Summer 8-13-2013

The Thought Experiments are Rigged: Mechanistic Understanding Inhibits Mentalistic Understanding

Toni S. Adleberg
Georgia State University

THE THOUGHT EXPERIMENTS ARE RIGGED:
MECHANISTIC UNDERSTANDING INHIBITS MENTALISTIC UNDERSTANDING

by

TONI ADLEBERG

Under the Direction of Eddy Nahmias

ABSTRACT

Many well-known arguments in the philosophy of mind use thought experiments to elicit intuitions about consciousness. Often, these thought experiments include mechanistic explanations of a systems' behavior. I argue that when we understand a system as a mechanism, we are not likely to understand it as an agent. According to Arico, Fiala, Goldberg, and Nichols' (2011) AGENCY Model, understanding a system as an agent is necessary for generating the intuition that it is conscious. Thus, if we are presented with a mechanistic description of a system, we will be very unlikely to understand that system as conscious. Many of the thought experiments in the philosophy of mind describe systems mechanistically. I argue that my account of consciousness attributions is preferable to the "Simplicity Intuition" account proposed by David Barnett (2008) because it is more explanatory and more consistent with our intuitions. Still, the same conclusion follows from Barnett's "Simplicity" account and from my own account: we should reassess the conclusions that have been drawn from many famous thought experiments.

INDEX WORDS: Consciousness, Agency, Intuitions, Thought experiments

THE THOUGHT EXPERIMENTS ARE RIGGED:

MECHANISTIC UNDERSTANDING INHIBITS MENTALISTIC UNDERSTANDING

by

TONI ADLEBERG

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2013

THE THOUGHT EXPERIMENTS ARE RIGGED:

MECHANISTIC UNDERSTANDING INHIBITS MENTALISTIC UNDERSTANDING

by

TONI ADLEBERG

Committee Chair: Eddy Nahmias

Committee: Neil Van Leeuwen

Daniel Weiskopf

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2013

For Richard Hirsch.

ACKNOWLEDGEMENTS

There are many people in the philosophy department at Georgia State, and particularly in my cohort, who I would like to acknowledge for making the past two years so enriching and fun. I wish I could name everyone but that might be a bit much. I thank Eddy Nahmias for being supportive, challenging, and really very kind throughout my time at Georgia State. I also thank Neil Van Leeuwen, who provided me with insightful and creative suggestions, and Dan Weiskopf, who always had sharp critical feedback. For helpful conversations and feedback on earlier drafts of my thesis, I'm grateful to Mike Abramson, Cami Koepke, Morgan Thompson, and attendees of the 2013 Meeting of the Southern Society for Philosophy and Psychology. I'm thankful for the support of family (especially my parents) and my friends (especially Anna and Carlene). Special thanks to Sam Sims for being here at the end. Finally, I'd like to acknowledge my dog Tofu for challenging my work in whatever ways she can (e.g., by pawing at my keyboard).

TABLE OF CONTENTS

1 INTRODUCTION

Many well-known thought experiments in the philosophy of mind elicit intuitions about consciousness in order to show that one theory of mind is more attractive than another. These thought experiments often include a mechanistic description of a system's behavior. I argue that when we understand a system as a mechanism, we are not likely to understand it as an agent. I also argue that understanding a system as an agent is necessary for generating the intuition that it is conscious. In short, by describing a system mechanistically, the author of a thought experiment discourages readers from understanding it "mentalistically," or as having mental states. My argument could be formalized as follows:

> (1) Whether a system is described mechanistically influences our intuitions about whether that system could be conscious.
>
> (2) Whether a system is described mechanistically is irrelevant to whether that system is in fact conscious.
>
> Therefore,
>
> (3) Our intuitions about the consciousness of a system that is described mechanistically should not be taken as a reliable indication of whether that system is in fact conscious.

If our intuitions do not track the truth, it follows that they should not be invoked in arguments for particular theories of mind.

1.1 Overview

In the remainder of Section 1, I explain my use of the term "intuition:" roughly, an inclination to make a particular judgment for reasons that are unavailable to introspection. I describe four examples of thought experiments commonly used to elicit intuitions in philosophy of mind and I suggest that each includes a mechanistic description of a system.

In Section 2, I argue that we are unlikely to view a system as an agent if its behavior seems to have a mechanistic explanation, especially if we are presented with the mechanistic explanation when we first encounter the system. I discuss the factors that influence intuitions about agency and I propose that we attribute agency to a system in order to explain and predict its behavior. My proposal is supported by evidence from Heider Simmel–style experiments, an experimental philosophy study conducted by Eddy Nahmias, D. Justin Coates, and Trevor Kvaran (2007), and an fMRI study from Anthony Jack and colleagues (2006).

In Section 3, I argue that we are unlikely to view a system as conscious if we do not view it as an agent. I draw from Arico, Fiala, Goldberg, and Nichols's (2011) AGENCY Model of folk intuitions about consciousness, according to which attributing agency to a system is necessary and sufficient for being inclined to judge that the system is conscious (336).

In Section 4, I explain why people are unlikely to have the intuition that the systems described in many famous thought experiments could be conscious. Thought experiments often include mechanistic explanations that discourage readers from positing mentalistic explanations. Notice that, in these cases, the failure to attribute consciousness to a system is due to the set-up of the thought experiment rather than any feature of the system.

In Section 5, I consider an alternative account of the psychology of consciousness attributions proposed by David Barnett (2008). Whereas I hold that we are unlikely to judge that a system is conscious if we do not understand it as an agent, Barnett holds that we are unlikely to judge that a system is conscious if we do not understand it as "simple." I argue that my account is more consistent with our intuitions than Barnett's. The same conclusion follows from Barnett's account and my own: whether or not we attribute consciousness to a system is influenced by factors that are irrelevant to whether or not that system is actually conscious.

I conclude that we should reassess many well-known arguments in the philosophy of mind that appeal to intuition. While thought experiments can help us learn about the psychology of consciousness attributions, the intuitions that many of them elicit should not be taken as evidence for or against any particular theory of consciousness.

1.2 What are "Intuitions"?

I use the word "intuition" to refer, roughly, to judgments or to overridable inclinations to make judgments that are generated for reasons that we cannot access introspectively. There seems to be increasing concern among philosophers that any discussion of the epistemic status of intuition is misguided, since the word "intuition" is used to describe many different things. Intuitions are often said to comprise a "heterogeneous" category, if they comprise a psychological category at all. Jennifer Nado (2012), for example, argues that the mental states we refer to as "intuitions" are likely "generated by several fundamentally different sorts of mental processes." Some mental processes may generate reliable intuitions and others may not (1).

My aim in this paper is in line with Nado's argument. I do not argue that all intuitions are unreliable; I am concerned with one particular kind of intuition: those elicited by common thought experiments in philosophy of mind. My argument does not require that these intuitions comprise an interesting psychological category or that they are produced by any dedicated cognitive faculty. I am only concerned with showing that they have one thing in common: they are influenced by an irrelevant factor: the type of description that a thought experiment provides.

1.3 Examples of Thought Experiments

Appeals to intuition are undeniably common in the philosophy of mind. The many thought experiments that philosophers have used to elicit intuitions include Block's Nation of China and Homunculi-Headed Robot, Davidson's Swampman, Leibniz's Mill, and Searle's Chi-

nese Room. The authors of such thought experiments are not always explicit about the intended role of intuition in their arguments, but many of them seem to be making arguments like the following:

> (1) If theory of mind M (e.g., identity theory, functionalism, behaviorism, etc.) is true, then system S would be conscious.
>
> (2) Intuitively, S is not conscious.
>
> Therefore,
>
> (3) Theory of mind M is not true.

Many philosophers might claim that they use intuition to gain the reader's sympathy for some view but deny that they are making an argument like the one above that "relies upon" intuition.[1] They may take themselves to be appropriately weighing intuitions against other factors. Sometimes, it may be appropriate to weigh the intuition that S is not conscious against the merits of theory M. If a certain theory entails, for example, that a rock is conscious, it seems reasonable to weigh our strong intuition that the rock is not conscious against the merits of that theory. Even though intuitions are not infallible, one might think that they should often bear *some* weight in our reasoning.

There are some cases, though, in which we should not assign much weight to our intuitions at all. Consider, for example, moral judgments that are made during the experience of disgust. Schnall, Haidt, Clore, and Jordan (2008) report that feelings of disgust are associated with more severe moral judgments. Subjects in their study were more likely to express moral disapproval of characters in vignettes while experiencing a foul odor, a disgusting environment, after recalling a disgusting experience, or after watching a disgusting movie. Presumably, whether a moral judgment-maker experiences disgust is not relevant to how severely others should be

[1] Huebner (2010) reports that many philosophers have expressed this objection to his paper.

judged, so the judgment-maker should be more skeptical of judgments made while disgusted than she should be of typical moral intuitions.

I argue that we should be skeptical of intuitions about certain thought experiments in philosophy of mind for the same reason that we should be skeptical of moral intuitions that are affected by disgust. Some of our intuitions about consciousness are influenced by factors that are irrelevant to the target of the intuition. We must understand precisely *which* intuitions about consciousness are influenced by irrelevant factors so that we are able to treat them more skeptically. I suggest that one influential irrelevant factor is whether a system is described mechanistically. In what follows, I provide four examples of thought experiments in the philosophy of mind that include mechanistic descriptions of systems and I examine the role they seem intended to play in their respective arguments.

1.3.1 Block's Nation of China and Homunculi-Headed Robot

Ned Block (1978) asks his readers to imagine that the citizens of China coordinate to instantiate the same set of functions as the squares of a Turing Machine table. In the same paper, he asks us to imagine that billions of tiny homunculi form a functional duplicate of a human brain and that they organize to manipulate a robot body. Block's two thought experiments are intended to elicit the intuition that the Nation of China and the Homunculi-Headed Robot would lack consciousness.

Because functionalism seems to imply that the Nation of China and the Homunculi-Headed Robot would be conscious, Block argues (quite cautiously) that our intuitions cast "prima facie doubt" on functionalism (278). He claims that "[i]f there is no reason to disregard this intuition, and in the absence of any good argument for Functionalism, we are justified in rejecting Functionalism, at least tentatively" (283).

1.3.2 Davidson's Swampman

In the Presidential Address to a Pacific Division Meeting of the American Philosophical Association, Donald Davidson asked his audience to imagine a dead tree in a swamp that is struck by lightning and transformed into his exact physical duplicate (443). The Swampman, Davidson imagines, would behave exactly as the original Davidson behaves. "But," he writes, "there *is* a difference. My replica can't recognize my friends; it can't *re*cognize anything. since it never cognized anything in the first place." Davidson claims further that the Swampman wouldn't be able to remember, produce meaningful language, or "have any thoughts" (444). He takes the Swampman to demonstrate that "people who are in all relevant physical respects similar [...] can differ in what they mean or think" (452).

1.3.3 Leibniz's Mill

Leibniz asks readers to imagine that a human body is enlarged so that we might enter it as we would enter a mill. He writes that "one should, when visiting within it, find only parts pushing one another, and never anything by which to explain a perception" (*Monadology*, section 17). Leibniz uses the intuition that nothing in a Mill could explain consciousness as evidence that mental states do not have mechanistic explanations. To explain mental states, he argues that we must look for a "simple substance," not a "composite" or a "machine" (*ibid*).

1.3.4 Searle's Chinese Room

John Searle (1980) famously describes a room with an inbox, an outbox, and a very long list of instructions. He asks us to imagine that an English speaker stands in the room and receives a written statement in Chinese. By consulting the very long list of instructions, the English speaker produces an appropriate response in Chinese. From the intuition that the man in the room would not understand Chinese, Searle concludes that the ability to produce an appropriate

response does not constitute understanding. Searle's objectors have suggested that the whole system of the Chinese Room might understand Chinese, though the man inside surely would not. Searle writes that it is "absurd" to attribute understanding to the whole system. He writes: "[i]t is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible" (419). In other words, he finds it *intuitively obvious* that the Chinese Room cannot understand Chinese, which he takes as evidence that functional theories of meaning, and perhaps functional theories of mind, are false. He expects a successful theory of mind to cohere with our intuitions about the phenomenology of different kinds of systems.

1.4 "Mechanistic" Descriptions

Notice that each of the thought experiments described above includes an explanation of a system's behavior in terms of the causal roles of its physical parts. For the purposes of this paper, I refer to a system whose behavior is causally determined by its parts as a "mechanism." I refer to a compositional causal explanation of a system's behavior as a "mechanistic explanation."

Block explains the behavior of the Homunculi-Headed Robot, for example, by describing the causal roles of the robot body and of the billions of homunculi that compose the system. Similarly, he explains the behavior of the Chinese Nation by describing the causal roles of the Chinese citizens. Davidson describes the Swampman as a collection of physical particles. Leibniz asks us to imagine that a human's behavior is explained by "parts pushing one another." Finally, Searle explains the behavior of the Chinese Room by describing the causal roles of the inbox, the outbox, the list of instructions, and the English speaker inside the room.

Importantly, the way that we *first* come to understand a system seems to influence our intuitions about its mental states. We first understand the Swampman, for example, as a collection

of physical particles and we have the intuition that it could not be conscious. Davidson then imagines the Swampman interacting with his friends and he imagines that the friends will understand Swampman as conscious. The reader, however, will not come to see Swampman as conscious so easily, since the reader understands Swampman as a mechanism first.

I intend to show that the intuitions elicited by the four thought experiments described in this section reflect the fact that the target systems are described mechanistically. If the target systems were described differently, we might have the intuition that they are conscious. Since the way that a system is described is irrelevant to whether it is conscious, I argue that we should not evaluate a theory of mind based upon its coherence with our intuitions.

2 MECHANISTIC UNDERSTANDING DISCOURAGES AGENCY ATTRIBUTIONS

My claim that our intuitions about thought experiments do not track the truth about consciousness is motivated by empirical studies of the factors that influence folk attributions of mental states. In this section, I briefly discuss the factors that influence folk attributions of agency according to the AGENCY Model proposed by Arico, Fiala, Goldberg, and Nichols (henceforth, AFGN). I argue that reading a mechanistic description of a system makes us unlikely to understand that system as an agent.

2.1 What is an "Agency Attribution"?

AFGN employ a technical use of the term "agent" that may not correspond to any vernacular expression (332). They describe the concept of AGENT with which they are concerned in terms of its partial causal network. We are inclined to attribute AGENCY to objects that display very simple cues like motion, having eyes, and exhibiting contingent behavior. (I take the term "contingent behavior" to mean something like "behavior that is not explainable or predicable in terms of obvious physical causes.") Making an attribution of AGENCY to an object inclines us to anticipate goal-directed behavior and to attribute desires and intentions to the object (332).

It seems as though we attribute agency in order to explain and predict a system's otherwise inexplicable or unpredictable behavior. AFGN are careful to distinguish the concept of AGENCY with which they are concerned from more sophisticated notions of agency. AFGN's concept of AGENCY does not include, for example, the ability to engage practical reasoning or morality (332). Understanding something as an agent, in AFGN's sense, seems to mean understanding it as the type of thing that can respond to its environment and cause its own actions.

It makes sense that simple cues like motion, contingent behavior, and having eyes might trigger AFGN's concept of AGENCY, since they do seem to be reliable indications of the ability to respond to the environment. Contingent behavior, in particular, seems to be the cue that might lead us to understand a system as a cause of its own behavior. If a system is behaving in a way that is not explicable in terms of obvious physical causes, observers might naturally conclude that there is no explanation for its behavior other than the system itself. Understanding a system as an agent, in AFGN's sense, might trigger the disposition to attribute mental states to the system because the ability to respond to the environment and cause one's own actions seems to indicate the capacity for goals, intentions, and desires.

2.2 Evidence from Heider and Simmel

There is a large body of evidence indicating that there is a concept of AGENCY that plays the causal role that AFGN describe. In one well-known study of mental state attributions, Fritz Heider and Marianne Simmel (1944) presented subjects with a two and a half minute animated film of a large triangle, a small triangle, and a circle moving around in a box. When the shapes move around at non-constant velocities, apparently interacting with their environment and each other, subjects interpreted the shapes as having intentions, desires, and beliefs. When asked to describe the behavior of the shapes, at least one subject refers to them as "men" and "women" and describes them as engaged in a complex social exchange (246-7).

One concern with the evidence from Heider Simmel–style experiments is that it may not be a direct indication of subjects' intuitions. Though subjects describe the animated shapes' behavior in terms of mental states, surely they do not *actually* think that a triangle is an agent. Perhaps the subjects were using agential language metaphorically but would not actually attribute agency to the shapes.

I do not argue, of course, that subjects in Heider and Simmel's experiments actually attribute agency to animated geometric shapes. I argue only that they have an *overridable inclination* to make the judgment that the shapes are agents. According to the AGENCY Model, the overridable inclination to judge that the shapes are agents suffices to trigger another overridable inclination to judge that the shapes are conscious.

My claim that subjects have a genuine intuition that the shapes have mental states and they are not just speaking figuratively is based partly on an appeal to phenomenology. Upon watching an animation of a big triangle and a small triangle repeatedly pushing against one another and then backing away, I, for one, am immediately inclined to think of the triangles as fighting. I know, of course, that triangles do not have minds and I can override my intuition that they do, but if I were to describe the triangles as "fighting," it would not be a figure of speech. It would reflect a genuine inclination to understand the triangles as agents.

Heider and Simmel note that, of the 34 subjects who were asked to "write down what happened in the picture," just one subject used geometric rather than agential language (246). This subject used the pronouns "it" and "they" to describe the shapes and their movements. Once near the end of her report, however, she referred to one of the triangles as "he." Presuming that a subject would use the pronoun "he" to refer to an agent and "it" to refer to an object without a mind, the single use of the word "he" near the end of the relevant subject's report seems to suggest that she was inclined to think of the shapes as agents but making an effort to override this inclination. The other 33 subjects freely describe the shapes in agential language.

2.3 Evidence from Nahmias and Colleagues

The motion of the shapes in the animations presented by Heider and Simmel is presumably an important cue for agency attributions, since subjects would not likely attribute agency to

motionless shapes. Motion *alone*, however, does not seem to trigger attributions of mental states. As AFGN point out, animations of shapes moving along linear paths at constant speeds do not tend to elicit attributions of mental states (330). It seems that we attribute agency to a system *only* when its behavior does not appear to have a mechanistic explanation. The behavior of an object moving at a constant velocity is consistent with the simple mechanistic explanation that the object has been pushed and no other force has intervened in its movement. Since we can explain the object's behavior mechanistically, we do not understand it as an agent.

Daniel Dennett (1973) has suggested that considering a system as a mechanism, or adopting a "mechanistic stance" towards it, seems to discourage "any explanation in terms of beliefs, desires, and intentions" (151).[2] A study conducted by Nahmias, Coates, and Kvaran (2007) seems to support the principle that, intuitively, "mechanism conflicts with mentalism" (221). Nahmias et al. presented subjects with various scenarios describing protagonists either in mechanistic, neuroscientific terms or in non-mechanistic, psychological terms like "thoughts, desires, and plans in the agent's mind" (228, 221).

Nahmias et al. found that subjects were significantly less likely to attribute free will and moral responsibility to the protagonists of the stories after reading the mechanistic descriptions of their behavior (229). They suggest that the reason for this effect is the conflict between the mechanistic stance and the mentalistic stance (221). Just as subjects are less likely to attribute free will or moral responsibility to systems described mechanistically, I suggest that subjects would be less likely to attribute *any* conscious states to systems that are described mechanistically. Indeed, one subject explicitly cites his intuition that a mechanistic system could not be conscious as justification for why that system could not be morally responsible: "If you have no con-

[2] See discussion of this quotation in Nahmias, et al. (2007, 221).

scious control over what you do," the subject explains, "how can you be held responsible for anything you have done? [i]t wasn't you, it was the chemical reaction that did it" (214).

2.4 Evidence from Jack and Colleagues

It may be that the mechanistic stance and the mentalistic stance are difficult to maintain simultaneously because the neural underpinnings of social and mechanical reasoning are reciprocally inhibitory. Jack et al. (2006) suggest that understanding the world mechanistically and understanding the world socially are activities of two different "cognitive modes." Activity of one mode seems to inhibit activity of the other.

To determine whether social and mechanical reasoning interfere with one another, Jack et al. conducted an fMRI study in which subjects were assigned to one of four conditions. Two of the conditions required "reasoning about the mental states of other persons" as presented in either written stories or video clips. The stories described the behavior of a protagonist with a false belief and the movies portrayed a man and a woman having an emotional discussion (387-8). Subjects assigned to both social conditions were questioned about the mental states of the characters in the stories or videos. The remaining two conditions required "reasoning about the causal/mechanical properties of inanimate objects." Subjects in the mechanical conditions were also presented with either written text or video clips (385). The mechanical videos demonstrated mechanical principles from the Video Encyclopedia of Physics. The mechanical texts were adapted from science puzzles. Here is an example of a mechanical text:

> A snowmobile is cruising over plains of white, hard packed snow. The driver steers the snowmobile in a straight line while at the same time pointing a flare gun straight into the air. The driver pulls the trigger, firing a bright flare into the air. Then, the driver immediately slams on his brakes. The flare flies through the air and then lands in the snow (388).

Subjects assigned to the mechanical conditions were questioned about the mechanical principle at work or whether a specific event would happen next. For example, subjects presented with the description of the flare were asked: "Will the flare land in front of the snowmobile?" (388). As they were engaged in the social or mechanical reasoning tasks, all subjects underwent functional Magnetic Resonance Imaging (fMRI).

The fMRI results from Jack et al. support the hypothesis that there is reciprocal inhibition between the neural substrates of social and mechanical reasoning tasks. Social reasoning tasks were correlated with the activation (compared to an estimated resting baseline) of the Default Mode Network (DMN), including regions in the medial prefrontal, medial parietal/posterior cingulate, lateral parietal, and superior temporal cortices. They were also correlated with the *de*activation (again, compared to an estimated resting baseline) of the Task Positive Network (TPN), including regions in the dorso-lateral parietal and lateral prefrontal cortices. Conversely, mechanical reasoning tasks were correlated with the activation of the dorso-lateral parietal and lateral prefrontal regions and the deactivation of the medial prefrontal, medial parietal/posterior cingulate, lateral parietal, and superior temporal regions (389).

Jack et al.'s findings provide neuroscientific evidence for the claim that it is difficult to engage in mechanistic and mentalistic reasoning simultaneously. The thought experiments I describe in Section 1.3 presumably engage readers' mechanistic reasoning since are similar to the texts presented to subjects in the "Mechanical Story" condition. They describe the physical parts of a system such that a reader may predict what the system will do next by reasoning about the interaction of its parts. If mechanistic reasoning inhibits social reasoning, as the fMRI results from Jack et al. suggest, someone who reads a mechanistic description of a system will be unlikely to think of that system in terms of mental states.

3 THE AGENCY MODEL: NO CONSCIOUSNESS ATTRIBUTIONS WITHOUT AGENCY ATTRIBUTIONS

AFGN's AGENCY Model consists of a "sufficiency thesis" and a "necessity thesis." According the sufficiency thesis, attributing agency to a system inclines us to attribute consciousness to the system (336). According to the necessity thesis, attributing agency to a system is necessary for generating the intuition that the system is conscious (337).

A quick survey of the everyday intuitions we have about consciousness seems to support the AGENCY Model. We generally attribute consciousness to entities we understand as agents like humans, insects, and other animals. We generally do not attribute consciousness to entities that we do not understand as agents like plants and inanimate objects. If a fellow human's agency is compromised (e.g. because she slips into a coma), we may no longer feel confident that she is conscious. Similarly, if a potted plant began exhibiting signs of agency (e.g. by scooting towards a watering can) we might have the intuition that it is conscious.

3.1 Evidence from Arico and Colleagues

AFGN conduct a study that seems to support the AGENCY Model. They present subjects with 120 "stimulus items." Each item consisted of one entity and one property attribution. The entities included words from eight categories: Mammals, Birds, Insects, Plants, Artifacts, Vehicles, Inanimate Natural Objects, and Moving Natural Objects. Property attributions included "Feels Anger", "Feels Happy", "Hunts", "Made of Metal", "Feels Pain", "Feels Pride", "Is A Living Thing", and "Is Colored White". Subjects were given two seconds per item to accept or reject the property attributions for each entity. Their responses were timed (339).

To analyze the data from their study, AFGN collapsed subjects' attributions of the properties "Feels Anger", "Feels Happy", and "Feels Pain", since they were interested in the attribu-

tion of simple conscious states (339). They compared the responses to attributions of simple conscious states to entities in the categories of insects, vehicles, moving natural objects, and plants. Since insects display the cues for the AGENCY concept and vehicles, moving natural objects, and plants do not, AFGN predicted that subjects would be more likely to attribute simple conscious states to insects. Consistent with their hypothesis, subjects were significantly more likely to accept the ascriptions of conscious states to insects, than to entities in non-agential categories. Subjects attributed pain, happiness, or anger to insects in 70% of the trials. They attributed pain, happiness, or anger to plants in only 10% of the trials, to vehicles in 6% of the trials, and to natural moving objects in 6% of the trials (399). When subjects *rejected* the attribution of simple conscious states to insects, they had longer response times than when they rejected the attribution of simple conscious states to vehicles or natural moving objects (339).

3.2 Reconciling the AGENCY Model with Evidence from Gray, Gray, and Wegner

AFGN point out that the AGENCY Model appears to be inconsistent with some of the findings from a well-known study conducted by Heather Gray, Kurt Gray, and Daniel Wegner (2007), but the apparent conflict can be resolved (AFGN, 333-4). Whereas AFGN suggest that attributions of agency are necessary and sufficient to incline us toward attributions of consciousness, Gray, Gray, and Wegner find that subjects tend to attribute agency and experience separately and by degree.[3] They report that subjects attribute a high degree of experience but a low degree of agency to entities like frogs and fetuses. Similarly, subjects attribute a high degree of agency but a low degree of experience to entities like God and robots (619).

The apparent conflict between Gray, Gray, and Wegner's findings and the AGENCY Model is resolved because, as AFGN explain, Gray, Gray, and Wegner employ relatively strict

[3] I take it that Gray, Gray, and Wegner use the term "experience" in the same way that AFGN use the term "consciousness."

criteria for what qualifies as an agency attribution. Gray, Gray, and Wegner determine that a subject has made an agency attribution when that subject attributes the capacities for self control, morality, memory, emotion recognition, planning, communication, and thought (Gray, Gray, and Wegner, 619). AFGN, on the other hand, employ a "minimal" notion of agency; they assume that subjects make agency attributions when faced with simple cues such as motion, eyes, and contingent behavior, but they remain non-committal about the content of an agency attribution.

Presumably, subjects attribute agency to frogs and fetuses in the minimal sense intended by AFGN but not in the stricter sense intended by Gray, Gray, and Wegner (334). Since frogs and fetuses are likely to be viewed as agents in the minimal sense, the intuition that they are conscious is consistent with the AGENCY Model.

AFGN do not discuss how they might account for subjects' attributions of mental states to God. Subjects in Gray, Gray, and Wegner's study make attributions to God based upon the following description:

> *God*. Many people believe that God is the creator of the universe and the ultimate source of knowledge, power, and love. However, please draw upon your own personal beliefs about God (7).

It seems to me that while the above description may elicit attributions of capacities like "morality" or "planning," it may not trigger the concept of AGENCY in the sense that AFGN intend. Perhaps if subjects had been presented with stories or movie clips in which a God character exhibits contingent behavior and motion, the relevant notion of agency would be triggered and the intuition that he has experience would follow.

According to the findings from Gray, Gray, and Wegner, subjects also attribute a moderate degree of agency but a very low degree of consciousness to robots. Robots seem to be a counterexample to AFGN's AGENCY Model because they may exhibit all of the signs of agency yet they are commonly understood to lack consciousness. I suggest, however, that the case of

robots is deceptive because our existing, general understanding of robots might lead us to override our intuitions upon encountering a particular robot.

Most of us understand that robots are mechanisms, which inclines us to think that they lack consciousness. Suppose I were to run into R2-D2 from *Star Wars* at my local coffee shop. I might feel strongly compelled to understand R2-D2 as agent because of its movement, its contingent beeping, and its camera that functions as an eye. However, I would likely classify R2-D2 as a robot because of its similarities to other robots I have seen (for example, it is made of metal). Because I believe that robots are mechanisms, I would likely conclude that R2-D2 is a mechanism and lacks consciousness, despite that its functional components are not immediately evident. Now suppose that the barista at the coffee shop had somehow never encountered or even heard of a robot before. In line with the AGENCY Model, I suggest that the barista would likely attribute agency and consciousness to R2-D2, since no mechanistic explanation of R2-D2 would be immediately apparent.

3.3 The Expertise Objection

Since the thought experiments in philosophy of mind are used to elicit *philosophers'* intuitions about consciousness, it may seem that studies of "folk" attributions of mental states are irrelevant. Timothy Williamson (2011), for instance, argues that philosophers are experts at thought experimentation just as scientists are experts at scientific experimentation. He would likely claim that principles gleaned from the folk psychology of consciousness attributions do not apply to philosophers. Perhaps philosophers are trained to consider only the relevant information when making counterfactual judgments.

I am not aware of any evidence that philosophers have different intuitions about consciousness than non-philosophers or that they arrive at those intuitions by a different process. It

seems to me that the burden of proof is on those who argue that philosophers have intuitions about consciousness that are more reliable than the intuitions of the folk. Even if experience with thought experiments does positively affect the reliability of philosophers' intuitions, there may be other factors that negatively affect the reliability of philosophers' intuitions. Philosophers, for example, are likely to be committed to various metaphysical positions that may bias them towards particular judgments about consciousness. Henceforth, I assume that what is true of the psychology of folk intuitions about mental states is also true of the psychology philosophers' intuitions about mental states.

If the sufficiency thesis of the AGENCY Model is correct, the intuition that a system could be conscious often results from the judgment that it is an agent. If the necessity thesis of the AGENCY Model is correct, the intuition that a system could *not* be conscious may often result from the judgment that it is *not* an agent. In other words, we will likely judge that a system is conscious when and only when we must view it as an agent in order to understand its behavior.

4 THE THOUGHT EXPERIMENTS ARE RIGGED

We can now see why so many well-known thought experiments in philosophy of mind discourage us from making consciousness attributions. To have the intuition that a system is conscious, we must understand it as an agent. To understand a system as an agent, its behavior must not appear to be explained mechanistically. Many philosophical thought experiments, including the Nation of China, the Homunculi-Headed Robot, the Swampman, Leibniz's Mill, and the Chinese Room, explicitly describe systems as mechanisms and encourage people to think of the systems in terms of the interaction of their parts rather than as agents.

Since the Nation of China system, for example, is described mechanistically, we do not understand it as an agent and hence do not understand it as conscious. Since the Chinese Room is described in terms of the functions of its physical parts, we do not understand it as an agent and hence do not understand it as conscious. Searle devised the Chinese Room in order to probe our intuitions about whether the system can understand language, not about whether the Chinese Room can be conscious. It seems plausible to me, though, that the reason we have the intuition that the Chinese Room cannot understand language is that we have the intuition that it is not conscious.[4] Intuitively, understanding seems to require consciousness.

Consider the following analogy. One might think of the cognitive system for reasoning about mechanisms as a microphone and the cognitive system for reasoning about mental states as an infrared light detector. If we suppose there is only one power source for the microphone and the infrared light detector, then the two systems are reciprocally inhibitory. When the microphone is on, the infrared light detector must be off, and vice versa. The mechanistic descriptions included in the thought experiments I describe activate our system for reasoning about mecha-

[4] Chalmers (1996) has also argued that, while Searle intended the Chinese Room as a thought experiment about intentionality, "it is fairly clear that consciousness is at the root of the matter" (301).

nisms just as a noise might lead us to turn on the microphone. When the microphone is on, we will not detect any infrared light. We should not conclude, however, that there *is* no infrared light. We should not draw any conclusions about the presence of infrared light when we are not using the infrared light detector. Similarly, when we reason mechanistically about a system, we will likely not detect consciousness but we should not draw any conclusions about whether the system is conscious.[5]

4.1 The Argument from Irrelevant Factors

My argument that our intuitions about consciousness do not track the truth if they are elicited by thought experiments that include mechanistic descriptions can be framed as an "argument from irrelevant factors." Arguments from irrelevant factors have a form relevantly similar to the following:

(1) Some factor, F, influences some subset of intuitions about topic T.

(2) F is irrelevant to the truth regarding T.

Therefore,

(3) The subset of intuitions that are influenced by F should not be taken as an indication of the truth regarding T.

Arguments from irrelevant factors have been employed in many other areas of philosophy. In moral psychology, for example, Joshua Greene (2000) has generated a great deal of discussion by arguing that our moral intuitions are unreliable because they track morally irrelevant factors like whether a harmful action is personal or impersonal.[6] Greene argues that we are more likely to have the intuition that harmful actions are wrong when they are personal rather than impersonal. Since the personal/impersonal distinction appears to be morally irrelevant, Greene

[5] I thank Neil Van Leeuwen for suggesting this analogy.
[6] Greene's argument was framed as an "argument from irrelevant factors" and criticized by Selim Berker (2009, 321).

concludes that we should be skeptical of our intuitions that personal harmful actions are wrong and that impersonal harmful actions are not wrong.

Arguments from irrelevant factors are prevalent in the history of philosophy. At the end of the nineteenth century, for example, the Michelson-Morely experiment led scientists to theorize that space is relative. Though relativity is counterintuitive, Henri Poincaré (1914) argued that human spatial cognition evolved for its adaptive value rather than its accuracy. He concluded that we should disregard our spatial intuitions and trust the predictions of the best available theory. In other words, Poincaré argued that our intuitions about space are unreliable because they are influenced a truth-independent factor: adaptivity.

The argument from irrelevant factors I advance here might be stated as follows:

> (1) Whether a system is described mechanistically influences our intuitions about whether that system could be conscious.
>
> (2) Whether a system is described mechanistically is irrelevant to whether that system is in fact conscious.
>
> Therefore,
>
> (3) Our intuitions about the consciousness of systems that are described mechanistically should not be taken as an indication of the truth about whether the system is in fact conscious.

It may be that any system, conscious or not, can be described such that its behavior appears to be explained mechanistically. Even a human can be described as a collection of cells whose behavior is largely caused by the activity of a hundred billion neurons. Or, as Leibniz asks us to imagine, a human can be considered as a machine. Though we would typically attribute agency and consciousness to a human, it is difficult to do so when we are considering that human as a mechanism, or as a system of interacting parts.

Just as it is possible to describe a human such that her behavior seems to have a mechanistic explanation, it may be possible to describe the systems in thought experiments such that their behavior does not have an obvious mechanistic explanation. When systems are not described mechanistically, we might have the intuition that they are conscious. Suppose that Searle had described a room that could converse in Chinese. Suppose he described its conversations with various interlocutors. Had he left out the mechanistic explanation of its behavior, it might have been easier to imagine that the room could have mental states.

Leibniz's Mill and Davidson's Swampman contrast with one another helpfully. Leibniz asks us to consider a human as a collection of parts while Davidson asks us to consider a collection of parts as a human. That is, *after* describing the Swampman mechanistically, as a collection of molecules, Davidson imagines that the Swampman moves into his house, starts writing articles on radical interpretation, and appears to recognize and converse with his friends. Davidson writes that anyone who interacts with the Swampman without knowing its mechanistic description would not be able to tell the difference between the Swampman and Davidson himself. Only when we consider the Swampman as a collection of parts do we have the intuition that it cannot be conscious. Now recall that Leibniz asks us to consider a *human* as a collection of parts as well and that, when we do so, it becomes difficult to consider the human as conscious. The Mill and the Swampman illustrate that our intuitions about consciousness track whether or not we are considering a system as a mechanism. They could not be tracking whether or not the systems are actually conscious, unless humans lose consciousness when we think of them as machines and Swampmen become conscious when we interact with them as would interact with a human.

The intuitions elicited by many thought experiments, then, are often determined by the psychology of mental state attributions and by the way that systems are described. They do not reflect whether the described systems are actually likely to be conscious or even whether we would judge them to be conscious under a different set of circumstances or a different description.

4.2 Suggestions for Future Research

It may be possible to test the hypothesis that mechanistic descriptions inhibit consciousness attributions empirically. AFGN's study, described in section 3.1, serves as a model. Subjects should be presented with stimulus pairs: one entity and one property attribution. The entities should be systems like the Nation of China, the Homunculi-Headed Robot, The Swampman, the Mill, and the Chinese Room. The property attributions should include simple conscious states like "feels happy" and "feels pain" as well as non-mental properties like "can walk" or "can speak."

Subjects should receive one of two treatments. One group of subjects should be encouraged to think of the systems mechanistically and the other should be encouraged to think of the systems non-mechanistically. In order to ensure that both groups of subjects are making judgments about the same hypothetical systems, it seems to me that both groups would have to be presented with the same mechanistic descriptions of the systems. All subjects, for example, should read a description of the Chinese Room similar to the one that Searle originally provides. Researchers might encourage one group to consider the systems mechanistically and the other group to consider the systems non-mechanistically through the use of images accompanying the written descriptions. To encourage subjects to consider the Chinese Room mechanistically, for example, the mechanistic description could be accompanied by images depicting the person in-

side the room consulting the list of instructions. To encourage subjects to consider the Chinese Room non-mechanistically, the mechanistic description could be accompanied by images of the room from the outside, perhaps as it communicates with Chinese speakers.

The AGENCY Model of consciousness intuitions predicts that subjects in the purely mechanistic group will not attribute conscious states to systems. Subjects encouraged to think of the systems non-mechanistically, however, will be much more likely to attribute conscious states to the systems. When they do reject the consciousness attributions, their response times of subjects in the non-mechanistic group will be longer. The need to present all subjects with the mechanistic descriptions of the systems may lessen the difference in responses between the two groups, but we still should expect some difference. After all, even those who believe that humans have true mechanistic descriptions will attribute consciousness to humans when thinking of them non-mechanistically. Similarly, I would expect that subjects would attribute consciousness to systems when thinking of them non-mechanistically even if they are given a mechanistic description.

5 A COMPETING HYPOTHESIS

David Barnett (2008) has recently presented an account of consciousness intuitions that is different from my own. He argues for three claims: first, he argues that many well-known arguments in the philosophy of mind rely upon thought experiments that elicit intuitions about consciousness. Second, he argues that intuitions about consciousness are constrained by what he calls the "Simplicity Intuition." Third, he argues that the Simplicity Intuition is false. Barnett's three claims entail the same conclusion that my account entails: we should reassess many arguments that appeal to our intuitions about consciousness.

I do not dispute the first or the third of Barnett's claims. In Section 5.1, I summarize his argument for the second claim, that our intuitions about consciousness are constrained by the Simplicity Intuition. In Section 5.2, I argue in favor of my own account of intuitions about consciousness. I argue that it provides a more systematic explanation of our intuitions than Barnett's Simplicity Intuition.

5.1 David Barnett's Simplicity Intuition

According to Barnett's Simplicity Intuition: "Our naïve conception of a conscious being demands that conscious beings be simple (i.e., that they not be composed of other things)" (309). Barnett begins his argument for the Simplicity Intuition by eliciting what he calls the "core intuition:" a pair of people cannot be conscious (312). Barnett supposes that the core intuition must have a source and he suggests five candidates: Number, Nature, Relation, Structure, and Simplicity. He then argues that, of the five candidate sources, only Simplicity can provide the "full explanation" of the core intuition. We do not attribute consciousness to a pair of people because a pair is composed of other things rather than "simple."

According to Number, we never attribute consciousness to a being with two parts or fewer. Number cannot be the "whole explanation" because increasing the number of parts of an object does not increase the likelihood that we will attribute consciousness to it (315). According to Nature, we never attribute consciousness to beings that are composed of other conscious beings. Nature cannot be the whole explanation because we are no more likely to attribute consciousness to a pair of carrots than we are to attribute it to a pair of people (315). According to Relation, we only attribute consciousness to beings whose parts relate to each other in a particular way. Barnett argues that Relation could not be the whole explanation because there is no reason why a pair of people could not relate to each other in the required way (315-316). According to Structure, we never attribute consciousness to mere collections of parts; rather, we attribute consciousness to structures composed of interrelated parts. Structure cannot be the whole explanation because we can “mentally impose structure” on a pair of people and we still do not see them as conscious (316).

Finally, Barnett considers whether some combination of Number, Nature, Relation, and Structure could explain the core intuition and concludes that none can. The rejection of Number, Nature, Relation, Structure, and any combination thereof leads Barnett to conclude that the Simplicity Intuition is the best explanation of the intuition that a pair of people could not be conscious. According to Simplicity, we never attribute consciousness to objects that are composed of other things, i.e. not "simple."

Barnett provides further evidence for the Simplicity Intuition by suggesting that it also explains many of the well-known intuitions that philosophers have elicited. The Simplicity Intuition is indeed consistent with our intuitions about Block’s Nation of China and Homunculi-Headed Robot, Davidson’s Swampman, Leibniz’s Mill, and Searle’s Chinese Room. We under-

stand all four of these systems as composites and we do not have the intuition that they are conscious.

5.2 The Simplicity Intuition Versus The AGENCY Model

According to Barnett, understanding a system as a composite prevents us from having the intuition that it is conscious. On my view, failing to understand a system as an agent prevents us from having the intuition that it is conscious. If attributing agency to a system required that we conceive of the system as simple, Barnett's proposal would be compatible with my own. In some cases, however, we understand a "composite" system as an agent. In other cases, we fail to understand a "simple" system as an agent.[7] Barnett's account and my own yield different predictions of the intuitions elicited by these two kinds of cases. If we understand a system as a composite agent, Barnett's account predicts that we will not understand it as conscious, while my account predicts that we will. If we understand a system as a simple non-agent, my account predicts that we will not understand it as conscious, while Barnett's account makes no prediction.

Two examples of intuitions about consciousness show that my account is more consistent with our intuitions than Barnett's. First, consider a "composite" system that we are likely to perceive as an agent since its behavior does not seem to be caused by its parts. Paul Bloom and Csaba Veres (1999) perform a Heider Simmel–style experiment, but they present subjects with *groups* of shapes moving on a screen rather than single shapes.[8] The groups of shapes move in an apparently unified way, but each shape in the group bears no apparent causal relation to the behavior of the group as a whole. In one control group, the groups of shapes were still. In another control group, the groups of shapes moved at constant speeds along linear trajectories (B4).

---

[7] For the remainder of this paper, I use quotation marks in the terms "'simple' system" and "'composite' system" as shorthand for "system *that we understand to be* simple" and "system *that we understand to be* a composite," respectively.

[8] Also see Bryce Huebner's (2010a) discussion of the Bloom and Veres experiment (3).

According to the Simplicity Intuition, subjects will not attribute mental states to any groups of shapes. On my account, subjects will likely attribute mental states to the groups moving along unexpected trajectories in order to explain their behavior, since the behavior does not appear to be explained in any other way. In fact, Bloom and Veres found that subjects did ascribe mental states including intentions and desires to the groups of shapes moving in agential ways (B4). Furthermore, when asked how many "characters" were in the animation, subjects' answers indicated that they counted a group of shapes moving together as one character (B6). In other words, it seems that subjects genuinely attributed mental states to the group, rather than to each individual shape in the group.

Though each group of shapes was presented clearly as a collection of parts that were not spatially continuous with one another, it is possible that subjects viewed each group as a single object. If so, the results from Bloom and Veres would not provide clear counterevidence to the Simplicity Intuition. It may be instructive to run a similar study in which it is more apparent that the interaction of the shapes in each group does not determine the behavior of the group as a whole. Bloom and Veres note that "[i]t is an open question whether the same results would emerge if each group was a swarm of distinct objects moving relative to one another" (B6).

Second, consider a "simple" system that we are not likely to perceive as an agent since its behavior seems to have a non-agential explanation. There are many examples of such systems, e.g. a candlestick, a statue, a rock, or a piece of furniture. Of course we do not typically attribute consciousness to such entities, but Barnett notes that we can easily imagine such systems as conscious when, for example, they are animated in Disney movies (314).

Why do we believe that a candlestick is not conscious when it sits on the table but we believe that it is conscious when it dances with the feather duster in *Beauty and the Beast*? The

Simplicity Intuition cannot explain the difference. We think of a candlestick as simple when we see it on the table and we think of it as simple when it is animated, so the Simplicity Intuition does not entail the failure to understand it as conscious in either case. My account, however, can explain the difference.[9] We are not inclined to attribute agency to the candlestick sitting still on the table because it is not behaving in a way that demands an agential explanation. An animated cartoon candlestick, however, might move about the screen. If its movements do not have an apparent mechanistic explanation, we will be inclined to posit an agential explanation for its movements and we will likely believe that it is conscious.

The AGENCY Model predicts our intuitions about both "composite" agents and "simple" non-agents, while the Simplicity Intuition yields a mistaken intuition about "composite" agents and yields no prediction about "simple" agents. Its predictive power and accuracy is just one advantage of the AGENCY Model relative to the Simplicity Intuition. In addition, the AGENCY Model seems explanatory in a way that the Simplicity Intuition is not.

5.3 Argument from Explanation

Barnett's Simplicity Intuition offers a method for predicting whether we will have the intuition that a system can be conscious. If we do not think of the system as simple, we will not think of it as conscious. If we do think of the system as simple, we might think of it as conscious. The AGENCY Model offers a method for predicting our intuitions as well (a more effective one, as I argued in the previous section), but it also does something else. It suggests a plausible explanation for why we have the intuitions that we do.

An explanation of our intuitions interesting for its own sake, but it also may provide abductive evidence for the truth of the AGENCY Model. The claims that (i) mechanistic under-

[9] This point was developed in discussion with Eddy Nahmias.

standing discourages agency attributions and (ii) that we don't attribute consciousness to systems that we do not understand as agents are likely correct, I argue, not just because they predict a wide range of our intuitions about consciousness but also because there seems to be a plausible explanation for *why* claims (i) and (ii) might be correct.

Jack et al. suggest two possible explanations for the reciprocal inhibition between mechanistic and mentalistic reasoning. First, they suggest that it may have evolved because it facilitates accurate predictions of the behavior of both agents and non-agents. They write: "It would be no less foolish to suppose that a person will continue in motion in a straight line unless acted upon by an external force, than it would be to suppose that a pool ball will alter its course because it wants to go into the pocket" (396). Second, they suggest that the reciprocal inhibition between mechanistic and mentalistic reasoning exists because mentalistic reasoning, and only mentalistic reasoning, is linked to moral concern (396).

Jack et al. favor the second of the two explanations. They note that moral cognition is correlated with activation in a similar set of regions to the DMN, which they take as evidence of the link between mentalistic reasoning and moral concern (396). Jack et al. hypothesize that the inhibition between the two cognitive systems is "driven by the need to differentiate members of our moral circle from objects suitable for manipulation" (396). That is, the reciprocal inhibition between mechanistic and mentalistic reasoning allows us to manipulate mechanisms without feeling concern, while pushing an agent, on the other hand, causes moral guilt.

Nahmias, Coates, and Kvaran also suggest that there is tension between the mechanistic stance and the mentalistic stance because only the mentalistic stance is associated with moral reasoning. They, however, focus on reasoning about moral agency rather than simply moral patiency. That is, Nahmias and colleagues suggest that subjects adopt an "objective attitude" to-

wards systems viewed from the mechanistic stance. Not only do they see such systems as objects that can be manipulated, "managed," or "handled;" they fail to see mechanistic systems as participants in the moral community (223). Non-mechanistic systems, on the other hand, are not viewed as objects that can be manipulated without moral concern and they are viewed as participants in the moral community. We ascribe moral responsibility only to non-mechanistic systems.

6 CONCLUSIONS

Any system, conscious or non-conscious, can be described in a way that prevents attributions of mental states to the system. Simply by explaining how a system works in terms of the interactions of its parts, the author of a thought experiment discourages us from attributing agency to the system. If AFGN's AGENCY Model is correct, the failure to attribute agency to the system prevents us from understanding it as conscious.

The intuitions elicited by descriptions of the functional decomposition of systems do not reflect the intuitions we might have in other circumstances nor do they seem to reflect the actual likelihood that a system is conscious. When a thought experiment includes such descriptions, then, we should understand the intuitions it elicits as determined by the psychology of mental state attribution and not necessarily as a reflection of any feature of the system.

I conclude that we should reconsider the morals drawn from many famous thought experiments including Block's Nation of China and Homunculi-Headed Robot, Davidson's Swampman, Leibniz's Mill, and Searle's Chinese Room. We should not be troubled if our intuitions about thought experiments like these conflict with otherwise promising theories. Our intuitions may help us to learn about the psychology of consciousness attributions, but we should not expect them to be consistent with any particular theory of mind.

The suggestion that we should not expect a theory of mind to cohere with all of our intuitions about consciousness implies that we should be careful in using appeals to intuitions about consciousness in arguments in philosophy of mind. How, then, should we proceed? As Eric Schwitzgebel (2013) writes, "[t]here is no conscious-ometer" (36). Aside from our intuitions, there is no way of determining which systems are conscious and which are not, which seems essential for developing an account of the necessary and sufficient conditions for consciousness.

Schwitzgebel writes that if two philosophers disagree about whether a frog is conscious, "no output from an fMRI machine or set of single-cell recordings is likely to resolve their disagreement" (37). If there is no empirical method for settling questions about the metaphysics of the mind, and if we should avoid appeals to intuition, it's not clear what methods philosophers of mind can rely upon.

Schwitzgebel's answer is to embrace "crazyism" about the mind. Crazyism is the view that something must be true about the metaphysics of mind that seems crazy or absurd. Schwitzgebel argues that commonsense metaphysics is incoherent and yields contradictions. Commonsense requires that either materialism, dualism, idealism, or a compromise/rejection position be true, yet commonsense conflicts with each of these four positions (29). Schwitzgebel argues for "universal dubiety" in the metaphysics of mind, the claim that *no* metaphysical position compels belief. He concludes that a philosopher should both advocate her preferred view and remain intellectually modest: she should "acknowledge dubiety" (35).

It may well be that our intuitions about consciousness are inconsistent and that, as a result, any theory of mind will seem crazy. Rather than acknowledge the universal dubiety of all metaphysical positions, however, my argument here might provide an example of another way a philosopher of mind might proceed. If we have particular reasons to doubt *certain kinds* of intuitions about consciousness, but not necessarily *all* intuitions about consciousness, perhaps not all positions in the metaphysics of mind are equally dubious. In other words, if we investigate the psychology of mental state attribution further, we may find that we have more reason to accept some seemingly "crazy" positions than others.

REFERENCES

Arico, A., B. Fiala, R. Goldberg, and S. Nichols (2011). "The Folk Psychology of Mental States." *Mind & Language* 26 (3): 327-352.

Barnett, D. (2008). "The Simplicity Intuition and its Hidden Influence in Philosophy of Mind." *NOÛS* 42 (2): 308–335.

Berker, S. (2009). "The Normative Insignificance of Neuroscience." *Philosophy & Public Affairs* 37 (4): 293-329.

Block, N. (1978) "Troubles with Functionalism," in *Readings in Philosophy of Psychology* 1980, edited by N. Block., Cambridge: Harvard University Press, 268–305.

Bloom, P. and C. Veres (1999). "The Perceived Intentionality of Groups." *Cognition* 71: B1-B9.

Chalmers, D (1996). *The Conscious Mind: In Search of a Fundamental Theory,* Oxford: Oxford University Press.

Davidson, D (1987). "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60 (3): 441-58.

Dennett, D (1973). "Mechanism and Responsibility." In *Free Will*, ed. Garry Watson. New York: Oxford University Press, 1982, 150-73.

Gray H., K. Gray, and D. Wegner (2007). "Dimensions of Mind Perception." *Science* 315 (5812): 619. Supporting Online Material, 2-10.

Greene, J. D. (2000). "The Secret Joke of Kant's Soul." In *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development* 2007, edited by W. Sinnott-Armstrong. Cambridge, MA: MIT Press.

Heider, F. and M. Simmel (1944). "An Experimental Study of Apparent Behavior." *American Journal of Psychology* 57: 243-259.

Huebner, B. (2010a). "Commonsense Concepts of Phenomenal Consciousness: Does Anyone Care About Functional Zombies?" *Phenomenology and the Cognitive Sciences* 9 (1): 133-155.

Huebner, B., M. Bruno, and H. Sarkissian (2010b). "What Does the Nation of China Think about Phenomenal States?" *Review of Philosophy and Psychology* 1 (2): 225-243.

Jack, A. I., A. J.Dawson, K. L. Begany, R. L. Leckie, K. P. Barry, A. H. Ciccia, and A. Z. Snyder (2013). "fMRI reveals reciprocal inhibition between social and physical cognitive domains." *NeuroImage* 66: 385-401.

Leibniz, G. W. (1714). *Monadology and Other Philosophical Essays*. Trans. & ed. by P. Schrecker and A. M. Schrecker. New York: Bobbs-Merrill Co., 1965.

Nado, J. (2012). "Why Intuition?" *Philosophical and Phenomenological Research.* doi: 10.1111/j.1933-1592.2012.00644.x

Nahmias, E., D. J. Coates, and T. Kvaran (2007). "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions" *Midwest Studies in Philosophy* 31 (1): 214-242.

Poincaré, H. (1914). *Science and Method*. Trans. F. Maitland. T. Nelson, London. 93-116.

Schnall, S., J. Haidt, G. L. Clore, and A. H. Jordan (2008). "Disgust as Embodied Moral Judgment" *Pers Soc Psychol Bull* 34 (8): 1096-1109.

Schwitzgebel, E. (2013). "The Crazyist Metaphysics of Mind," draft updated June 5, 2013. Accessed online: http://www.faculty.ucr.edu/~eschwitz/SchwitzPapers/CrazyMind-130605.pdf

Weinberg, J. M., C. Gonnerman, C. Buckner, and J. Alexander (2010). ''Are Philosophers Expert Intuiters?'' *Philosophical Psychology* 23 (3): 331–55.

Williamson, T. (2011). "Philosophical Expertise and the Burden of Proof." *Metaphilosophy* 42 (3): 215-229.