledge of Content.” *Mind* 99, no. 395 (1990): 447-51.

———. “Semantic Answers to Skepticism.” In *Skepticism: A Contemporary Reader*, edited by Keith DeRose and Ted A. Warfield, 43-60. Oxford: Oxford University Press, 1999.

———. “Trying to Get Outside Your Own Skin.” *Philosophical Topics* 23, no. 1 (1995): 79-112.

Burge, Tyler. “Individualism and Self-Knowledge.” *The Journal of Philosophy* 85, no. 11 (1988): 649-63.

Dretske, Fred I. “Epistemic Operators.” *The Journal of Philosophy* 67, no. 24 (1970): 1007-23.

Evans, Gareth. *The Varieties of Reference*. Oxford: Oxford University Press, 1982.

Frege, Gottlob. “Sense and Reference.” *The Philosophical Review* 57, no. 3 (1948): 209-30.

Heil, John. “Privileged Access.” *Mind* 97, no. 386 (1988): 238-51.

Humphries, Paul W., and James H. Fetzer. “Introduction.” In *The New Theory of Reference: Kripke, Marcus, and Its Origins*, edited by Paul W. Humphries and James H. Fetzer, vii-xiii. Dordrecht: Kluwer Academic Publishers, 1999.

Kripke, Saul. “Identity and Necessity.” In *Identity and Individuation*, edited by Milton K. Munitz, 135-64. New York: New York University Press, 1972.

———. *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Lewis, David K. "Counterpart Theory and Quantified Modal Logic." *The Journal of Philosophy* 65, no. 5 (1968): 113-26.

Nagel, Thomas. *The View from Nowhere*. Oxford: Oxford University Press, 1986.

Nozick, Robert. *Philosophical Explanations*. Cambridge: Cambridge University Press, 1981.

Olson, Eric T. "An Argument for Animalism." In *Personal Identity*, edited by Raymond Martin and John Barresi, 318-34. Oxford: Blackwell, 2003.

Putnam, Hilary. "Is Semantics Possible?" In *Mind, Language and Reality*, edited by Hilary Putnam, 139-52. Cambridge: Cambridge University Press, 1979.

———. "It Ain't Necessarily So." *The Journal of Philosophy* 59, no. 22 (1962): 658-71.

———. "The Meaning of 'Meaning'." In *Mind, Language and Reality*, edited by Hilary Putnam, 215-71. Cambridge: Cambridge University Press, 1979.

———. *Reason, Truth and History*. Cambridge: Cambridge University Press, 1981.

———. "Replies." *Philosophical Topics* 20, no. 1 (1992): 347-408.

Russell, Bertrand. "Knowledge by Acquaintance and Knowledge by Description." In *Mysticism and Logic: And Other Essays*, edited by Bertrand Russell, 209-32. London: Longmans, Green and Company, 1919.

Searle, John R. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1983.

———. "Proper Names." *Mind* 67, no. 266 (1958): 166-73.

Smith, Peter. "Could We Be Brains in a Vat?" *Canadian Journal of Philosophy* 14, (1984): 115-23.

Smith, Quentin. "Marcus, Kripke, and the Origin of the New Theory of Reference." *Synthese* 104, no. 2 (1995): 179-89.

Sprevak, Mark, and Christina McLeish. "Magic, Semantics, and Putnam's Vat Brains." *Studies in History and Philosophy of Science* 35, (2004): 227-36.

Warfield, Ted A. "A Priori Knowledge of the World: Knowing the World by Knowing Our Minds." In *Skepticism: A Contemporary Reader*, edited by Keith DeRose and Ted A. Warfield, 76-90. Oxford: Oxford University Press, 1999.

Weatherson, Brian. "Are You a Sim?" *The Philosophical Quarterly* 53, no. 212 (2003): 425-31.

Wright, Crispin. "On Putnam's Proof That We Are Not Brains-in-a-Vat." *Proceedings of the Aristotelian Society* 92, (1992): 67-94.