Mindreading and Self-Knowledge

This paper is a back door argument for an inner sense model of introspection against pure transparency models. I use some of the parallels between self-knowledge and mindreading (non-testimonial knowledge of other persons) to argue that some form of introspection parallel to perception is required in order to acquire new self-knowledge by reflection.[1] I argue that a problem that afflicts theory theory and simulation theory accounts of mindreading in their pure forms is also problematic for accounts of self-knowledge that deny that there is a form of introspection that can be understood in terms of a perceptual analogy.

The Input Problem and Mindreading

There are two dominant camps in the academic discussion of mindreading. Theory theory claims that human beings mindread by theorizing about the unobservable mental states that could explain the behaviour one observes. Simulation theory claims

---

[1] Drawing inferences from the nature of mindreading to the nature of self-knowledge or vice versa is not unique. The most common connection drawn between mindreading and self-knowledge appears to be that an account of knowledge of others in terms of the development of a mentalistic theory can explain self-knowledge as well. For a development of this thesis, see Dennett, Daniel (1991). *Consciousness Explained*. Boston: Little, Brown, & Co. and for a sympathetic review of such proposals see chapter four of Stich and Nichols, Shaun and Stich, Stephen (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford: Clarendon Press. Simulation theorists also tend to rely on a relationship between mindreading and self-knowledge insofar as simulation theorists tend to posit a fairly traditional notion of introspection in their account of the manner in which simulation works. Cf. Goldman, Alvin (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press. For a notable exception amongst simulatin theorists, see Gordon, Robert (1995). "Simulation Without Introspection or Inference from Me to You." *Mental Simulation: Evaluations and Applications*. Eds. Martin Davies and Tony Stone. Cambridge, Mass: Blackwell. 53-67.

that one creates a working model of the mind of another person and that one feeds imaginary experiences into this model in order to mindread others.

I claim that neither theory theory nor simulation theory are adequate as they stand. Both suffer from "the input problem".[2] Each of the dominant theories cannot account for the way in which an experienced object gets classed in such a way that the machinery of theory theory or simulation theory can be brought to bear on that object. Both theories need there to be a process other than those they posit that can track the presence of persons. Both theories need something like a perceptual mechanism to bridge the gap between the mechanisms they posit and sensory stimuli in which another person figures.

For our purposes, one can suppose that one or the other of these two views of mindreading would be sufficient when coupled with a mechanism that could handle the perception of the presence of persons and some of their mental states. Such a perceptual mechanism must provide the right kind of inputs to the cognitive machinery that theory theory and simulation theory posit. What I will show is that, if this line of argument is cogent, the same line can be applied to self-knowledge in a way that militates in favor of a quasi-perceptual model of introspective self-knowledge.

For mindreading, one can formulate the input problem for theory theory as follows.

> (1) The application of general psychological principles, a folk psychology, is sufficient as an explanation of how mindreading is possible.
>
> (2) The inputs that the basic principles of folk psychology are applied to must not already be represented as being of or about a person (otherwise 1 is contradicted).

---

[2] For an in depth discussion of the input problem and the role it should play in the mindreading debate, see author 1.

(3) We do not have a good reason to think that there are basic principles of folk psychology that transform inputs that are not of or about a person into outputs that are.

Therefore, we do not have a good reason to think that theory theory can provide a sufficient explanation of mindreading, and premise one is false.

The crucial premise is the third one.

Basic folk psychological principles have to be of the form, "If there is an x that behaves in manner y, then x Φ's that z". For example, Jerry Fodor gives us the example "Ceteris paribus (i.e. in normal circumstances), people act in a way that will satisfy their desires if their beliefs are true."[3] Folk psychology has to start with some observable behaviour that can be used as an antecedent from which to infer one or more of the mental states of another person. The tricky part is this. If the mechanisms of theory theory are to be sufficient by themselves to account for mindreading, then the basic folk psychological principles need to not have any reference to the presence or mental states of an agent in the antecedents of those basic principles.

Take the principle, "If x looks hungry, then x wants food". In order to use this principle to infer that anything wants food, one needs a way of establishing that the antecedent holds. One needs to specify a value for "x" and, preferably, one that looks hungry. If this principle were a basic one, however, "x" could not already be represented as a person and "looking hungry" could not already be represented as a state that holds of agents. Otherwise, one would already be smuggling in some unexplained knowledge of the other person, namely that there exists at least one other person and that person is behaving in a certain manner. If the antecedent of this principle has mentalistic content

---

[3] Fodor, Jerry (1995). "A Theory of the Child's Theory of Mind." *Mental Simulation: Evaluations and Applications*. Eds. Martin Davies and Tony Stone. Cambridge, MA: Blackwell. 109-122. 111.

already, then one cannot use the conditional without some prior mindreading ability that can determine that the antecedent holds.

One might try to find a more fundamental candidate for a basic psychological rule. For example, one might think that the theory theorist Simon Baron-Cohen is on the right track when he posits an eye direction detector (EDD) that operates by a rule such as this, "If x looks like an eye (e.g. "has concentric circles of coloration"), then x is an eye and belongs to an agent that is aware of what is in front of that eye."[4] Baron-Cohen also posits an intentionality detector that operates by a rule such as, "If x moves and one does not detect something moving the object, then x is an agent, moving where it intends to go."[5]

The problem is that whenever one tries to find a still more basic rule than the last rule that one tried, one either finds that the rule is false/ does not plausibly correspond to actual folk psychological practice or that mentalistic content has been snuck into the antecedent. For example, one needs a way of refining ED since even a moth's wings can have concentric circles of coloration and blind persons have eyes but are not aware of what is in front of them. If, however, the origins of one's representing others as visual attenders comes from ED, it's not clear where the data is going to come from to refine ED.[6] One needs some understanding of what it means to be an attender that can be decoupled from having eyes that can be used to refine the conditional in ED. One cannot posit the identification of false positives for free. Similarly, we do not observe anything moving a magnet or a plant. If ID is the origin of representing others as agents who

---

[4] Baron-Cohen, Simon (1995). *Mindblindness*. Cambridge, MA: MIT Press. 38-39.
[5] Baron-Cohen 31-32.
[6] After all, the moth is still going to navigate its environment in a way that evidences a sensory awareness of its environment.

intend things, it is not at all clear how we could rule out magnets and plants from this class. Likewise, it is not clear how ID could be modified to account for intentionally allowing oneself to be moved by another person, as when one takes a charge in basketball or when a cat mother patiently allows its kittens to climb over it and tug its tail.

A perceptual mechanism that can unlock the antecedents of the basic principles of a folk psychology is necessary in order for the mechanisms and abilities posited by theory theory to be of any use in identifying other agents and their mental states.

An unsupplemented simulation theory also cannot plausibly do the work of mindreading. To be a complete explanation of mindreading, the simulation theory has to claim that one somehow uses a model of another person without recourse to data that includes any mental features. For, if using a mental model of another person is a complete explanation of mindreading, then no mindreading ability should be required that is not a simulation ability. Input that is not conceived of as being of or about a person is used to somehow modify one's own mind so that a mental simulation can be run.

Notice, however, that an ability to use a working model or even to assemble a model is of no use in understanding the world unless something directs the creation of the model which tracks the nature of what is modeled. A computer simulation of a chess game between two masters, for example, is dependent upon being fed inputs that correspond to the nature of the game of chess and the strategic preferences of the chess players in question. Having the ability to host or carry out a chess simulation is of no use in understanding the world of chess if there is not some other ability or mechanism that allows one to set up and run a useful simulation.

Similarly, if George rounds a corner and encounters Suzie crying, the ability to run a simulation where George imagines what he would be thinking or feeling if he was crying is of no use in understanding what Suzie is going through unless George catches on to at least two different features of his environment. George has to have a non-simulation ability to recognize Suzie as a potential candidate for simulation and must recognize the features of Suzie that need to figure in a simulation of her. George needs an ability to process sensory stimuli in a way that allows his simulation abilities to be of use.

A full defense of the claim that theory theory or simulation theory cannot plausibly get around the input problem without supplement is not possible here. What has been said should be enough to illustrate what I have in mind. In the case of both theory theory and simulation theory, the best move available to them is to supplement what they already posit with a perceptual ability to identify persons and the mental states most closely tied to their sensible behaviour. In the next section, I will argue that the input problem applies to theories of self-knowledge as well and exerts a similar pressure for the inclusion of some perception-like ability in an account of self-knowledge.

The Input Problem and Self-Knowledge

I want to focus on cases of self-knowledge where one discovers new truths about oneself through the course of one's experience, where it is the quality of one's experience which serves as the evidence for what one discovers. Thus, we will set to one side self-knowledge one gains on the basis of testimony by others, and we will not concern ourselves with whatever self-knowledge one may gain in an a priori manner (e.g. Shoemaker REF). Somewhat more subtly, I will be setting aside cases where one

discovers something about oneself de re but not de dicto.[7] Rather, I am concerned with cases like the following: one discovers that one is cold, that one likes sushi, that one believes Alberto Contador should die alone, and that one cannot really remember what it's like to be in preschool. For ease of reference, let us call the class of self-knowledge that I am gesturing at "empirical self-knowledge".

Within the philosophy of self-knowledge, there are two proposals that appear most popular for accommodating empirical self-knowledge. To put the matter crudely, one proposal is that we look inward to acquire empirical self-knowledge and the other is that we look outward. On what may be called an observational or inner sense model, empirical self-knowledge is acquired in a way that is similar to perception. How similar the two processes are taken to be can vary. One claim that I will focus on is that there is a class of introspective self-knowledge that is analogous to perception in that just as perception allows one to be aware of the presence of objects in one's environment and some of the qualities of those objects so introspection allows one to be aware of the presence of various kinds of mental states in one's mind and some of the qualities of those mental states.

In contrast, on the transparency model, one attains empirical self-knowledge not by considering oneself but by considering what one's mental states are supposed to be about. Gareth Evans gives the following well-cited articulation of this idea.

> "[I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks

---

[7] For example, if Sam hears a group of people laughing at some guy who sat in paint, Sam might thereby discover that "That guy is going to be put out when he discovers the paint on his trousers". Sam is the guy who sat in paint, but there is a kind of self-knowledge that Sam lacks in this case, at least at the time when he hears the people laughing.

> me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?'"[8]

For the transparency model, the evidence that one needs for self-knowledge comes from attending to the object or content of the mental state. The evidence that one is aware of feeling cold is one's being aware that it is cold. The evidence that one likes sushi comes from thinking about sushi as desirable, not from thinking about thinking. After all, it is odd to think of mental states as things one can take a gander at. Perhaps to think of mental states as objects that can be the object of quasi-perceptual states is to make a category mistake. Thus, self-reflection on this model is thought to be transparent in that reflection on mental objects ends up consisting in reflection on non-mental ones. As Fred Dretske puts it, the knowledge in question is "obtained—indeed can only be maintained by awareness of non-mental objects".[9]

Though this insistence on the transparency of self-reflection is where a transparency model starts, it cannot be where it ends. If I have a thought of the form "Sushi is flavorful", that thought may be evidence for predicating a liking of sushi to myself. "Sushi is flavorful" by itself is not self-knowledge, however, because it is not about me and my mental states. At best, it is knowledge about Sushi. One needs a further cognitive step that recognizes "Sushi is flavorful" or the qualia associated with sushi's flavor as evidence for my liking sushi. There must be some kind of ability that the transparency theorist can appeal to that takes self-reflections that do not involve any

---

[8] Evans, Gareth (1982). The Varieties of Reference. Oxford: Oxford University Press. 225.

[9] Dretske, Fred (1994). "Introspection." Proceedings of the Aristotelian Society 94:263-278. 264.

mental terms as an input and yields outputs that have mental terms which are indexed to oneself.

The way that Dretske and other adherents[10] to the transparency model make the connection is through an appeal to theorizing ability. Just as one might infer from someone else's (non-sarcastic) statement that sushi is flavorful that the person in question likes sushi, so one can infer from the thought content "Sushi is flavorful" that one likes sushi. One will recognize that this approach to empirical self-knowledge exactly corresponds to the theory theorist's approach to mindreading. Both attempt to explain mentalistic understanding in terms of theorizing that is brought to bear on non-mental inputs.[11]

When evaluating the transparency model, one wants to know how it is that one is making the inferences that get one self-knowledge. Without an ability to be aware of oneself as the haver of thoughts, how is one in a position to infer from "sushi is flavorful" to "I like sushi"?

---

[10] Cf. Dretske 1994. Dretske, Fred (1999). "The Mind's Awareness of Itself." Philosophical Studies 95: 103-124. Fernandez, J. (2003). "Privileged Access Naturalized." *The Philosophical Quarterly* 53: 352-372. Byrne, A (2005). "Introspection." Philosophical Topics 33: 79-104.

[11] One might also see simulation-style mechanisms appealed to in order to extend the range of application of a transparency model. For example, one might wonder whether one likes sushi. Perhaps the matter is not clear to one initially. One may then try to imagine oneself eating a nice piece of sasazushi and finds oneself thinking that sushi is very flavorful. Thus, even when the non-mental objects that might factor into self-knowledge are not present, one might think a simulation mechanism of some kind removes the pressure to appeal to some ability to perceive beliefs, desires and the like. Rather, one simply needs to be able to imagine non-mental objects. One might even be able to invoke simulation in a way that would circumvent the need for theorizing in some cases, but I don't know of a simulation story that can undergird the transparency model without help from theorizing.

Suppose that the conditional one is using is "If sushi is flavorful, then I like sushi." If we are to preclude any empirical self-knowledge that involves a quasi-perceptual awareness of one's mind, then sushi being flavorful cannot already involve an awareness of one's own mind. Otherwise, we are owed an explanation of where the self-knowledge already necessary to satisfy the antecedent of the conditional is coming from.

Insofar as one can make sense of "being flavorful" without there being any implicit reference to a mentalistic qualia, there is no reason to think that one could reasonably infer that there is a person who likes the taste of sushi and that person is oneself. Without a prior ability to be aware of one's taste sensations as belonging to oneself, it's unclear how one could infer that one likes sushi from this input. One could try to change the conditional to be investigated, searching for a more friendly option. As in the case of mindreading, however, one finds that, to the extent that one comes up with a conditional that is more plausibly true or reliable, the conditional in question is one whose antecedent already contains self-knowledge to be explained. For example, the conditional "If sushi tastes good, then I like sushi" is more plausible, but it is implausible that one can make sense of the antecedent without positing an implicit reference to oneself (aka "if sushi tastes good (to me)").[12]

[12] Furthermore, there is no reason to think that simulation ability can provide inputs that theorizing cannot. An ability to simulate one's encounters with non-mental objects is only of use if one can produce a model of oneself engaging the non-mental objects. It is unclear where one would get the information to set up a simulation on a transparency model if one wanted to make simulation and not theorizing the ground floor of one's account of self knowledge. After all, one can have many different mental attitudes towards non-mental objects. It is also not clear how one would appropriate the output state of a simulation if one is denying any ability to be aware of the mental state that the simulation produces.

Consider, for example, the phenomena of thought insertion, often associated with schizophrenia. Someone suffering from thought insertion experiences some of their thoughts as if they come from some source other than herself. Someone suffering from thought insertion can have the thought "sushi is flavorful" or "sushi tastes good", and yet clearly not be in a position to infer who, if anyone, likes sushi.[13] Notice, one need not posit that victims of thought insertion suffer a deficiency in having thoughts with normal contents nor need one posit that such persons lack inferential ability. Nevertheless, there is something missing for such persons that allows them to not identify some of their thoughts as their own and may even lead them to think the thoughts *could not* come from themselves. Most people do not suffer from thought insertion and have no problem figuring out to whom their thoughts belong. I find it most plausible that the reason for this is that most people have an awareness of their thoughts as being their own.

In conclusion, in this paper, I hope to have established that the transparency model of empirical self-knowledge suffers from a problem which its rival, the inner sense model, does not suffer from. The transparency model needs to account for the inference procedure that actually generates self-knowledge, and it needs to do so without appealing to the abilities that its rival posits.

[13] Even if someone suffering from thought insertion was able to infer that it was she herself who was having the thought, this would not be a typical case of empirical self-knowledge. The victim would be finding an indirect way of discovering what most of us seem to discover through some more direct means.