Philosophy in Practice

VOLUME 7 – SPRING 2013

CALIFORNIA STATE UNIVERSITY, LOS ANGELES
DEPARTMENT OF PHILOSOPHY

Philosophy in Practice

Volume 7 – Spring 2013

Contents

Articles:

Acknowledgments

Philosophy in Practice is produced by students at California State University, Los Angeles. The editorial staff wishes to thank all the authors for submitting and refereeing articles, the faculty of the philosophy department for agreeing to supervise individual contributions, and the University for its generous financial support.

Editors:

Chuck Dishmon
Melvin J. Freitas
Douglas C. Wadle

Faculty Advisor:

Dr. Michael K. Shim

California State University, Los Angeles Philosophy Faculty

Talia Bettcher (2000–), Chair, Ph.D. University of California, Los Angeles. History of Modern Philosophy, Philosophy of Self, Gender and Sexuality

Mohammed Abed (2008–), Ph.D. University of Wisconsin, Madison. Ethics, Social and Political Philosophy, Philosophies of Violence, Genocide and Terrorism

Mark Balaguer (1992–), Ph.D. City University of New York. Philosophy of Mathematics, Metaphysics, Meta-ethics, Philosophy of Language, Logic

Anna Carastathis (2009–), Ph.D. McGill University. Social and Political Philosophy, Feminist Philosophy, Philosophy of Race, Critical Race Theory, Postcolonial Theory, Phenomenology

Richard Dean (2009–), Ph.D. University of North Carolina, Chapel Hill. Ethics, Kant's Moral Philosophy, Applied Ethics

Foad Dizadji-Bahmani (2013–), Ph.D. London School of Economics, United Kingdom. Philosophy of Science, Philosophy of Physics, Philosophy of Probability

Ronald Houts (1983–), Ph.D. University of California, Los Angeles. Metaphysics, Epistemology, Logic

Henry R. Mendell (1983–), Ph.D. Stanford University. Ancient Philosophy, History of Ancient Mathematics and Science, Philosophy of Science, Metaphysics

David Pitt (2003–), Ph.D. City University of New York. Philosophy of Mind, Philosophy of Language, Metaphysics

Joseph Prabhu (1978–), Ph.D. Boston University. Philosophy of Religion, 19th and 20th Century German Philosophy, Moral and Social Philosophy, Indian and Comparative Philosophy

Sheila Price (1964–). Recent Philosophy, Comparative Religions, Medical Ethics, Environmental Ethics.

Michael K. Shim (2007–), Ph.D. State University of New York, Stony Brook. 20th Century Continental Philosophy, Phenomenology, Husserl, Modern Philosophy, Philosophy of Mind, Philosophy of Language

Kayley Vernallis (1993–), Ph.D. University of California, Berkeley. Moral Psychology, 19th and 20th Century Continental Philosophy, Feminist Philosophy, Ethics, Aesthetics, Gender and Sexuality

Emeritus Professors

Sidney P. Albert (1956–1979). Aesthetics, Ancient Philosophy.

Thomas Annese (1961–1992). Epistemology, Modern Philosophy.

Sharon Bishop (1967–2004). Ethics, Political Philosophy, Philosophical Psychology, Feminist Ethics.

Donald Burrill (1962–1992). Ethics, Philosophy of Law, American Philosophy.

Ann Garry (1969–2011). Feminist Philosophy, Philosophical Methodology, Epistemology, Applied Ethics, Wittgenstein, Philosophy of Law

Ricardo J. Gómez (1983–2011). Philosophy of Science and Technology, Philosophy of Mathematics, Kant, Latin American Philosophy

George Vick (1967–1997). Metaphysics, Phenomenology, Existentialism, Philosophy of Religion, Medieval Philosophy.

Professor Spotlight: Foad Dizadji-Bahmani

When asked how to pronounce his last name, Professor Foad Dizadji-Bahmani disarmingly told us that even he hasn't the slightest idea how to say it, and simply advises his students to call him "Foad" (pronounced FOE-add). And if you insist on formality, he will reluctantly accept "Professor Foad" as an alternative. But whatever you choose to call him, the newest member of our department's faculty is most certainly a philosopher. This was immediately apparent when we began our interview by asking him, "Where are you from?", since he then proceeded to question the question as only a philosopher would. Does this interrogatory refer to the place of one's birth? Does it refer to one's nationality? Does it refer to one's current place of residence? Is it meant to place one's accent (as Foad's is distinctively English)? Or is it meant to identify one's race? After a while of this, we could only wonder how difficult this interview was going to be. But, in the end, this is actually a difficult question in Foad's case. He was born in Iran but left at the age of two with his mother to seek political asylum in Berlin, where he lived for six years before moving to London. To explain the move to the UK, Foad told us that his mother had studied at UCLA, so she was proficient in English, and had thus wanted to move to an Anglophone country. Foad himself is fluent in German, Farsi, and English.

From this point Foad's biography is a bit more conventional, though his person may not be. Foad first received a B.Sc.

in physics and philosophy (joint honors) and then an M.A. in the philosophy and history of science (with distinction) at the University of Bristol. He then went on to receive his Ph.D. in Philosophy at the London School of Economics while continuing his work in the philosophy of science and the philosophy of physics. His dissertation supervisors included Dr. Roman Frigg (LSE), Dr. Miklós Rédei (LSE), Prof. Harvey Brown (Oxford), and Prof. Craig Callender (UCSD). After receiving his Ph.D., Foad was appointed as a Fellow at the London School of Economics before coming here to Cal State L.A. One quick look at his curriculum vitae will ensure you that the department has made a fine choice in hiring someone who is already an exceptional scholar and teacher. As one anonymous Cal State student recently put it, "If his British accent doesn't hook you, then his passion for philosophy will!"

Then again, if the "inter-theoretic reduction of thermodynamics to statistical mechanics" is not your cup of tea, it might interest you to know that Foad is now the proud owner of a mint green 1966 Ford LTD. Thus, you might soon see him on a classic PCH road trip, learning to surf, checking out the Sequoias, or furthering his research in Bayesian probability at the MGM Grand, Las Vegas. Foad is also a big time foodie, admitting that he nearly came to tears when biting into a freshly made burrito from a local street vendor. It was apparently that good. He also claims to have recently had an out-of-body experience testing out the fare at Umami Burger. Foad is also a big jazz fan, being particularly fond of John Coltrane's A Love Supreme (the live version), and he's now missing his regular jazz fixes at The Hackney Cut in London where he's a good friend of the owners. More practically, Foad offers aspiring CSULA graduate students the following sage piece of advice: when you're working on your dissertation, you must keep reading good literature, as this will help keep you focused as well as open-minded, in terms of problem-solving approaches. In this regard Foad thinks that David Foster Wallace's book Infinite Jest is the best book ever written (including all 388 of its endnotes).

In short, you need only talk to Foad a short time to know that he's started an adventure here at Cal State L.A. Though he's never before taught a class with 150 students, he's found the students here to be very earnest in their desire to learn. Fortunately for us, he didn't go to law school, despite having been accepted a couple places, and chose instead his enduring love for philosophy. Foad told us that if he had it all to do over again, he would definitely study philosophy, but his second choice would now be architecture. He has an abiding interest in art, which his local friends have indulged by purchasing him a pass to LACMA. And, should you meet Foad, you'll quickly notice that he has quite the sense of humor. Just ask him about his accidental trip through Skid Row on the way to Karaoke the other night. Or let him apprise of you of the fact that, when you move from England to the U.S., your credit history is completely wiped out (despite using the very same financial institution in both countries). Keep in mind, however, that some of his stories may be embellished, as the editors have not been able to independently verify his claim to have recently wrestled a mountain lion. Similarly, many of his humorous claims about Professor Dean (his office mate) may be apocryphal at best. But all laughs aside, we are certainly fortunate to have Professor Foad here at the Cal State L.A. Philosophy Department.

— C.D., M.J.F. & D.C.W.

Representational Theories of Phenomenal Consciousness in Animals

Nathaniel Greely

A few decades ago it was standard practice to perform surgery on pets without anesthetic (Allen 2010). Part of the rationale for this was the surprisingly common notion that animals don't have the type of consciousness necessary to feel pain and suffer. If animals don't have what we call "phenomenal consciousness", then they can't feel pain the way we do, and what sounds like anguished barking is nothing more than stimulus-response. Most pet owners today would shudder at such a practice. It has become commonplace to assume that at least some nonhuman animals possess phenomenal consciousness; that is, the sort of mental states with a "felt" quality, like pain, smell, or color experience.

Many contemporary philosophers, like contemporary pet owners, are much more amenable to the idea of phenomenal consciousness in animals than their predecessors. Very few, however, have offered any theories about precisely where on the phylogenetic scale phenomenal consciousness arises. Representational theories of consciousness like those of Fred Dretske (1995) and Michael Tye (1995) offer an account of mentality that could arise in very simple systems, and as such are particularly suited to address the question of animal consciousness. Dretske prefers to leave the question open, claiming that consciousness arises "At the same time a poor man becomes rich as you keep giving him pennies" (Dretske 1995, p. 168). Tye, on the other hand, makes the very specific claim that phenomenal consciousness arises with the ability to form simple beliefs, and that this occurs in animals that are able to learn and modify their own behavior (1997). He claims that bees and fish form beliefs, and that therefore they are phenomenally conscious.

I will argue there are two problems with Tye's thesis. Firstly, I will attempt to show that Tye's claim that phenomenal consciousness is contingent upon the ability to form beliefs is precarious because it relies on his solution to the pathological case of blindsight, and I will argue that his solution is problematic. Secondly, I will review phenomenological arguments by Hubert Dreyfus (2005), which suggest that learning may not be contingent upon beliefs, casting doubt on whether learning could be an empirical indicator of belief in animals, as Tye claims. Third, I will argue that instead, variations on the standard empirical approach already commonplace in scientific research on pain, while they could not definitively prove the existence of phenomenal consciousness in animals, are the next best thing. They show correlations that provide us with the most fruitful places for future philosophers and scientists to look for consciousness in animals, if it is ever to be found (Allen 2010).

Because Tye's claims are contingent upon a series of arguments that ultimately trace back to the very foundation of representational theories of mind, I will proceed in five sections. In Section 1, I will briefly address why animal consciousness is of interest to philosophers of mind and how representational theories are particularly pertinent. In Section 2, I will sketch out the nature of representational theories of mind and show that they necessitate a commitment to phenomenal content externalism and a denial of higher order theories of phenomenal consciousness. In Section 3, I will explain Tye's own PANIC theory of phenomenal consciousness, his solution to the problem of blindsight, and his position on phenomenal consciousness in animals. In Section 4, I will spell out my two objections to Tye's theory and, in Section 5 I will offer a suggestion for a more fruitful line of empirical inquiry.

Section 1: Why Study Animal Consciousness?

Consciousness itself remains an elusive topic, and many philosophers are resigned to an unbridgeable "explanatory gap" between the facts studied by the objective sciences and the inherently subjective aspects of consciousness. If we cannot be objectively certain that our closest human friends are not unconscious “zombies”, regardless of how much we might know about their behavior and their brains, then why speculate on the status of animals? Even the most optimistic among us must admit that it is entirely possible that this explanatory gap may never be closed, but we must also admit that other phenomena which appeared equally mysterious in the past have since been explained. We may be cautiously optimistic at least of the possibility that the problem of explaining minds may one day go the way of the problem of explaining life, and our present arguments over consciousness may seem as quaint as the arguments over vitalism a century ago. If we are to have any hope of closing the explanatory gap, we must begin to construct the theoretical framework. As with the problem of life, the answers may well be found in simpler biological entities. Therefore, it’s not unreasonable to investigate, to the best of our ability, “what it’s like” to be a bat, or a honeybee, or a fish.

Section 2: Representational Theories of Consciousness

Michael Tye claims that in fact there is something that it is like to be a fish, as well as a honeybee. His claim is based on considerations that relate to the whole of his representational theory of mind, and so in order to understand his position, it is important to understand representationalism and its commitments. Representational theories of mind, like those of Tye and Dretske, claim that mental states are representations of states of affairs in the world, in much the same way that states of a speedometer are representations of the speed of a vehicle (Dretske 1995, p. 2). Many natural phenomena covary. The size of a piece of metal varies reliably with temperature, expanding when hot and shrinking when cool. But what makes a system representational is the fact that it is designed to indicate certain states of affairs in the world, as is a column of mercury in a thermometer. Once a thermometer has been given the purpose of representing temperature by the person who built the device, it becomes possible for it to misrepresent in a way that a piece of metal on its own cannot. A piece of metal always follows the laws of nature, but a thermometer can break down and fail to fulfill its man-made purpose. This possibility of misrepresentation is one of the classic marks of mentality. Dretske claims that we can see a sort of purpose, or teleology, in the process of natural selection (1995, pp. 49–50). Certain spontaneously occurring mechanisms in an organism, which correlate with states of the world, are perpetuated and multiplied due to the survival value of this correlation. This survival value creates the "purpose" which makes a certain mental state representational rather than simply covariant, and also makes it possible for them to misrepresent. Dretske's account shows how a simple mental mechanism which correlates with a state of affairs in the world, say a detector of a certain type of food, could arise by natural selection, being of such benefit to a primitive organism that it would be more likely to thrive, survive, and reproduce.

Because of this straightforward relationship between mental states and the world, representational theories of consciousness are committed to content externalism of the type initially championed by Hilary Putnam. Putnam's arguments were strictly concerned with linguistic meaning, but were expanded by later philosophers to include intentional states, and expanded further by Dretske to include phenomenal mental states as well. Says Dretske, "The Representational Thesis is an externalist theory of the mind. It identifies mental facts with representational facts, and though representations are in the head, the facts that make them representations—and, therefore, the facts that make them mental—are outside the head" (Dretske 1995, p. 124). Putnam uses the "Twin Earth" argument to show that identical mental states in physically indistinguishable subjects can have different content depending upon the subject's environment. If this is the case, and many philosophers are convinced it is, then mental content isn't constituted wholly by intrinsic features (Putnam 1973). Putnam describes a world nearly identical to earth with a "molecule for molecule" duplicate of each individual on earth.

The only difference between the two worlds is that the substance we both call "water", though phenomenally indistinguishable on both planets, is composed of different molecules, H2O on Earth and XYZ on Twin Earth. This means that an individual on earth viewing water and having a "water thought" would have different mental content from an individual on Twin Earth who, though in an identical brain state, is thinking of XYZ rather than H2O. This view of mental content is essential for the representational theorist, for whom all mental states are "intentional", meaning they are about states of affairs in the world.

Dretske acknowledges that not all philosophers are content externalists, but nonetheless takes Putnam's argument as at least possibly true. He then argues that, "if an externalist theory of thought can be true, an externalist theory of experience can also be true" (Dretske 1995, p. 127). Dretske admits that he has no knock-down argument for externalism about phenomenal content, or "experience' as he calls it. The argument he offers in *Naturalizing The Mind* is too lengthy to recreate here, and it is not meant to convince any hardened opponents of content externalism. For the purposes of this paper it is sufficient simply to point out that representational theories of mind are committed to phenomenal content externalism. If mental states are representations, then colors and other phenomenal content are not in the mind, they are in the objects that the mental states represent.

Because they are committed to phenomenal content externalism, representational theories of consciousness are also necessarily opposed to higher-order theories of phenomenal consciousness. Higher order theories of consciousness are used to account for common occurrences like absent-minded driving. We often find we are able to do a great many complicated tasks while our attention is on other things, like driving a car while daydreaming. While not necessarily unconscious, the absent minded driver seems to be functioning at least on a lower level of consciousness than when she is snapped out of her daydream into a more attentive state. After the driver is brought to awareness, she "sees" many objects that escaped her attention before, like other cars, though obviously she was still able to somehow see enough to navigate around them while in her lower, "absent-minded" state of consciousness. This ability to snap in and out of attentiveness to our phenomenal states could be explained by some higher order awareness or "spotlight" that can be shifted to focus on our lower order phenomenal states. This type of higher order theory is known as "higher order experience" or "higher order perception" theory. As we will see later, an appeal to higher order perception might help explain the phenomenon of "blindsight". Unfortunately, higher order perception is incompatible with representationalism, as Dretske points out:

> All one can become aware of by scanning (monitoring - choose your favorite word) internal affairs are activities of the nervous system. That, after all, is all that is in there. All that is in the head are the representational vehicles, not the contents, the facts that make these vehicles into thoughts and experiences (Dretske 1995, p. 108).

For the representational theorist, mental states transparently represent states of affairs in the world, and their contents aren't in the head, so there is nothing for this inner spotlight to scan but grey matter. There are other theories of higher order consciousness, including "higher order thought", which claims that cognitive states like beliefs and judgments are necessary for phenomenal consciousness (Carruthers 2011). While representational theorists don't deny the existence of higher order cognitive states, they do deny that they are always necessary for phenomenal consciousness. Rather, as we will see shortly, Tye claims that phenomenally conscious states are "poised" or available to cognitive states for the formation of beliefs, but needn't always result in the formation of beliefs.

Section 3: Tye's Theory of Phenomenal Consciousness

Now we are prepared to take a look at Tye's theory of phenomenal consciousness and what implications it has for phenomenal consciousness in animals. Tye uses the acronym PANIC to represent the essential features of phenomenal content, the characteristic experiences and feelings that constitute phenomenal consciousness. He claims that, “experience and feeling arise at the level of the outputs from the sensory modules and the inputs to a cognitive system. It is here that phenomenal content is found” (Tye 1995, p. 137). “Sensory modules” are systems like the eye and its associated neural machinery, and the “cognitive system” is that part of the mind responsible for propositional attitudes like belief and desire. Tye claims that phenomenal content occurs at the intersection of these two systems.

Tye’s PANIC theory claims that phenomenal content is Poised, Abstract, Non-conceptual, Intentional Content. By poised Tye means that phenomenal states are available to make a difference to propositional attitudes like beliefs and desires. “To say that the contents stand ready in this way is not to say that they always do have such an impact. The idea is rather that they supply the inputs for certain cognitive processes whose job it is to produce beliefs (or desires) directly from the appropriate nonconceptual representations, if attention is properly focused and the appropriate concepts are possessed” (Tye 1995, p. 138). The claim that these contents need not always make an impact on cognitive states distinguishes Tye’s view from higher order thought theories, which according to Dretske (who also rejects them), “maintain that what makes an experience (the sort of mental state we are here concerned with) conscious is that the creature whose experience it is believes, knows, or somehow conceptually represents this experience (or itself as having this experience)” (Tye 1995, p. 106).

By ‘abstract’ Tye means having no concrete contents. “Since different concrete objects can look or feel exactly alike phenomenally, one can be substituted for the other without any phenomenal change” (Tye 1995, p. 138). Instead, “What is crucial to phenomenal character is the representation of general features or properties” (Tye 1995, p. 138). A representation, like a speedometer’s representation of speed, is an abstract indicator of the concrete property, not the concrete property itself, and as such it can even misrepresent. “Whether or not you have a left leg, for example, you can feel a pain in your left leg” (Tye 1995, p. 138).

The N in PANIC stands for ‘nonconceptual’. “The claim that the contents relevant to phenomenal character must be nonconceptual is to be understood as saying that the general features entering into these contents need not be ones for which their subjects posses matching concepts” (Tye 1995, p. 139). For example, we can distinguish far more shades of red than for which we have concepts. Beliefs and other cognitive states require concepts, but phenomenal states do not.

The IC in PANIC stands for intentional contents. For Tye and representational theorists in general, all mental states are intentional, since they represent states of affairs in the world. Says Tye, “All states that are phenomenally conscious—all feelings and experiences—have intentional content” (1995, p. 93). For Tye, inner experiences like pains are intentional as well, they just represent states of affairs within the body instead of outside it.

Tye’s position on animal consciousness hinges on his PANIC theory and his commitment to a particular solution of the problem of blindsight. Blindsight patients have damage to their brains, causing blind spots in their visual field. When prompted, however, they are able to make guesses about objects in this blind spot that are correct to a degree far higher than chance. They are hesitant to make these guesses, and claim to have no idea as to the correct answers; but when they do guess they are largely correct. Clearly these subjects are receiving information about the objects in their blind field, but this information is not available to phenomenal consciousness. One way that philosophers have accounted for this ability is to appeal to higher order theories of consciousness. Higher order perception theory (HOP) can explain blindsight as an inability to focus the “inner spotlight” on lower order phenomenal states in order to bring them to awareness. What these patients have, according to HOP, is an awareness problem, not a vision problem (Carruthers 2011). As we saw in Section 2, representationalists like Tye cannot appeal to HOP, so he must offer another solution.

To make the problem more acute, Tye adopts Ned Block's idea of "super blindsight" (Block 1997). A super-blindsight patient is a hypothetical sort of blindsight patient who has trained herself to make accurate guesses at will about objects in her blind spot, without prompting. Such a patient would be behaviorally indistinguishable from a person with normal vision. What then would be the difference? She would have beliefs about objects in her blind spot, but "… the impact on the belief system here is both anomalous and indirect. First, an act of will is required; then a guess is generated; then the guess comes to be believed" (Tye 1995, p. 143). This is a very different process than that of normal vision. "So there are no nonconceptual contents (pertaining to the blind field) that are appropriately poised in super-blindsight subjects any more than there are in ordinary blindsight subjects. This is why, according to the PANIC theory, experience is lacking in both" (Tye 1995, p. 143). He claims that what blindsight (and super blindsight) patients lack is not higher-order awareness, but the ability to form beliefs about the phenomenal experiences in the blind areas. Keep in mind, he is not saying that these patients are unable to form beliefs because they can't see the objects, but quite literally they can't see these objects because they can't form beliefs about them. This is because the functional role of a phenomenal state is to supply visual information to the cognitive system, and if the state cannot perform its functional role, then it ceases to be phenomenal. "Phenomenal states lie at the interface of the nonconceptual and conceptual domains. It follows that systems that altogether lack the capacity for beliefs and desires cannot undergo phenomenally conscious states," and, he claims, "[t]his approach solves the problem of super blindsight" (Tye 1995, p. 144). Here we have a subtle distinction between higher order thought, which claims that phenomenal states require a belief or other cognitive state be formed about them in order to be conscious and PANIC, which states that phenomenal states must be "poised" or available to the cognitive system. Phenomenal consciousness for Tye is not contingent on any particular beliefs being formed, but it is contingent on the ability to form beliefs. I intend to show in section four that this claim is unintuitive and only motivated by its utility as a solution to the problem of blindsight.

If phenomenal consciousness does in fact depend upon the ability to form beliefs, then perhaps Tye has found an indicator of phenomenal consciousness in all creatures. But how might we tell if an animal has beliefs? This is far from simple, but Tye believes it can be done, because for Tye the ability to learn and respond to situations in a unique, non-reflexive manner requires beliefs.

> They (PANIC) are also states that form the outputs of sensory modules and stand ready and available to make a direct difference to beliefs and desires. It follows that creatures that are incapable of reasoning, of changing their behavior in light of assessments they make, based upon information provided to them, by sensory stimulation of one sort or another, are not phenomenally conscious. Tropistic organisms, on this view, feel and experience nothing. They are full-fledged unconscious automata or zombies, rather as blindsight subjects are restricted unconscious automata or partial zombies with respect to a range of visual stimuli (Tye 1997, p. 301).

So if we find animals that can do more than respond to stimuli in a fixed way and instead modify their behavior and even predict novel situations, then we have learning and therefore belief, and therefore phenomenal consciousness. Tye finds this ability in fish and bees. “Fish do not typically react in a purely reflexive manner. The behavior they produce often depends upon their evaluations or judgments of the deliverances of their senses and their immediate goals” (Tye 1997, p. 304). He cites experiments in which fish learned to avoid eating other fish that had been artificially colored and injected with a bad-tasting chemical, while continuing to eat the naturally colored fish of the same species. In other studies, bass learned to ignore minnows behind a glass wall. “Cumulatively, the evidence seems best explained by supposing that fish often make cognitive classifications or assessments, directly in response to the information conveyed to them by their senses, and that these, together with their goals, often determine their behavior" (Tye 1997, p. 305).

This process is made possible by "simple beliefs". "In this sense, given the facts adumbrated earlier about fish behavior, it seems to me very plausible to suppose that fish form simple beliefs on the basis of immediate sensory representations of their environments" (Tye 1997, p. 306). (What exactly he means by "simple beliefs" is not clear, but I will attempt to sort this out later.) "So," concludes Tye, "fish are the subjects of states with PANIC. They are phenomenally conscious" (1997, p. 306). Tye also cites experiments in which honeybees are able to learn, and so determines that honeybees are phenomenally conscious as well.

Tye's thesis is a bold one, and as he notes, "There may be some reluctance to say that fish have beliefs" (1997, p. 306). The claim does seem implausible prima facie and one's first impulse may be to question what Tye means by beliefs. He introduces the term "perceptual beliefs" in this paper, and defines it in relation to "perceptual concepts". He elaborates his idea of perceptual concepts in *Ten Problems*, dividing them into "indexical concepts" which simply point out a quality, are not predicative, and therefore not available for retrieval by memory, and "predicative concepts" like "red", which can be accessed by memory (Tye 1995, p. 167).

So as regards fish,

> Perceptual beliefs are (roughly) representational states that bring to bear such concepts upon stimuli and that interact in rational ways, however simple, with one another and other representational states the creature generates in response to its needs, thereby determining behavior. Perceptual beliefs are like inner maps by which the creature steers (Tye 1997, pp. 305–306).

Since these perceptual concepts are available to the memory of fish, they must be of the predicative sort. So should we take it that perceptual beliefs, which operate on these concepts, are propositional? Tye, in his article on animal consciousness, avoids stating explicitly that “perceptual beliefs” are propositional attitudes or sentences, but it is clear from *Ten Problems* that this is his position as regards belief simpliciter. Speaking of the symbolic nature of intentional states, he writes, “In the case of beliefs, the symbol structures are sentences, but not, I claim, in the case of sensations” (Tye 1995, p. 100). So whatever the difference between “perceptual beliefs”, “simple beliefs”, and plain old beliefs, they all at the very least seem to be cognitive states that employ predicative concepts, either in sentences or “maps”.

SECTION 4: OBJECTIONS TO TYE’S THEORY

I have two objections to Tye’s thesis. One is more controversial, and I will deal with it last. It is that it is questionable that beliefs are required for learning, hence we cannot infer that a creature has beliefs just because it can learn. This objection, however, is only relevant if phenomenal consciousness is contingent on beliefs. But, and this is my primary objection, I do not believe Tye has shown it to be the case. I will address this more important objection first.

Tye admits that it may be difficult to accept the idea that fish have beliefs, and he attempts to mitigate this by his description of “perceptual beliefs” and “perceptual concepts”, which I have described in the last section. As I pointed out, if we take his more complete characterization of perceptual concepts from *Ten Problems*, we see that for these concepts to be available to memory and belief they must be of the predicative type, not the indexical type; and so Tye must attribute some sort of predicative nature to fish beliefs.

We thus have the counter-intuitive claim that fish possess propositional attitudes of some sort, employing predicative concepts. Such a claim requires strong support; but the support Tye offers hinges on his solution to the pathological case of blindsight. Gabriele Jackson (2011), in “Motor-Intentionality and the Role Of the Pathological in Maurice Merleauy-Ponty’s Work,” questions this all too common practice of drawing conclusions about normal subjects from pathological cases. She quotes Merleau-Ponty:

> How are we to discover by means of it [i.e. the pathological case] what function, found in the normal person, is absent in the patient? There can be no question of simply transferring to the normal person what the deficient one lacks and is trying to recover. Illness… is a complete form of existence and the procedures which it employs to replace normal functions which have been destroyed are equally pathological phenomena. It is impossible to deduce the normal from the pathological… by a mere change of the sign. We must take substitutions as substitutions, as allusions to some fundamental function that they are striving to make good, and the direct image of which they fail to furnish (Merleau-Ponty 1945, pp. 123-124).

What Merleau-Ponty warns against is precisely what Tye does when he assumes that the blindsight patient operates as a normal patient who lacks the ability to form beliefs about certain parts of their visual field. He does not consider the possibility that these abnormal patients have developed abnormal ways of compensating for their visual deficiencies that operate in a much different way than in the normal patient. In fact, he does not even consider that there could be any other explanation beside the "awareness deficiency" which he discards and the "belief deficiency" which he promotes. This might be excusable if blindsight were simply a side note, a problem for which he feels his overall theory might offer an explanation. But in fact his entire thesis about phenomenal consciousness in humans and in animals hinges on his account of blindsight.

In describing his commitment to phenomenal concepts being "poised" (the P in PANIC) Tye states,

> The reason I take the above view of basic perceptual experiences is not simply that it explains certain facts about perceptual illusions [Müller-Lyer diagrams]. In addition, it accommodates our pretheoretical conception of the role

> of experiences as the bedrock for many beliefs and judgments. It is also motivated by a desire to have an account of phenomenal consciousness that fits the facts of blindsight (1997, p. 295).

Müller-Lyer diagrams, the famous diagrams in which arrows appear to be of different sizes even though they are the same size, do not show that phenomenal consciousness is dependent upon belief. All they show is that belief is not determined by perceptual experience. Nor does the fact that experiences are the bedrock for beliefs entail that the capacity for beliefs is necessary for one to have phenomenal consciousness. So, in fact, the only part of Tye's theory that suggests the counter-intuitive conclusion that in order to have phenomenal experience, we must be able to form beliefs about what we see is his account of blindsight, which relies on a simplistic answer to a pathological case.

Even if we were to grant Tye his account of blindsight, and therefore of PANIC, there still exists a problem with his assumption that learning in animals is indicative of their ability to form beliefs. Perhaps for many of us the idea of fish beliefs is patently absurd and needs no further refutation. However, I would like to point out some controversial, yet intriguing evidence that belief is not always necessary for learning even in humans, much less fish. Philosophers in the phenomenological tradition like Dreyfus and Merleau-Ponty have long been battling the idea that learning and skill are conceptual "all the way down". Though they do not deny that some of our mental activity is conceptual, and that conceptual states are available to us at any time, they argue that the bulk of human activity, skill, and learning are non-conceptual. "We directly perceive affordances and respond to them without beliefs and justifications being involved," claims Dreyfus (2005, p. 59). Instead, he claims that we use "motor-intentionality", by which our bodies interact directly with the world. Presumably the brain is often involved in some way, but he does not characterize this activity as "mental" at all. He claims that we do not even use implicit concepts for such activity and that much of the time our way of getting around in the world is no different from that of animals who lack concepts altogether. If human activity is largely belief-free, it makes little sense to attribute beliefs to fish and bees simply because they are able to act on a higher level than stimulus-response.

Much of what makes these phenomenological accounts of learning convincing is their descriptive power, which matches up well with our intuitive notions about how we experience the world. Unfortunately I cannot recreate this in a short paper. Dreyfus does attempt to bridge the gap between the analytic and phenomenological communities, and in doing so offers more concise arguments in addition to phenomenological ones, but from the outset I must admit that they are largely anecdotal. Rather than focus on pathological cases, Dreyfus cites highly skilled individuals like master chess players, who operate so quickly in very complicated situations that it seems absurd to assume that they are consulting some database of concepts, or as Tye suggests, a belief map. "Indeed," claims Dreyfus, "if learners feel that they can act only if they have reasons to guide them, this attitude will stunt their skill acquisition" (2005, p. 52). Though humans are able to learn through concepts and rules, the basic method we share with infants and animals is to "acquire skills by imitation and trial and error" and even if we sometimes use concepts to learn, the nonconceptual again becomes dominant once we attain the level of skill (Dreyfus 2005, p. 52). "Animals, prelinguistic infants, and everyday experts like us all live in this space" (Dreyfus 2005, p. 57).

> We need to consider the possibility that embodied beings like us take as input energy from the physical universe and process it in such a way as to open them to a world organized in terms of their needs, interests, and bodily capacities without their minds needing to impose a meaning on a meaningless Given… nor their brains converting the stimulus input into reflex responses (Dreyfus 2005, p. 49).

So for Dreyfus learning and skill are neither a matter of reflex, nor do they require a map of propositional attitudes.

Dreyfus supports his claim by showing that when conceptual states are imposed onto instances of skillful coping, they actually impede performance. If our learned skills were truly a case of referring to an implicit, internal map of concepts, then making these concepts explicit should not impede the activity. He cites as an example the baseball player Chuck Knoblauch, who developed the nasty habit of thinking about the mechanics of catching a ball. His playing deteriorated except in instances where the moves were so difficult that he did not have time to slip into this contemplative mode. "What he couldn't do was field an easy routine grounder directly to second base, because that gave him time to think before throwing to first… Indeed, he became such a full-time rational animal that he had to be dropped from the team, and he never returned to baseball" (Dreyfus 2007, p. 354). This evidence is, of course, anecdotal, but in longer works like *The Phenomenology of Perception* and *Being and Time*, Merleau-Ponty and Heidegger are able to develop intuitively satisfying accounts of human learning and skill that do not depend on belief. This phenomenological account conflicts directly with Tye's highly unintuitive account that requires fish to form beliefs in order to make adjustments to their environment at a level above pure reflex.

Section 5: A Better Way

If Tye's method of deducing animal consciousness is problematic, then is there a better way to search for phenomenal consciousness in animals? There is no direct method, of course, of peering into the consciousness of other beings; but many of us nonetheless feel confident in attributing consciousness to other humans, primates, and even our pets. We are able to do this with other humans because we each know about our own phenomenal states directly, and we are able to infer from other persons' similar behavior and similar biological structure that they have experiences similar to ours. It would be strange if natural selection were to develop near-identical systems that create states with radically different intrinsic qualities. It is natural to assume consciousness in animals with whom we interact daily and in whom we find similar behaviors and biological structures to our own. As I noted in the introduction, for this reason veterinary surgeons now routinely anesthetize animals for surgery.

> Much of the research that is of direct relevance to the treatment of human pain, including on the efficacy of analgesics and anesthetics, is conducted on rats and other animals. The validity of this research depends on the similar mechanisms involved and to many it seems arbitrary to deny that injured rats, who respond well to opiates for example, feel pain (Allen 2010, p. 29).

I don't mean to promote all such experiments, I'm sure many are no doubt quite cruel. I mean to point out that assuming that animals feel pain in the same way we do, if they exhibit sufficient similarities, has been quite fruitful without any elaborate philosophy of mind to guide it.

Both Tye and Dretske point out, and have to account for, the fact that human experience of color is consistent even across shifting ranges of actual frequency input on the retina. At dusk, the entire spectrum shifts, so that the frequencies we see as red at midday are different than the frequencies we see as red at dusk. The evolutionary advantage of this phenomenon is clear enough, for we are able to recognize objects, say a ripe fruit, as the same color despite variations in lighting. At first, this seems to point to the existence of qualia that are intrinsic to the viewer. Dretske and Tye are quick to explain this phenomenon in representational terms, as a second-order processing of the initial wavelength information. Indeed, such processing needn't necessarily result in any qualitative state whatever. However, I would suggest that if such a system correlates reliably with phenomenal states in humans, it is reasonable to suppose that a similar system in an animal would produce a similar qualitative state. It would be stranger to suppose that evolution would produce two similar systems that produce vastly different effects. Simple behavioral tests could detect such second order visual systems in animals, if they exist, and might suggest which animals experience color like we do.

Representational theories of consciousness have an advantage over other theories of mind in that they offer an explanation of mentality that could arise by natural selection in very simple systems. They provide for the gradual development of consciousness rather than a sharp distinction between humans and other animals. This agrees with our modern scientific intuitions, and an account that attributes consciousness to fish and bees seems to make it less mysterious that humans could be conscious. Unfortunately, Tye's views about fish and bees rely on two controversial claims, that beliefs are required for learning and that the capacity to form beliefs is required for phenomenal consciousness. The former, though still the orthodox view, is challenged, I think formidably, by phenomenologists like Dreyfus. The latter, that seeing requires believing, relies on an explanation of the pathological case of blindsight that does not take into account the caveats associated with pathological cases nor the possibility of other explanations. I have argued that instead, the standard scientific approach of using similarities in behavior and biological structure to predict similarities in experience, while far from offering any certainty or hope of closing the explanatory gap, is still the best approach. By this method we can at least find the best places on the phylogenetic scale to look for consciousness, should scientific and philosophical developments ever reach the point where they might explain it.

Bibliography

Allen, Colin. (2010) "Animal consciousness," in: *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/entries/consciousness-animal/

Block, Ned. (1997) "On a confusion about a function of consciousness," in: N. Block, O. Flanagan, & G. Guzeldere (Eds), *The Nature of Consciousness*, pp. 375-416 (Cambridge: MIT Press)

Carruthers, Peter. (2011) "Higher-order theories of consciousness," in: *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/entries/consciousness-higher/

Dretske, Fred. (1995) *Naturalizing The Mind* (Cambridge: MIT Press)

Dreyfus, Hubert. (2007) “The return of the myth of the mental,” *Inquiry* 50(4), pp. 352-365

_____. (2005) “Overcoming the myth of the mental: How philosophers can profit from the phenomenology of everyday expertise,” *Proceedings and Addresses of the American Philosophical Association* 79(2), pp. 47-65

Jackson, Gabrielle. (2011) “Motor-intentionality and the role of the pathological in Maurice Merleau-Ponty’s work,” Retrieved from www.bu.edu/conscious/2011papers/Jackson.doc

Merleau-Ponty, Maurice. (1945) *Phenomenology of Perception* (New York: Routledge Classics)

Putnam, Hilary. (1973). “Meaning and reference,” *Journal of Philosophy* 70(8), pp. 699-711

Tye, Michael. (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge: MIT Press)

_____. (1997) “The problem of simple minds: Is there anything it is like to be a honey bee?” *Philosophical Studies* 88(3), pp. 289-317

CONCEPT POSSESSION, AND COLOUR-AND-SHAPE CONCEPTUALISM[1]

Adam Sanders

I. INTRODUCTION

The debate concerning whether or not perception bears conceptual content has been a central focus within the philosophy of mind for quite some time.[2] The *content* of a given perceptual state can be properly understood to be the way that state represents the environment (Bermúdez 2007, p. 56). On one side of the dispute, nonconceptual theorists have traditionally maintained that perceptual experience is constituted and exhausted by nonconceptual content. The notion of *nonconceptual* content, as it is used in the context of perception, can be thought of as a kind of content that merely represents physical objects and properties in the environment without depending on the concepts that specify that content (Bermúdez 2007, pp. 68-69). In contrast, conceptual theorists have argued for the claim that the content of perception can be conceptual.

The content of a given thought can be taken to express some proposition that has conceptual content as its constituent parts. Stephen Schiffer characterizes propositions as *that-clauses*, which are abstract, mind-and language-independent entities that have truth conditions (2006, pp. 1-4). The idea that propositions and their conceptual content are abstract objects (i.e., they are both non-temporal and non-spatial entities) has its roots in the traditional Fregean view of propositions. Gottlob Frege asserts that *thoughts* are abstract objects that do not depend on any particular mind or language for their existence, whereas *ideas* are token mental instantiations that correspond to thoughts (1956, pp. 299-302). According to this view, a subject can mentally bear the idea *that snow is white*, which corresponds to an abstract proposition composed of the abstract concepts *snow* and *white*. The Fregean view of concepts is controversial, in that if concepts are abstract objects, then they are causally inefficacious; and this mystifies the relation between concepts and their corresponding mental states (Margolis and Laurence 2007, p. 580).[3] For the purpose of this paper, I will skirt these issues and assume that concepts are mental representations that express what is putatively held to be abstract.[4]

Arguments for the conceptual content of perception are usually motivated by epistemological concerns, in the sense that in order for beliefs with empirical content to be justified and grounded in perceptual experience, the content of perception must be characterized in conceptual terms that can stand in the correct justificatory relation with beliefs (Kelly 2001, p. 402). Kevin Connolly has advanced an account called *Colour-and-Shape Conceptualism* that attempts to accommodate this epistemological concern by contending that perceptual content is inherently conceptual. Colour-and-Shape Conceptualism is the view that in order for a subject to perceive a given color or shape, the possession of the concept for that color or shape must occur at the sub-personal level prior to the conscious perception of it (Connolly 2011, p. 243). In order for Colour-and-Shape Conceptualism to be plausible, the sub-personal level of the visual system must satisfy the necessary and sufficient conditions for possessing color and shape concepts.

My thesis in this paper is to argue that Colour-and-Shape Conceptualism is untenable, in that the conditions for concept possession that the view necessarily depends upon are not sufficient for crediting the sub-personal level of perception with color and shape concepts. The argument for my thesis follows a three-fold approach, and will proceed as follows. In the second section following this introduction, I will briefly discuss the fineness-of-grain argument and how Kevin Connolly attempts to overcome the argument with Colour-and-Shape Conceptualism. In the third section, I will discuss the desiderata of *intentionality*, *categorization*, and *publicity* required for a theory of concepts. In the fourth section, I will argue that the conditions for concept possession that Colour-and-Shape Conceptualism requires do not jointly satisfy the desiderata for concepts, and that the view is therefore untenable.

II. Fineness-of-Grain and Colour-and-Shape Conceptualism

In his Varieties of Reference, Gareth Evans posed the following question: “Do we really understand the proposal that we have as many color concepts as there are shades of color that we can sensibly discriminate?” (1982, p. 229). This question has since been expanded upon and developed into the *fineness-of-grain argument*, which is routinely employed in philosophical literature to demonstrate the nonconceptual content of perception. The fineness-of-grain argument contends that since perceptual content has an extremely fine-grained quality that outstrips our limited conceptual repertoire, perception must be nonconceptual (Bermúdez 2007, p. 60). The argument can be expressed as follows:

(1) A given state is conceptual if and only if the content of that state is completely characterized in conceptual terms.

(2) There are perceptual states with fine-grained contents that are not completely characterized in conceptual terms.

(3) Therefore, perception is not inherently conceptual.

One caveat that should be addressed is that the fineness of grain argument does not entail that perception is *essentially* nonconceptual. Richard Heck claims that perceptual states are usually differentiated in terms of being concept-independent, while belief and cognitive states are considered to be concept-dependent (2000, pp. 484-485). It might be argued that characterizing perceptual states as concept-independent does not entail that those states must necessarily exclude conceptual content. Jeff Speaks illustrates this point by differentiating between *absolutely* nonconceptual states and *relatively* nonconceptual states. Speaks claims “a mental state has absolutely nonconceptual content if and only if that mental state has a different kind of content than do beliefs, thoughts, and so on” (2005, p. 360). In contrast, Speaks contends that “a mental state of an agent *A* (at a time *t*) has relatively nonconceptual content if and only if the content of that mental state includes contents not grasped (possessed) by *A* at *t*” (2005, p. 360). In other words, a subject’s perceptual state might be constituted, at least partly, by conceptual content if the perceiving subject is in a state to grasp the relevant concepts that characterize their perceptual experience.

In contrast to the view that perception is relatively nonconceptual, Kevin Connolly has advanced a strong conceptualist account of perception contending that the conclusion that the fineness-of-grain argument attempts to establish is false. Connolly argues for the thesis that he calls *Colour-and-Shape Conceptualism*, which states that a subject is capable of having a perceptual experience of a given color or shape if, and only if, she possesses the concept of that type of color or shape first (2011, p. 243). Connolly’s thesis is an entailment of the view that he characterizes as *Conceptualism*, which claims that “a subject can have a perception of some property only if she possesses a concept of that type of property” (2011, p. 244). In other words, the capacity to perceive the rich amount of colors and shapes that constitute our visual experiences is made possible only by a prior possession of the concepts for those types of colors and shapes.

Colour-and-Shape Conceptualism necessarily hinges on the conditions of satisfaction for concept possession. According to Connolly, to possess a concept *F* is to have the capacity to type-identify *F* things as being *Fs* (2011, p. 246). For example, a subject’s capacity to type-identify a given token instantiation of the property red, as being of the *type* red, is sufficient for the possession of the concept *red*. In Connolly’s view, this criterion for possessing shape and color concepts is satisfied at the sub-personal level of perception. Connolly defines sub-personal concept possession in the following way: an organism sub-personally possesses a concept *C* of a color or shape *F* if and only if the organism is able to type-identify *F*-things prior to consciousness (2011, p. 251).

Sub-personal type-identification of colors and shapes is an innate process, in which the visual system categorizes incoming sensory information according to color and shape types prior to any conscious perceptual awareness of those colors and shapes. In other words, the ability to be consciously aware of color and shape properties in perception is a direct consequence of prior sub-personal type-identification proceeding from the neuronal and sub-personal level of the visual system. For example, Connolly claims that the visual system is able to identify and reconstruct the two dimensional and upside down image on your retina prior to your perceptual experience of that image (2011, p. 248). If the sub-personal level of the visual system is capable of type-identifying shapes and colors prior to conscious perceptual experience, then the visual system satisfies the conditions for possessing the concepts of those shapes and colors.

The type-identification of *F*-things being *Fs* that occurs at the conscious and personal level of perception is simply the end state of the classification process (Connolly 2011, p. 248). For example, the ability to consciously identify red things as being of the *type* red at the personal level is made possible by the prior possession of the concept *red* at the sub-personal level of perception. The hard-wired sub-personal categorization of sensory information occurring in the visual system is a necessary condition for the formation of conscious perceptual experiences of colors and shapes. Therefore, our perceptual concepts are not outstripped by perception, but rather, our prior sub-personal possession of color and shape concepts are necessarily involved in the production of our perceptual experiences.

At face value, the ability to type-identify *F*-things as being *Fs* might seem like an inadequate condition for possessing some concept *F*. More precisely, the criterion would credit the possession of concepts to anything that can engage in type-identification behavior. For instance, there are obvious cases of machines that can type-identify things; and it seems counter-intuitive to credit these machines with concept possession (Connolly 2011, p. 246). The result is a *reductio ad absurdum*, in that concept possession would most likely be a ubiquitous phenomenon in nature. However, Connolly claims that if concepts are either mental objects or abstract entities, then the worry about crediting concepts to things like classifying-machines will go away (2011, p. 246). If concepts are stipulated as mental representations that express what is putatively held to be abstract, then only subjects with mental representations can possess them.

III. Desiderata for Concepts

Concepts are characterized as the constituents of propositions that thoughts can express, and are fundamental to both thought and language. In addition to their sub-propositional roles, concepts are thought to have other essential properties. Jesse Prinz has contended that there is a list of desiderata encompassing these essential properties of concepts that any given theory of concepts should ideally explain (2002, pp. 2-3). I will not attempt to delve into a full account of all of the desiderata that Prinz has advanced. Rather, I will discuss three core features of concepts that I consider to be both uncontroversial and essential to any theory of concepts. These desiderata are *intentionality*, *categorization* and *publicity*.

(i) ***Intentionality***. Concepts can be thought of as referential devices, in that they pick out or denote objects and properties that fall under them. Prinz claims that "to say that concepts have intentionality is to say that they refer, and those things to which they refer, I call their intentional contents" (2002, pp. 3-4). Another way of putting it is that concepts have semantic properties, in that they represent things. Similarly, Tim Crane has asserted that "all mental states exhibit what has been called 'aboutness' or 'directedness': they are about or directed on things" (2009, p. 454). For example, a conceptual state like a belief *that-P* has as its intentional object the proposition *P* along with the concepts that constitute *P*. If concepts are tokened mental representations, then the content of those representations are constituted by what they are directed at. However, it is crucial to note that the intentional properties of concepts do not necessarily refer to physically existing things, in that they can also refer to possible or abstract objects (Prinz 2002, p. 4). For example, the concept *three* represents a number; and numbers and other mathematical objects are usually held to be abstract objects. For the purpose of this paper, empirically observable concepts like the color concept *red* will typically refer to physical things that can be represented in perception.

(ii) ***Categorization***. Another important feature of concepts is that they individuate categories. According to Prinz, a given concept's reference to a set of things is a semantic relation, whereas the set of things to which a concept is *taken* to refer is an epistemic relation (2002, p. 9). In other words, concepts have an epistemological function, in that they are invoked in order to classify objects and properties into their appropriate categories. Prinz claims that categorization encompasses two distinct, but closely connected abilities: (1) *category identification*, which is manifested when a given subject identifies the category under which some object belongs; and (2) *category production*, which is manifested when a given subject is able to identify which attributes or properties some object possesses if it is a member of a given category (2002, p. 9). To illustrate this epistemic feature further, a given subject's possession of the *kangaroo* concept enables both her ability to track and identify kangaroos as falling under the category *kangaroo*, as well as her ability to identify the appropriate characteristic properties that a given thing must possess in order for it to count as a member of the category kangaroo.

(iii) ***Publicity***. The final desideratum for concepts that will be discussed is the publicity requirement. In the introduction of this paper, concepts were stipulated to be mental representations that referred to what is putatively held to be abstract. The publicity desideratum applies more flesh to this original formulation by further elucidating the fact that concepts must be both mind-and language-independent entities. Prinz claims that "concepts must be capable of being shared by different individuals and by one individual at different times" (2002, p. 14). For instance, the singular proposition *that-P*, along with its constituent concepts, is not dependent on any particular mind or language to token or express it. According to Prinz, people are capable of understanding each other's words in virtue of the fact that they associate the same concepts with those words; and thus concepts must be sharable in order for communication to be possible (2002, p. 14). Given the shareable quality of concepts in language and thought, any plausible theory of concepts must account for this phenomenon.

Common sense intuition would seem to require that any plausible theory concerning concepts and their possession should adequately accommodate the aforementioned desiderata of *intentionality*, *categorization* and *publicity*. It is my intent in the following section to argue that the criterion for concept possession that Colour-and-Shape Conceptualism necessarily requires does not jointly satisfy these desiderata.[5]

IV. Conditions for Concept Possession

Colour-and-Shape Conceptualism follows from the claim that the sub-personal level of perception possesses fine-grained concepts for shapes and colors. The sub-personal level of the visual system is the antecedent cause for conscious perceptual experience of colors and shapes. More precisely, the visual system systematically processes and categorizes inputs of electromagnetic radiation into sensory information that can be represented by perceptual states. The contention is that the possession of some concept *F* follows from the ability to type-identify *F*-things as being *Fs*. If the operant processes of the visual system function according to the ability to type-identify inputs of sensory information, then the sub-personal level of perception possesses fine-grained concepts for colors and shapes *ex hypothesi*. For any given color or shape concept *F*, the argument for the sub-personal level's possession of *F* can be expressed as follows:

(1) If type-identification of *F*-things as being *Fs* occurs, then concept possession of *F* occurs.

(2) The sub-personal level of perception does type-identify *F*-things as being *Fs*.

(3) Therefore, the sub-personal level of perception possesses the concept *F*.

The above modus ponens argument is obviously valid. However, the argument is not a sound one. It is my intent in this section to argue that the mental ability to type-identify *F*-things as being *Fs* is not a sufficient condition for the possession of *F* at the sub-personal level of vision.

Connolly explicitly asserts that the possession of the concept *ostrich* at the personal level requires the ability to type-identify ostriches as ostriches; and this same kind of type-identification of shapes and colors occurs in the visual system prior to the conscious perception of those shapes and colors (2011, pp. 246-248). However, the type-identification that occurs at the personal level manifests itself differently than the type-identification that proceeds at the sub-personal level of vision. More precisely, the ability to type-identify ostriches as being ostriches at the personal level is a mental ability involving a judgment that some object *x* is an *F*, where *F* denotes the concept *ostrich*. In contrast, the visual system is a hard-wired causal mechanism that begins with the categorization of sensory information, and ends with an output of perceptual states that represent that information. Therefore there would seem to be equivocation occurring between these two manifest forms of type-identification.

While it might be appropriate to credit the possession of a given concept *F* to a subject that demonstrates the mental ability to type-identify *F*-things as being *Fs* at the personal level, it isn't intuitively obvious that the kind of type-identification occurring in the visual system is also indicative of the sub-personal level's possession of fine-grained color and shape concepts. This raises a pressing question: What individuates a fine-grained concept for a given type of color or shape such that the sub-personal level of the visual system can type-identify things according to that concept? The only two possible answers that come to mind are either:

(1) Concepts for colors and shapes are innate sub-personal mental abilities in perception; or (2) Concepts for colors and shapes are fine-grained and innate sub-personal perceptual states that guide the type-identification occurring in the visual system. Neither of these views seems plausible.

Admittedly it seems that a subject's capacity to possess a given color or shape concept presupposes some sort of mental ability. However, the type-identification that occurs at the sub-personal level of perception would be the wrong sort of mental ability for individuating fine-grained color and shape concepts. Wayne Davis argues that concepts differ fundamentally from abilities, in that we can conceive concepts; but it is a category mistake to claim that what we are conceiving is some ability (2005, p. 144). For instance, a subject's possession of the color concept *red* entails that she has the capacity to conceive *red*. However, it is not the case that when a given subject is conceiving the concept *red* that they are conceiving the sub-personal mental ability that produces the tokening of a perceptual state that represents redness. Moreover, concepts have semantic properties that can be true or false of objects, whereas abilities do not (Davis 2005, p. 145). Therefore, the view that color and shape concepts are individuated by abilities involving type-identification at the sub-personal level of perception should be rejected.

The last refuge for Colour-and-Shape Conceptualism is the view that the sub-personal level of perception possesses fine-grained color and shape concepts in the form of innate perceptual states that not only guide the type-identification behavior occurring in the visual system, but can also be tokened in perceptual thought. The *prima facie* problem with this view is that it requires a perceiving subject to have an incalculable amount of innate concepts. Moreover, this view seems entirely implausible, in that it can only account for the desideratum of *intentionality*, while failing to accommodate the desiderata of *publicity* and *categorization*. For example, if *L* is some particular wavelength of visible light, *P* is perceived by a subject, *S* is a mental state, and *Rxy* is a representation relation, then we can express this idea in the following logical form:

$$(\forall x)[(Lx \,\&\, Px) \rightarrow [(\exists x)(Ryx \,\&\, Sy)]]$$

What this demonstrates is that *y* is just a specific sub-personal perceptual state that can represent a *particular* instance of some fine-grained sense datum before it is made consciously available as a mental representation in perceptual experience. It does not show, however, that *y* itself is directed at some abstract *category* or *type* that tracks particular instances of a perceivable property. This point is similar to Fred Dretske's claim that a given subject's conscious experience of some object *x* does not necessarily entail that she is aware of the fact that *x* is an *F*, where *F* is some concept (1993, p. 266). This principle is applicable to the sub-personal level underlying conscious perceptual states. The ability to recognize some state *y* as representing a category *F* would seem to require a further mental event that occurs during or after the conscious perception produced by *y*. To claim otherwise would be begging the question. Therefore the view would militate against the desideratum of categorization.

If color and shape concepts are individuated and possessed innately as sub-personal perceptual states that guide type-identification behavior, then concept possession entails psychologism. *Psychologism* is the view that all of our concepts, logical laws, and mathematical objects reduce to psychological facts. Edmund Husserl clearly expresses the problem with psychologistic accounts as follows:

> If the laws of logic have their epistemological source in psychological matter of fact, if, e.g., as our opponents generally say, they are normative transformations of such facts, they must themselves be psychological in content, both by being laws for mental states, and also by presupposing or implying the existence of such states. This is palpably false. No logical law implies a 'matter of fact', not even the existence of presentations or judgments or other phenomena of knowledge (2001, pp. 38-39).

Psychology is an empirical science that traffics in the production of inductive generalizations stemming from particular events and facts. The reduction of concepts, logic and mathematics to psychological facts relativizes that which is intended to be objective (Husserl 2001, p. 52).

According to the aforementioned desiderata, concepts are held to be abstract entities that are objective and independent of any particular mind or language for their existence. If color and shape concepts are individuated as innate sub-personal perceptual states that are the teleological ends to hard-wired causal processes in the visual system, then Colour-and-Shape Conceptualism conflicts with the desideratum of publicity.

It is true that a given subject's ability to process sensory information into color and shape percepts is inherently a product of human psychology. However, it is a category mistake to identify innate sub-personal perceptual states with the sort of mental states that instantiate concepts. More precisely, sub-personal perceptual states that propagate fine-grained color and shape percepts are functionally differentiated from the mental states that correspond to empirical concepts that *refer* to those percepts. It would thus be a mistake to identify color and shape concepts with the sub-personal perceptual states that produce their referents.

The ability to type-identify things would thus seem to be an insufficient condition for the possession of color and shape concepts. According to Jerry Fodor, conditions for concept possession that rely on mental abilities involving sorting or categorizing are dispositional or epistemic conditions, in that they involve a *knowing that* or *knowing how* (2004, p. 29). Applied to Colour-and-Shape Conceptualism, the visual system's ability to type-identify *F*-things as being *Fs* is a dispositional ability. However, using type-identification as a sufficient condition for the possession of fine-grained color and shape concepts relies on a circular argument. According to Fodor, the argument proceeds by claiming that (1) the ability to type-identify *F*-things as being *Fs* depends on the possession of the concept *F*; and (2) to possess the concept *F* requires an ability to type-identify *F*-things as being

Fs (2004, pp. 39-40). I take Fodor's circularity objection to be fatal to the view that the ability to type-identify F-things as being Fs is a sufficient condition for the possession of a given color or shape concept *F*. Therefore, the view that the sub-personal level of perception possesses every fine-grained color and shape concept would seem incoherent.

An alternative (and more plausible) condition for the possession of some empirical concept *F* would be the ability to simply conceive *F*.[6] Tyler Burge claims that concepts are sub-components of thought contents; and to possess a concept is just to have the capacity to think thoughts that contain that concept (1993, p. 309).[7] Both Jerry Fodor and Wayne Davis subscribe to this condition for concept possession. For example, Fodor claims that in order for a subject to possess the concept *dog*, she must be able to think of dogs as such (2004, p. 31).

If this alternative condition for concept possession were to supplant the previous condition of being able to type-identify things, then Colour-and-Shape Conceptualism would entail something similar to what John McDowell calls *demonstrative* concepts. According to McDowell, a demonstrative concept for some color property can take the form of a demonstrative act like thinking *that shade*; and when a given subject has a perceptual experience of a color property that she lacks the general concept for, a demonstrative concept can be used to conceptually characterize the content of that perceptual state (1996, p. 57). Demonstrative concepts are a controversial issue, in that it is not conclusive whether these concepts actually exist.[8] For the purpose of this paper I will remain neutral on the plausibility of demonstrative concepts. However, if there are such things as demonstrative concepts, then their possession would only occur during a tokening of them in conscious thought. This consequence, however, would still be incompatible with Colour-and-Shape Conceptualism, which necessarily requires the possession of fine-grained color and shape concepts at the sub-personal level of perception *before* those colors and shapes can be perceived.

Given the discussion thus far, there doesn't seem to be any compelling reason to accept the claim that the sub-personal level of perception possesses fine-grained color and shape concepts simply because it has the capacity to type-identify sensory information. The ability to type-identify *F*-things as being *Fs* is demonstrably insufficient for an adequate possession condition of a given color or shape concept *F* at the sub-personal level of perception. If this is the case, then the first premise of the above argument for concept possession at the sub-personal level is false; and thus the argument is not sound. Therefore, Colour-and-Shape Conceptualism does not necessarily follow, and would be an untenable view.

V. Conclusion

In this paper I have attempted to show that Connolly's account of Colour-and-Shape Conceptualism is untenable. Colour-and-Shape Conceptualism is the view that a subject can perceive a color or shape property if, and only if, she possesses the concept for that color or shape prior to conscious perception. According to Connolly, to possess a concept *F* is to be capable of type-identifying *F*-things as being *Fs*. Connolly contends that the sub-personal level of the visual system does type-identify *F*-things as being *Fs*; and thus the sub-personal level of perception must possess concepts.

Connolly's Colour-and-Shape Conceptualism is implausible for several reasons. The view that the sub-personal level of perception possesses innate and fine-grained color and shape concepts does not jointly satisfy the desiderata of intentionality, categorization and publicity. Moreover, Fodor demonstrated that possession conditions for color and shape concepts that depend solely on the ability to type-identify is circular, in that it relies upon the following two claims: (1) To possess a concept F depends on the ability to type-identify *F*-things as being *Fs*; and (2) the ability to type-identify *F*-things as being *Fs* depends on the possession of the concept *F*. Given this, the ability to type-identify *F*-things as being *Fs* at the sub-personal level of perception would not be a sufficient condition for the possession of some fine-grained color or shape concept *F*. If this is true, then the argument for the possession of fine-grained color and shape concepts at the subpersonal level of perception is not sound. Therefore, Colour-and-Shape Conceptualism would be an untenable position to endorse.

Notes

1. I would like to thank Mark Balaguer, Jay Conway, David Pitt, Michael Shim, and Douglas Wadle for their insightful comments that guided me during the course of writing this paper.
2. See, for instance, José Bermúdez (2007), Kevin Connolly (2011), Fred Dretske (1993), Sean Kelly (2001), John McDowell (1996), and Jeff Speaks (2005).
3. See Margolis and Laurence (2007) for more information regarding the ontology of concepts, and the problems associated with positing the existence of abstract objects in theories about concepts.
4. I say *putatively* because someone could hold a fictionalist view about concepts. *Fictionalism* is the view that concepts are merely fictions that are used in order to represent the physical objects of the world in both thought and language.
5. See Jesse Prinz (2002) for an account of the other desiderata for concepts not mentioned here. These other desiderata are *scope*, *cognitive content*, *acquisition*, and *compositionality*.
6. See, for instance, Tyler Burge (1993), Wayne Davis (2005), and Jerry Fodor (2004).
7. I am agnostic about whether there should be one single criterion for concept possession. It seems plausible that different kinds of concepts might require different possession conditions. It could also be the case that an adequate set of conditions for possessing some concept *F* would require both the ability to type-identify *F*-things as being *Fs*, and the ability to conceive *F*. However, it seems intuitively obvious that the ability to type-identify *F*-things as being *Fs* is not sufficient by itself for crediting the visual system with the possession of some color or shape concept *F*.
8. See Sean Kelly (2007) for an argument against demonstrative concepts.

Bibliography

Bermúdez, José. (2007) "What is at Stake in the Debate on Nonconceptual Content," *Philosophical Perspectives* 21, pp. 55-72

Burge, Tyler. (1993) “Concepts, Definitions, and Meaning,” *Metaphilosophy* 24(4), pp. 309-325

Connolly, Kevin. (2011) “Does Perception Outstrip Our Concepts in Fineness of Grain?,” *Ratio* 24(3), pp. 243-258

Crane, Tim. (2009) “Is Perception a Propositional Attitude?,” *The Philosophical Quarterly* 59, pp. 452-469

Davis, Wayne. (2005) “Concept Individuation, Possession Conditions, and Propositional Attitudes,” *Nous* 39(1), pp. 140-166

Dretske, Fred. (1993) “Conscious Experience,” *Mind* 102, pp. 263-283

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

Fodor, Jerry. (2004) “Having Concepts: a Brief Refutation of the Twentieth Century,” *Mind and Language* 19(1), pp. 29-47

Frege, Gottlob. (1956) “The Thought: A Logical Inquiry,” *Mind*, 65, pp. 289-311

Heck, Richard. (2000) “Nonconceptual Content and the ‘Space of Reasons’,” *The Philosophical Review* 109(4), pp. 483-523

Husserl, Edmund. (2001) *The Shorter Logical Investigations* (London: Routledge)

Kelly, Sean. (2001) “Demonstrative Concepts and Experience,” *The Philosophical Review* 110(3), pp. 397-420

Margolis, Eric, and Laurence, Stephen. (2007) “The Ontology of Concepts—Abstract Objects or Mental Representations?,” *Nous* 41(4), pp. 561-593

McDowell, John. (1996) *Mind and World* (Cambridge: Harvard University Press)

Prinz, Jesse. (2002) *Furnishing the Mind* (Cambridge: MIT Press)

Schiffer, Stephen. (2006) “Propositional Content,” in: E. Lepore and B. Smith (Eds), *The Oxford Handbook of Philosophy of Language* (Oxford: OUP) (Manuscript)

Speaks, Jeff. (2005) “Is There a Problem about Nonconceptual Content?,” *The Philosophical Review* 114(3), pp. 359-398

Conceptualism and Causal Explanation of Empirical Conceptual Content

Douglas C. Wadle

Conceptualism is the view that perceptual states[1] carry—or can carry—conceptual contents. Anti-conceptualism is the view that the contents of perceptual states are non-conceptual. In what follows, I will argue that the traditional argument for conceptualism requires a modification in the form of an additional premise, if the conclusion (or something very near to it) is to follow. I formulate and then defend a premise that, I think, meets this requirement.

I. Conceptual Content

Traditionally, conceptual content is identified with the content of a thought—particularly with the sort of thought that can be shared by two or more thinking subjects (i.e., such contents are not mind-dependent)—and the contents of thoughts are taken to be propositions (e.g., that the sweater is blue) (Frege 1990 [1892], 1956 [1918/1919]).[2] Another way of describing conceptual content is to say that it is the sort of content that is composed of concepts (Crowther 2006, p. 250; Laurence and Margolis 1999, p. 4; Prinz 2004, pp. 12-14). These characterizations are compatible. Propositions, however understood (e.g., singular propositions, general propositions, sets of possible worlds, etc.), are abstract objects existing outside of space and time, and concepts, too, are usually understood as abstract objects.[3] That is, there is no metaphysical problem resulting from the claim that concepts are constituents of propositions.

Of course, one might reject this platonistic (i.e., abstract object-based) account while continuing to believe in conceptual contents—e.g., one might be a nominalist and espouse a version of content externalism (i.e., the conceptual content is socially determined, as described in Putnam's twin earth examples (1973, 1975)). Any reference to conceptual content herein should be taken as referring to the sorts of contents that are composed of concepts and which are mind-independent contents of our thoughts while remaining neutral with respect to the nominalist/platonist debate.

There are constituents of thoughts that are not concepts (at least on some of the common views—e.g., among adherents of singular propositions), such as the sweater (i.e., the thing itself) in the thought that the sweater is blue. The form of this content is either (*Sa* & *Ba*), if one thinks that the definite description carries the content that *a* is a sweater, or *Ba*, if one thinks that the definite description just denotes the singular term, *a*. (Which view of definite descriptions one subscribes to doesn't matter for present purposes.) Those things that might qualify for concept-hood, in the example, are the predicates *S* and *B*, which correspond to the properties of *being a sweater* and *being blue* and, therefore, to the concepts SWEATER and BLUE.

I belabor the point because it will help to clarify the sort of conceptual content on which this paper will focus. These are empirical conceptual contents, by which I mean something roughly equivalent to the content of Quine's observation sentences:

> Viewing the graded notion of observationality as the primary one, we may still speak of sentences simply as observation sentences when they are high in observationality. In a narrow sense, just 'Red' would qualify; in a wider sense, also 'Rabbit' and 'The tide is out'. It is for observation sentences in some such sense that the notion of stimulus meaning constitutes a reasonable notion of meaning (1960, p. 44).

Here, "stimulus meaning" corresponds to some perceptual content that would cause one to assent to a question asking if such perceptual content is present—e.g., the stimulus meaning of "red" will be all the perceptual contents that would cause competent language users to answer in the affirmative when asked, “Is this red?” (more properly, just “Red?”); the stimulus meaning for “The tide is out” will be the range of perceptual contents that would lead to affirmative answers to the question “Is the tide out?”. So, empirical conceptual contents are, roughly, stimulus meanings. More precisely, empirical conceptual contents are those contents whose non-logical constituents are reducible to concepts, which we can call observation concepts, corresponding to perceivable properties, like *being blue*, and to perceived particulars, which function as individual terms denoting objects and the like.

I will not propose any rule for drawing a principled distinction between what does and does not count as an observation concept. I am certain that the boundaries will be fuzzy (this is why Quine admits degrees of observationality). For instance, I can imagine someone disputing the fact that *being a sweater* is a wholly perceivable property and, hence, that SWEATER is an observation concept. I do assume that there are such concepts, though, and that they comprise the conceptual constituents of empirical conceptual contents. Now we can turn to the traditional argument for conceptualism.

II. The Traditional Argument

The argument relies on the notion of intentionality, which is the relation of aboutness—i.e., some content, the intentional content, is about something else, its intentional object.[4] Conceptual content is considered a species of intentional content (or, for those that don’t believe in non-conceptual content, the two are identical). For instance, the conceptual content that the sweater is blue expresses the fact that the thing under consideration (the intentional object) is blue and (perhaps) that it is also a sweater. Now we can formulate the traditional argument in favor of conceptualism:

(P1) Conceptual content is a species of intentional content.

(P2) Intentionality is a causal relation in which the intentional object—that which the thought content is about—causes the intentional content. (At least where the intentional object is a physical object or state of affairs.)

(P3) Perception is a causal relation between an immediately present physical state of affairs and a perceptual content-bearing state (i.e., some version of the causal theory of perception is true).[5]

(P4) Physical states of affairs cannot be the *immediate* cause of any non-perceptual content-bearing state (i.e., there's no way for the external world to directly impinge on our consciousness than through perception).[6]

(C) Therefore, if a state bears conceptual content about an immediately present physical state of affairs or object (i.e., not a remembered or imagined state of affairs), then that state just is a perceptual state.[7]

I take the premises to be pretty widely endorsed among philosophers. The most contentious of these is (P2), which will be subject to further clarifications in section 3. (P2) is one way to state the causal theory of content (Stampe 1977; Dretske 1981, 1983, 1988; Fodor 2003 [1987]).[8] The causal theory of content finds a form of rudimentary intentionality in certain causal relations between physical systems (e.g., a barometer expresses the current air pressure due to the effects of that air pressure on the barometer) and includes perception among those systems. Therefore, observation concepts can be entirely causally explained. The causal content theorist hopes to extend this account of observation concepts to all concepts, thereby showing intentionality to be causal through-and-through:

> So what's proposed is a sort of foundationalism. The semantics of observation concepts is indeed special: First, in that—given an intact observer—the nomologically suffi-

> cient and semantically relevant conditions for their tokenings are specifiable 'purely externally'; viz., purely psychophysically. And second, in that all the other semantically relevant symbol/word linkages run via tokenings of observation concepts. 'Horse' means *horse* if 'horse' tokening are reliably caused by tokenings of psychophycical concepts that are in turn caused by instantiations of psychophysical properties for which instantiations of *horse* are in fact causally responsible. (Fodor 2003 [1987], p. 295)

The causal theory of content has been objected to on the grounds that it cannot explain content about abstract objects or vacuous terms, but—say whatever you like about the prospects for extending the semantic account beyond observation concepts—what concerns us at present are observation concepts.[9] But it is not my intent to argue for (P1)–(P4). I want to address a more pressing worry for the conceptualist; namely, that (C) does not follow from these premises, even if they are true.

The fact that (C) doesn't follow is clear from the initial plausibility of Davidson's suggestion that perceptual states may causally instantiate conceptual content states (2002, p. 143). I want to offer reasons for rejecting this claim. In doing so, I will argue for the inclusion of the following premise in the argument so that the conclusion (or something very near to it) *will* follow: (P5) perceptual content-bearing states do not cause the instantiation of further non-perceptual content-bearing states. In the course of the discussion, this will be revised a bit, but it is a good enough approximation with which to begin. Without (P5), the conceptualist argument fails regardless of the truth of (P1)–(P4) and, even if any one of (P1)–(P4) turn out to be false, (P5) is still independently interesting.

With the addition of (P5), (C) does follow. (P3) shows that perception, like intentionality, is a causal relation between an object and a content-bearing state. (P4) restricts the possible immediate content-bearing effects of a physical state of affairs to a perceptual content-bearing state. (P5) imposes a stop on any further causation of content-bearing states, due to physical states of affairs, beyond the perceptual content-bearing state. Then, the only thing that can satisfy the definition of intentionality in (P2), with respect to presently perceived physical states of affairs, is perception. So, any intentional content regarding presently perceived physical states of affairs must be borne by a perceptual content-bearing state. Therefore, by (P1), any conceptual content regarding presently perceived physical states of affairs must be borne by a perceptual content-bearing state because (P1) just says that conceptual content is a form of intentional content.

Before defending (P5), I will attempt to clarify some loose ends and ambiguities in the argument as currently formulated. I then argue that, in order to reject (P5), the anti-conceptualist must accept that all the pieces required to make conceptualism true are in place but still maintain that there is some redundant mechanism overlaid on top of those pieces that does the work of actually expressing empirical conceptual content.

III. Clarifications

I want to begin my clarifications by addressing a potential concern with the conceptualist argument: that it entails that we can only think about blue sweaters and such when we are having some perceptual episode that has the blue sweater as its content. This is not correct. The argument does not entail that the perceptual content *is* the conceptual content. To get that result we would need to add a further premise that states that the perceptual state bears no content other than its perceptual content, but we don't want that result (because it's obviously wrong)[10] so we can do without this further premise. The addition of (P5) might seem to force us to say that these are identical because the perceptual state is blocked from causing a further conceptual content-bearing state, but this does not preclude the possibility of the given perceptual state also expressing the conceptual content, which is just what the conceptualist wants.

Conceptualists want to say that a perceptual episode bears conceptual content but that this content could be expressed by some other kind of mental state *under the right conditions*—e.g., seeing the blue sweater might lead one to entertain the content that the sweater is blue, but so might someone telling one that the sweater is blue or your remembering the fact that the sweater is blue, etc. The right conditions are a (purported) causal connection to some observer's actual experience of the relevant conceptual content-bearing perceptual state.

Furthermore, for any empirical conceptual content, *p*, to be true there must be a (genuine) causal chain from the physical object or state of affairs, o, that is the intentional object of *p*, to *p*. Call this the rule of first contact. When we consider the traditional argument for conceptualism, it should be understood that what is at issue is the ability of a perceptual state to serve as first contact in establishing any empirical conceptual content about an immediately present physical object or state of affairs—particularly in establishing the truth conditions for that conceptual content. Without this initial contact, it seems that no empirical thoughts will ever have truth conditions grounded in the physical world.

What (C) does entail is that the conceptual content expressed by the perceptual state is identical to or includes what is expressed in thoughts about that object or state of affairs. Add to this the condition that the given perceptual state necessarily expresses that content and we arrive at the thesis known as content conceptualism. The problem I find with content conceptualism is that it offers no account of why, when we are in a given perceptual state, we do not access all of its potential conceptual contents. Nor does it explain the role of concept acquisition for the expression of the relevant content.

Recently, an alternative version of conceptualism, state conceptualism, has arisen (Crowther 2006; Speaks 2005). The idea, as Crowther has put it is that "Where *S* has an experience, *e*, with the content *p*, *p* is a conceptual content iff in order for *S* to be undergoing *e*, *S* must possess the concepts that characterize *p*" (2006, p. 252). For instance, seeing a blue sweater (or hearing "The sweater is blue") can cause, in the perceiver, the instantiation of the conceptual content that the sweater is blue if, and only if, she possesses the concepts SWEATER and BLUE. I will add the claim, in keeping with the causal theory of content, that these will, at least, reduce to observation concepts of a sort that can only be acquired through first-hand perceptual experience.[11]

From this it follows that—for all non-innate observation concepts—no one, at any point along the causal chain specified in the rule of first contact, can entertain empirical conceptual contents without having already acquired the concepts out of which that content is composed, and these will reduce, at least in part, to observation concepts of a sort that can only be acquired through first-hand perceptual experience. Call this the rule of fundamental observation concept acquisition. The mechanisms behind both of these rules will be discussed in more detail in section 5. The reformulation of the conceptualist argument, at which I will arrive later in this section, is an argument for (primarily non-nativist) state conceptualism—i.e., one that uses the two rules in tandem.

We now turn from the conclusion to the premises. I have said that I accept (P1)–(P4) as uncontroversial. This requires some clarification, particularly with respect to (P2). What it means to say that intentionality, where the intentional object just is a physical object or state of affairs, is a causal relation is, only roughly, that the intentional object causes the instantiation of the intentional content. More precisely, the intentional object causes or instantiates a mental state that *expresses* the intentional content. The nature of this expression, for conceptual content, is a relationship between the mental state and, on the platonistic account, an abstract object (i.e., the proposition expressed). On this view, the expresses-relation cannot be a causal relation because it occurs between a mental state and an abstract object, and causal relations only obtain between physical objects (i.e., there is causal closure of the physical). Therefore, intentionality—even when restricted to content about things in the physical environment—cannot be wholly a species of causation. However, it does retain a causal component: the causation of a mental state by some object.

The nominalistic view of conceptual content, by contrast, has the merit of making the expresses-relation far less mysterious than it seems on the platonistic view (and in such a way that vindicates the intentionality of empirical conceptual contents as causal through-and-through). The basic idea is that there is a causal chain running from an ostensive fixing of a term's reference (Kripke 1980, Putnam 1973) to present-day usage by a linguistic community. An individual is connected, causally, to that chain in virtue of his acquisition of that reference through subsequent instances of ostensive fixing, thereby forming a mental representation that is sufficient for expressing the content. Notice that both views (and fictionalism regarding conceptual content, too) depend on the three-place characterization of intentionality: an intentional object causes (where that object just is a physical object or state of affairs) a mental state that expresses the conceptual content, whether understood as a proposition, a composite of causally fixed social meanings, or a non-existent (fictional) object). The revised argument is concerned with the causal connection between the intentional object and the mental state (the first two elements of the three-place relation of intentionality) for empirical conceptual contents.

Of course, mental states are not obviously physical, either, so the causal closure of the physical presents a problem here, too, but one that can be easily circumvented by accepting the sUniversity Presservenience of mental states on neural states, where neural states *are* physical things. Given these clarifications we can modify the argument as:

(P1) Conceptual content is a species of intentional content.

(P2') Intentionality includes a causal relation between an intentional object (the cause) and a neural state (the effect), where the neural state expresses (via a supervening mental state) content about the object. (At least where the intentional object is a physical object or state of affairs.)

(P3) Perception is a causal relation between an immediately present physical object or state of affairs and a neural state, where the neural state results in a supervening mental state that has perceptual content.

(P4') Physical objects or states of affairs cannot be the immediate cause of any neural state upon which a non-perceptual content-expressing mental state supervenes.

(P5') The neural states upon which perceptual content-bearing mental states supervene do not cause further non-perceptual neural states upon which conceptual content-expressing mental states supervene.

(C') Therefore, a mental state expresses empirical conceptual content (i.e., conceptual content about empirical physical object or state of affairs) if and only if it supervenes on a perceptual neural state.

What I mean by a perceptual neural state just is a neural state composed of arrangements of neurons that are involved in perception, as characterized in (P3). A non-perceptual neural state is one that isn't composed of arrangements of neurons that are involved in perception. Now we can define two senses of perceptual states (i) mental states that are states comprised of perceptual content, and (ii) mental states supervening on perceptual neural states. To distinguish the two, I will use "perceptual state" to refer to perceptual states in sense (ii) and "perceptual content-bearing mental states" or, simply, "perceptual content" for sense (i). The core question in the conceptualism debate has been the relationship between perceptual content-bearing mental states and conceptual content-bearing mental states, with conceptualism maintaining that they are of the same sort and anti-conceptualism maintaining that they are of a different sort. If that sort is defined at the level of neural state, upon which such content supervenes, then (ii) is clearly a form of conceptualism.

Given our two rules for causal instantiation of content and (P5'), we find that fundamental observation concepts are acquired, via first contact, for each possessor of those concepts and that these are stored in the perceptual systems of the brain. More precisely, a neural state expressing such a concept is instantiated and stored in the perceptual systems of the brain. (This is what I will mean when I speak of concept possession in the remainder of this paper.) This allows us to drop the restriction to immediately present physical objects and states of affairs, a fact which is reflected in (C').

Notice, in this respect, that memories of perceptual contents are perceptual states, as I define them, because they supervene on neural states involved in perception. This does *not* entail that there is any occurrent sensory phenomenology as those memories (or even novel thoughts built on perceptual memories) are entertained. In terms of our rules of causal connection, we can say that memories, like perceptual states of one's immediately present physical environment, express content according to the rule of first contact (provided the perceiver possesses the requisite concepts) and the formation of novel thoughts, general thoughts, and other exercises of imagination express empirical conceptual content according to the rule of fundamental observation concept acquisition—i.e., they employ observation concepts as constituents, and those observation concepts either are fundamental observation concepts or are composed of observation concepts that can only be acquired through direct perception).

If my reformulation of the argument is sound, then anti-conceptualism entails that the neural states underlying the expression of observation concepts bear no causal relation to the physical things (their purported objects) instantiating the corresponding properties in perception because they violate the rule of first contact, and it is the rule of first contact that establishes truth-conditions for empirical conceptual contents. Without truth-conditions, these contents cannot be true. Furthermore, because truth is a necessary condition for knowledge,[12] any theory of empirical conceptual contents that entails that these contents cannot be true also entails that such contents cannot be candidates for knowledge. Therefore, anti-conceptualism is committed to skepticism concerning empirical conceptual contents, and—because these just are the sorts of contents that express things about the physical world—anti-conceptualism entails skepticism about the physical world.

Since we are granting, as true, the causal theory of perception and the causal theory of content (for empirical conceptual contents) along with the fact that there is no other access point for states of affairs to impinge upon our consciousness than our perceptual systems, the anti-conceptualist needs a causal connection between the neural states upon which perceptual content-bearing mental states supervene and mental representations (MRs), where mental representations are neural states that express conceptual content, via a supervening mental state, if he is to avoid skepticism.

We can refer to non-perceptual MRs as $\mathrm{MR}_{\mathrm{np}}$s, and perceptual MRs (whether or not we believe in them) as MR_{p}s. We can call a neural state an NS, where a perceptual NS is an NS_{p} and a non-perceptual neural state is an NSnp. Then we can present the dispute as whether an NS_{p} instantiates an MR_{p} (the conceptualist position) or an $\mathrm{MR}_{\mathrm{np}}$ (the anti-conceptualist position). This should make clear the significance of (P5'). (P5') straightforwardly denies the anti-conceptualist response (i.e., Davidson's response) to the problem of causal explanation of empirical conceptual content.

IV. Direct Causation

If the anti-conceptualist wants to reject (P5'), she must show that an NS_{p} can, and does, cause an $\mathrm{MR}_{\mathrm{np}}$. There are two ways in which this could be established. The first is to show a direct, or one-to-one, causation from a particular NSp to a particular MRnp. The second is to show a many-to-one causation (i.e., causation via an abstraction) from a set of NS_{p}s to a particular $\mathrm{MR}_{\mathrm{np}}$. I think that direct causation is not very plausible, but explaining why helps to set the stage for the discussion of many-to-one causation.

Direct causation entails that we have the content of a thought, e.g., that the sweater is blue, in virtue of the fact that the singular NS_{p} underlying the perceptual state in which the sweater is perceived causes an MR_{np} expressing the content that the sweater is blue. If this were true of the perceptual state, taken as a whole, it would entail that every possible perceptual state—all those that our NS_{p}s can produce—have corresponding MR_{np}s for every conceivable thought that could be had about that perceptual state. That is, the infinitely many potential perceptual states must each have infinitely many corresponding conceptual contents about them already present as MR_{np}s. They must be already present because, if the MR_{np} is unique to an individual NS_{p}, then there is no means of acquiring that MR_{np} because it is only instantiated once, when the particular perceptual episode supervening on that NS_{p} occurs. (Direct causation of this form doesn't work for MR_{p}s, either, for the same reason.) In the case of our blue sweater, that particular perceptual episode would have its very own MR_{np} expressing that the sweater is blue, as well as that the sweater is fuzzy, that the sunlight on the sweater is bright, etc. No other perceptual state involving a blue sweater will access any of these same MR_{np}s unless it is identical to the original perceptual episode, and this—if not strictly impossible—is surely close enough to impossible for the point to stand.

Perhaps, then, the MR_{np} can be caused by the particular arrangement of primitive components of a given perceptual state—i.e., qualia, which I construe here as irreducible perceivable qualities. A particular shade of a particular color is a classic example of a quale. In this case, the irreducible perceptual properties identifying the sweater as a sweater and as blue are instantiated by NS_{p}s that are *not* unique in the way that the total perceptual state is (i.e., the NS_{p} triggered by a particular shade of blue is the same for all instances of that shade), and so they do not require the absurd multiplication of conceptual contents seen in the preceding example. Though there still should be more than one possible MR_{np} for each NS_{p}, the number is substantially reduced and, we might suppose, this will alleviate the problem of too many innate concepts.

There are two responses to this move. The first is that there is still an infinite number of qualia,[13] each one of which requires more than one corresponding MR_{np} (e.g., BLUE, DARK BLUE, ROYAL BLUE, etc.), so the problem is not actually solved. (This objection applies to direct causation of an MR_{p}, too.) To solve the problem will require the kind of many-to-one causation to be discussed in the next section. The second response is that, assuming there weren't an infinite number of qualia and restricting the output of any NS_{p} instantiating some quale to a single MR_{np} (to avoid the problem just raised), then the content of that MR_{np} can only be what it is like to experience the given quale (because it is a primitive). If that is so, then attempting to compose, out of these primitive qualia-representing MR_{np}s, some intentional content regarding the larger perceptual state or state of the world presented through that perceptual state will only result in a description of the particular features of that perceptual state (composed of indicators for each quale present in that perceptual state), not in some more abstract content (i.e., conceptual content) regarding the properties of object-kinds (e.g., that the sweater is blue).

V. CAUSATION VIA ABSTRACTION

The anti-conceptualist could attempt to reject (P5') by taking a range of NSps as being individually capable of causing the same MRnp. Just such an answer is proposed by Dretske:

> Perception is a process by means of which information is delivered within a richer matrix of information (hence in analog form) to the cognitive centers for their selective use. Seeing, hearing, and smelling are different ways we have of getting information about s to a digital-conversion unit whose function is to extract pertinent information from the sensory representation… (2003 [1980], p. 30).

That is, describe a mechanism taking the rich input of perception and outputting a value (0 or 1—false or true) for the perceived object, *a*, concerning *a*'s *F*-ness. Since this mechanism must be causal, according to the causal theory of content, it must operate on the physical stuff of perceptual states (NS_ps) or indicators of those NS_ps (NS_{np}s) of the sort discussed in the preceding section. What we will need to do, then, is to define a set, *S*, of NSs (of one or the other sort) that are able to cause a given MR, and then see if we can determine whether this is an MR_{np}, as the anti-conceptualist maintains, or an MR_p, as the conceptualist believes.

S must either be possessed innately or it must be acquired through abstraction.[14] It is important not to think of *S* as a set of *finite* states defining a perceivable property, *F*, because a conjunction cannot accommodate the variable range of features abstracted—e.g., a set of shades of blue, joined by a conjunction, cannot deliver blueness with respect to some shade falling between two shades in *S*.[15] So, we can think of S as an infinite set containing a range of possible NS_ps. But this creates a problem. If *S* has an infinite number of members, then it cannot be the case that we ever actually have S complete in our heads.

The way to answer this concern is to posit the existence of an innate principle for ordering NS_ps. Then we would only require a set, *R*, of NS_ps defining a partition of the full range of the innate ordering. Call this partition *P*. Minimally, we must assign one ordering principle to each quale-type (e.g., color, pitch, temperature).[16] Call such an ordering principle *Q*. *P* can be a partition of a simple one-dimensional continuum, as it is in the case of a partition of a single *Q*, or it can be extraordinarily complex, having as many dimensions as it has *Q*s to be partitioned.[17] For simplicity's sake, I will focus on one-dimensional *P*s. Taking color as an example, we can say that the whole range of possible neural states (*S*) underlying color phenomenology are ordered according to an innate principle (*Q*), particular to color perception (hue perception, to be more exact), which can be segmented into units (*P*s) corresponding to color concepts such as BLUE, GREEN, and RED by a set of NS_ps (*R*) that defines the boundaries of *P*.

The *Q* for color might be imagined like the color wheel used by artists—a continuous variation of hue from red to violet, bleeding back into red, represented as a circle. The members of some *R* will be experiences of color phenomenology taken as falling under the same category that can be plotted on the circumference of the wheel. *P* will be the arc along the surface of the color wheel between the outermost *R*s—i.e., it will be the region of continuous variation in color phenomenology containing all the members of *R*. A point to bear in mind is that one can possess some *P* without possessing a term denoting that *P*. Also, notice that, for innate concepts, *R* will just be a set of innate boundary cases for a given *P*. It seems possible that there are innate concepts, but an *entirely* nativist picture of abstraction will not work because it entails that very possible *P* is innate—i.e., every possible combination of *Q*s are partitioned into every possible *P*—and that commits us to an impossibly large number of innate contents. So, acquisition is central to abstraction, and it is by abstraction that we hope to explain the causal connection between perceptual inputs and the neural underpinnings (MRs) of the expression of conceptual content about those perceptual inputs.[18]

An immediate concern with the proposed abstraction process is that it doesn't give any account of why any two or more perceived things are ever associated together such that they comprise an *R*. This is precisely the worry that led Laurence and Margolis to suggest the existence of innate fine-grained general representations, low-level abstractions that can be combined, in acquisition, to form higher-level abstractions. They describe a sample case as follows.

> In this case, a learner comes equipped for the task with general representation for different shades of white (among other colors), as well as an innate similarity metric that organizes her color space. Then upon encountering different instances of white things (snowballs, paper, milk, etc.) she would represent those particular shades, and through a process of positive and negative feedback, develop a representation that incorporates all of the shades that received a positive signal and none of the shades that received a negative signal (Laurence and Margolis 2012, p. 21).

My only complaint with this is that I don't see what work the innate representations are doing. If there is positive and negative feedback in the acquisition of WHITE, this is, presumably, feedback concerning proper identification of white things using the word “white”. But why isn’t the high correlation of instances in which one sees someone else describing various shades as white by using the unvarying perceptual item, “white”, sufficient (in combination with the innate ordering principle/similarity metric) to get the abstraction going?[19] This is not to say that there are no innate concepts (fine-grained or otherwise). It’s just to say that they don’t seem to be necessary for abstraction. Either way, we have two proposals for answering the objection, both of which seem to me to be plausible.

One might also worry that my account of abstraction cannot accommodate vagueness, and, surely, our color concepts offer a good example of vagueness—e.g., there is no point along the color continuum at which green becomes yellow. However, vagueness can be accommodated. For any set of *P*s on a given *Q*, it is highly unlikely that the *R*s defining *P*s have contiguous members (contiguous in terms of articulating a just noticeable difference along the given *Q*) such that the upper boundary of one *P* is contiguous with the lower boundary of the next *P*. Then there are regions of ^{Q} that remain undefined and, probably, predications of *F*-ness, for an *a* falling within an undefined region of *Q*, will correspond to a best-match with a partitioned region of *Q*. Alternatively, the relevant *P*s might overlap, and an *a* falling within the region of overlap will exhibit vagueness with respect belonging to one *P* or another (where the *P*s are not thought to be able to coincide).

Now we have the machinery to explain how we go from the perception of *a* to a mental state (supervening on a perceptual neural state) expressing the content *Fa*. I don’t have the space to mount a full description of what I have in mind, but I can sketch the outline in a sentence or two. Roughly, we can get from *a* to *Fa* if *a*, or some set of its features, are attended to with respect to some *Q*. Then, *a* will be assigned either no value (0 for each *F* denoting a *P* within *Q*) or a positive value (1 for the *F* denoting the *P*, to which a makes the best match within *Q*).

In fact, we can now describe two ways in which a neural state underlying perceptual contents can cause the instantiation of an MR that expresses empirical conceptual content. The first (corresponding to the rule of first contact) is that the NS underlying the perceptual content is attended to in the manner described above. We can call this post-acquisition causation. The other (corresponding to the rule of fundamental observation concept acquisition) is to say that the $\mathrm{NS_p}$ plays a causal role, along with other (similar) $\mathrm{NS_p}$s, in the acquisition of the abstraction, *P*. That is, the NS functions as a member of *R* in partitioning some *Q*. We can call this in-acquisition causation.[20]

Notice that in neither case is any particular $\mathrm{NS_p}$ necessarily tied to only one MR—that would just be the one-to-one causation discussed and dismissed in the previous section. In other words, there is nothing preventing overlapping *P*s and, where they do overlap, any perceptual content that falls within the region of overlap could be attended to with respect to either one (or both) of the given *P*s. This overlap might occur as a total containment of one *P* within another (e.g., ROYAL BLUE is contained within BLUE) or it might occur at boundary cases (e.g., CHARTREUSE will overlap both YELLOW and GREEN—indeed, YELLOW and GREEN might overlap at their boundary without an additional color concept being invoked).

So much for the machinery. Now we can see if it can be made to work in a way that allows the anti-conceptualist to dispense with (P5'). To do this we have to determine whether or not it is plausible that the abstraction (i.e., *P*) is housed outside the perceptual systems and whether or not it is plausible that there is some non-perceptual output (i.e., an $\mathrm{MR_{np}}$) from a perceptual *P*. If either of these is plausible, then (P5’) is false.

The conceptualist says that both *R* and *Q* are housed within the perceptual system and that, therefore, *P* is perceptual, too, and perceptual states can express conceptual content corresponding to the abstractions encoded as *P* by *R* and *Q*. That is, *P* just is an $\mathrm{MR_p}$ expressing the content, regarding some perceived object, *a*, that it is *F*, where *F*, denotes the perceivable property defined by *P* (e.g., *being blue*). The anti-conceptualist can make claim (a): that *R*, *Q*, or both are contained outside the perceptual system and, hence, that *P* is not (entirely, at least) perceptual, or he can make claim (b): even if *R* and *Q* (and, therefore, *P*) are perceptual, *F*-ness is expressed by an MR_{np} that is caused by an a falling along *P*—i.e., there's something we might call an $\text{MR}_{\text{P-indicator}}$. If claim (a) is true, then (P5') is false with respect to in-acquisition causation. If claim (b) is true, then (P5') is false with respect to post-acquisition causation. I now hope to show that neither (a) nor (b) is plausible.

The preceding section pursued the unlikely approach of defining a one-to-one causation with the intent of doing some work for us down the line. What it has done is restrict the range of possible answers to the present question. Either *R* is comprised of NS_{p}s or it is comprised of NS_{np}s that are indicators of the features encoded in corresponding NS_{p}s—call these $\text{NS}_{\text{q-indicator}}$s. Therefore, *Q* must be a means of ordering the quale-types encoded in NS_{p}s or in $\text{NS}_{\text{q-indicator}}$s. Now a few of the ways of formulating (a) can be quickly rejected. The position in which *R* is a set of $\text{NS}_{\text{q-indicator}}$s while *Q* is housed within the perceptual systems—i.e., *Q* orders NS_{p}s—commits an obvious category error. If *Q* orders NS_{p}s, then $\text{NS}_{\text{q-indicator}}$s are the wrong sort of thing to be ordered by *Q*. We would need a non-perceptual *Q* to order $\text{NS}_{\text{q-indicator}}$s. The same problem holds for the position in which NS_{p}s are ordered by a non-perceptual *Q*, so we can reject that, too.

The claim that abstraction operates on $\text{NS}_{\text{q-indicator}}$s outside of the perceptual system (i.e., both *R* and *Q* are non-perceptual) suffers from the following problem. A rule for ordering instances of a given quale-type according to the temporal order of their occurrence for a given individual—i.e., the first color you saw is rated more similar to the third color you saw than it is to the seventeenth color you saw—will be just as good as ordering them according to some similarity relation between the NS_{p}s underlying the phenomenal experience of the color continuum. So, we could have a color concept that was wildly discontinuous with our actual perception of colors and it would seem just as good a color concept as BLUE. Of course, we don't have such color concepts and we don't think that such a color concept would be a perfectly good color concept. The obvious response is to claim we have an innate set of $\text{NS}_{\text{q-indicator}}$ *Q*s, analogous to the *Q*s for the perceptual system, ensuring that our similarity ratings match the continua we find for variation within a given quale-type in our perceptual systems. But, if we are worried about parsimony, this is a poor response. It requires that we completely replicate our perceptual systems elsewhere in the brain. We no longer have to suppose that *P* is housed in the perceptual system, but we do have to suppose that all the ingredients of *P* (i.e., *R* and *Q*) have correlates in the perceptual systems. So why not just accept a perceptual *P*?

A variant of the $\text{NS}_{\text{q-indicator}}$ proposal might hold that, rather than possessing a conceptual content system that runs in parallel with the perceptual systems, we have an intervening NS_{np} between the inputs of early stage perception and its phenomenal outputs. This could be responsible for imposing an arbitrary *Q* of the sort proposed, while outputting perceptual states in which the quale-type ordered by that *Q* are phenomenally similar in the expected ways. This suggestion is fraught with problems, not the least of which is that it short circuits the causality from world to perceptual content-bearing mental state, such that we cannot ever be certain that the appearance of some thing actually corresponds to a feature of the physical object itself (e.g., surface reflectance properties) rather than some arbitrary rule programmed into our brains. And that imposes an epistemic gap between us and the truth-conditions allegedly secured by the rule of first contact because we don't know if we are having *genuine* contact with the physical world in any given perceptual episode. This, alone, is cause enough to reject this proposal.

Proponents of the MRP-indicator view (i.e., claim (b)), the final available anti-conceptualist position, must endorse the entire mechanism required by the conceptualist, including the existence of a partition, *P*, of *Q* in the perceptual system such that *a* can be isolated along that partition for positive identifications of *F*-ness, while denying that the partition in the perceptual system does anything more than pass information along to a non-perceptual system that is responsible for expressing *F*-ness.[21] The analogy that comes to mind is a bell on a shop door. The bell rings whenever someone enters the shop, alerting the shopkeeper to a customer's presence, in case she was, say, in the storeroom. The $\text{MR}_{\text{P-indicator}}$ is analogous to the bell, but in this case the shopkeeper never has her eye off the door, so the bell (i.e., the $\text{MR}_{\text{P-indicator}}$) is redundant. It is certainly possible that there is such redundancy built into our brains, but in the absence of good reasons for believing this, we should prefer the more parsimonious explanation—conceptualism.

Furthermore, I'm not sure that $\text{MR}_{\text{P-indicator}}$s actually express the conceptual content that is required of them. To avoid falling into the problem besetting the completely nativist picture of abstraction, these $\text{MR}_{\text{P-indicator}}$s can only express that the *P* to which I have been assigned is firing because this view doesn't allow for acquisition in tandem with the perceptual systems. To allow this tandem acquisition is just to fall back on the $\text{NS}_{\text{q-indicator}}$ view. If we are asked to justify why it is that we believe, e.g., that the sweater is blue, we will have to fall back on the content of *P*, not the content of its associated $\text{MR}_{\text{P-indicator}}$. But then this view isn't really anti-conceptualist at all.

If I am correct about all of this, we get the following result. There are three possible views about the causal connection of the intentional objects of empirical thoughts to the neural states underlying the expression of the content of those thoughts. In order of plausibility they are (1) the MR_{p} view (i.e., (P5') is true), (2) the $\text{MR}_{\text{P-indicator}}$ view (i.e., the shopkeeper's bells, which might just be a roundabout way of endorsing (P5')), and (3) the $\text{NS}_{\text{q-indicator}}$ view (i.e., the completely non-perceptual abstraction view, on which (P5') is false). The only one of the three that does not commit itself to redundancy is (1). The only one that clearly makes (P5') false is (3)—i.e., the least plausible (because the least parsimonious) of the three possibilities. Then, barring the advent of empirical evidence to the contrary, we should accept (P5') as a strong inference to the best explanation.

At this point, it should become clear why (P5') is phrased as

"The neural states upon which perceptual content-bearing mental states supervene *do not* cause further non-perceptual neural states upon which conceptual content-expressing states supervene" as opposed to "The neural states upon which perceptual content-bearing mental states supervene cannot cause further non-perceptual neural states upon which conceptual content-expressing states supervene". In fact, they *might*. It just so happens that we have no good reason to believe that they do. The inference that (P5') is correct becomes even stronger when we consider recent research connecting brain lesions in perceptual regions of the brain with conceptual deficits (Simmons and Barsalou 2003).

VI. Conclusion

If we accept (P5'), then—on the assumption that (P1)–(P4) are true—(C') follows. That is, conceptualism is true. But even where the unmodified version of the argument for conceptualism has been accepted, conceptualism has not been widely endorsed. This is because conceptualism has traditionally been considered only in its strongest form, content conceptualism, and the proponents of content conceptualism have not always been clear on the distinction between the perceptual content of a perceptual state and conceptual content expressed by that state. These two form of content have often been discussed as though they were one, and this leads to obvious problems—e.g., it seems true that the visual content of seeing a blue sweater does not vary with the possession or lack thereof of the concepts SWEATER and BLUE. Then the visual content must be distinct from the conceptual content.[22]

I don't think this is a correct interpretation of the content conceptualist thesis, but it won't matter here because the version of conceptualism that my reformulated argument defends is state conceptualism. On state conceptualism the objection clearly fails—it's not that the visual content of the blue sweater is identical to the content of that the sweater is blue. It is, rather, that the visual content of the blue sweater expresses the content that the sweater is blue if one happens to possess the concepts SWEATER and

BLUE. We can make this more precise by saying that the perceptual content yields a perceptual state that expresses this content if one has acquired (or otherwise possesses) abstractions that function as MRs expressing the observation concepts SWEATER and BLUE, and—I have argued—such MRs, if possessed, are housed in the perceptual regions of the brain.[23] Even innate MRs for observation concepts must be stored in the perceptual system. Otherwise—due to (P5')—no NSp could causally instantiate them.

This is a wider claim than the one made by the traditional conceptualist (content or state), which is just about conceptual contents concerning present states of affairs. It entails that, when you tell me, for instance, that you have a blue sweater in the top drawer of your dresser, if I am to understand you, it is (at least partially) in virtue of abstractions (e.g., SWEATER, BLUE, IN, DRESSER) that are stored in my perceptual system and that have been correlated with further perceptual items, e.g., "sweater", "blue", "in", and "dresser". In hearing the words of your sentence, the mental representations (i.e., the perceptual abstractions) associated with those words are activated—though not necessarily in a mental image-producing way.

Finally, I have said that among the possible views, we should prefer conceptualism for its parsimony. However, I have only been concerned with the expression of empirical conceptual contents. There may well be cause to endorse views (2) or (3) if it can be shown that perceptual MRs are insufficient for the expression of abstract conceptual contents such as that patience is a virtue or that seven is prime.[24] Then, one could argue that a unified account of intentionality is better than a hybrid one and, since there are accounts (i.e., the $\text{MR}_{\text{P-indicator}}$ and $\text{NS}_{\text{q-indicator}}$ views) of how it could be the case that MR_{np}s express conceptual content about empirical states of affairs, it follows that conceptualism is false. I don't think that any such fact has been established, and whether or not it can be established seems to me to hang upon our ability to draw a principled distinction between empirical and non-empirical conceptual content. The possibility of drawing such a principled distinction, then, seems to be the next point to consider in the conceptualism debate.

Notes

1. I will use "perceptual state" to refer to perceptual or (inclusive) proprioceptive (i.e., motor) states. What makes a perceptual state a perceptual state, in this sense, will be discussed in a bit more detail below.
2. This view, introduced by Frege, has been standard in analytic philosophy ever since. For substantiation of this claim, see Laurence and Margolis (1999, pg. 4).
3. See, for instance, Peacocke (1995). This tendency to associate concepts with abstract objects is so strong that, even where they attempt to argue against the view, Laurence and Margolis wind up endorsing a non-standard version of it, in which concepts are mental representation types (1999, pp. 5-8). Types, as I understand them, are a form of abstract object.
4. In some recent writing the term "intentional content" has been used as I am using "intentional object". Usually, this just seems to be a matter of semantics. For instance, Jesse Prinz uses "intentional content" as I use "intentional object", but he then introduces the term "cognitive content" to denote what I call "intentional content" (2004, pp. 3-7). There is no substantive disagreement here. I prefer my distribution of meanings to these terms, which I share with, among others, Devitt (1990), Husserl (2001 [1900/1901]), Putnam (1973, 1975), and Searle (1983). We should also note that, on a singular proposition picture, the intentional object is a constituent of the intentional/conceptual content, but it never just is that content because what makes the content conceptual is that it has concepts as constituents and individual particulars of the sort included in singular propositions aren't concepts (see section 1).
5. See Grice (1961), Tye (1982), and Noë (2003) for a sampling of views in support of the causal theory of perception.
6. (P4) is a, usually unarticulated, background assumption of the argument. I choose to articulate it for the sake of clarity. I take it as uncontroversial simply because to do otherwise is to endorse some sort of extra-sensory perception, and I reject—and take it that I am safely within the majority in doing so—the existence of extra-sensory perception.
7. For a detailed version of this argument, see (McDowell 1994, pp. 3-10; Searle 1983, pp. 45-57).
8. The view that intentionality is a species of causation is widely accepted in analytic philosophy and, beyond its specific formulation as the causal theory of content, is particularly associated with the project of "naturalizing the mind" (e.g., Dretske 1981, 1983, 1997; Harman 1990; Tye 1995).
9. Responses to these, and other, criticisms of the causal theory of content have been offered. See, for instance (Rupert 199, 2001, 2008; Ryder 2004;

Usher 2001). Furthermore, as Fodor and Dretske are anti-conceptualists, I can hardly be accused of biasing thee case by endorsing their account of the intentionality of observation concepts.

10. One reason that it's obviously wrong will be discussed in section 6, but another is it cannot be true on any non-psychologistic view of conceptual content, and psychologism is false (Frege 1956 [1818/1919]).

11. Certainly this is true of BLUE, (that is one take-home lesson of Jackson's (1982) Mary, the vision scientist thought experiment). One cannot have a color concept capable of identifying things as that color unless one has had actual experiences with that color leading to the acquisition of that color concept. SWEATER, on the other hand, might not be acquired this way. For instance, someone might be able to acquire SWEATER by a description—e.g., something like a shirt that is knit from yarn and is, usually, worn over a shirt. If one has concepts SHIRT, KNIT, YARN, etc. then one could acquire SWEATER from such a description. At bottom, though, there will be at least some components of one's SWEATER concept—and I think a very large percentage of those components—that are acquired through direct perception.

12. Even the Gettier (1963) problem for the traditional analysis of knowledge as justified true belief seems to be a problem concerning justification. If one doubts this, we can simply ask "Can someone know that 2 + 2 = 5?" I suspect that the overwhelming answer will be "no".

13. One might worry that, if there are an infinite number of qualia, we need an infinite number of NS_ps to accommodate them and, so, this count of perception is impossible as well. In fact, qualia are composed of complex NS_p that code for coarsely coded feature and feature relation descriptors (Treisman 1998, p. 27). These subsidiary NS_ps are not phenomenally available in perception. Furthermore, there are a finite number of them.

14. See Laurence and Margolis (2012), Barsalou (2003), and Simmons and Barsalou (2003) for recent accounts of how an acquisitional abstraction mechanism could work. Though the first (Laurence and Margolis) comes from a mildly nativist viewpoint and the other (both proposals from Barsalou) from an empiricist one, the general principles are the same: define a means of assessing similarity in some feature domain by giving a principled means of ordering the total range of possible variation in that feature and then allowing a means of partitioning that domain into discrete units. The account I give of abstraction, below, is intended to be compatible with both views.

15. This point is not original. See, for instance, (Laurence and Margolis 2012, p. 10).

16. This won't result in a problem of too many innate ordering principles because, though each ordering principles might, itself, cover an infinite range of possible qualia, there are a finite number of quale-types, and each ordering

principle can be defined as a rule for combinations of the (finite number of) coarsely-coded features and feature relation descriptors which combine to form qualia of the given quale-type (see note 13, above). The accounts of abstraction from Simmons and Barsalou (2003) and Laurence and Margolis (2012) both invoke something like my *Q*. For Simmons and Barsalou, there is a single principle, the similarity-in-topography principle, by which all like perceptual states can be organized. This proposal maintains low-level domain specific orderings analogous to what I am proposing, as well as explaining similarity ratings of more complicated perceptual phenomena of the sort that would need to be ordered by one of my multi-dimensional *Q*s (see note 17, below). I find their proposal plausible and parsimonious but not conclusively proved. Laurence and Margolis offer a set of innate, low-level abstractions to form the basis of the abstraction process. That is, we come with innate partitionings of *Q* for basic perceptual features such as color or curvature (i.e., just those sorts of perceivable features that I am saying require, minimally, an ordering principle, *Q*) and build further abstractions (in a manner they do not fully specify) out of those (2012, pp. 20-21).

17. These multi-dimensional *Q*s will need to include dimensions ordering primitive feature *relations* as well as primitive features (qualia). There is good empirical support for the existence of such feature relation codings in early vision (see, in particular Kosslyn 1975, 1988, 1995, and Kosslyn and Pomerantz 1977).

18. Just how central a role is to be played by acquisition depends on how far one is willing to take nativism, given the restriction against total nativism. Clearly, though, a large portion of our abstractions are *not* innate (i.e., they are acquired).

19. As stated above, I do not think that it is necessary to have a term denoting the concept in order to possess that concept. All I am committed to is that there is some non-varying perceptual, or some already formed concept, instances of which are positively correlated with the percepts of the kind that are taken as comprising an R.

20. If there are innate MRs—i.e., *P*s—expressing observation concepts, then they are in place to satisfy the rule of first contact. It is only with innate concepts that the rule of first contact can operate independently of the rule of fundamental observation concept acquisition.

21. This is, essentially, the modularity of mind thesis—i.e., the thesis that certain brain functions, including perception, are modularly encapsulated and isolated from any input from cognitive processes. For arguments against modularity from empirical data, see (Churchland 1988; Prinz 2004, pp. 113-114). I will argue against the view on other grounds.

22. See Speaks (2005) for a lucid defense of both forms of conceptualism against the traditional anti-conceptualist arguments, including this one.

23. Perceptual content yields a perceptual state in just the way I sketched out in section 5- i.e., by attending to the perceptual content in virtue of some *Q* and making a best-match (or no-match) with an existing partition, *P*, of that *Q*.

24. That seven is prime, in particular, seems problematic, but there is always the escape hatch of mathematical fictionalism.

Bibliography

Barsalou, Lawrence W. (2003) "Abstraction in perceptual symbol systems," *Philosophical Transactions of the Royal Society* 358, pp. 1177-1187

Churchland, Paul M. (1988) "Perceptual plasticity and theoretical neutrality: a reply to Jerry Fodor," *Philosophy of Science* 55, pp. 167-187

Crowther, T. M. (2006) "Two conceptions of conceptualism and nonconceptualism," *Erkenntnis* 65(2), pp. 245-276

Davidson, Donald. (2001) *Subjective, Intersubjective, Objective* (Oxford: Oxford University Press)

Devitt, Michael. (1990) "Meanings just ain't in the head," in: Boolos (Ed) *Meaning and Method: Essays in Honor of Hilary Putnam*, pp. 79-104 (Cambridge: Cambridge University Press)

Dretske, Fred I. (1981) *Knowledge and the Flow of Information* (Cambridge: MIT Press)

_____. (1983) "Précis of knowledge and the flow of information," *Behavioral and Brain Sciences* 6, pp. 55-63

_____. (2003 [1980]) "Sensation and perception," in: Gunther (Ed) *Essays on Nonconceptual Content*, pp. 25-106 (Cambridge: MIT Press)

_____. (1988) *Explaining Behavior: Reasons in a World of Causes* (Cambridge: MIT Press)

_____. (1997) *Naturalizing the Mind* (Cambridge: MIT Press)

Gettier, Edmund. (1963) "Is justified true belief knowledge?," *Analysis* 23, pp. 121–123

Grice, H. P. (1961) "The causal theory of perception," *Proceedings of the Aristotelian Society*, Suppl. 35, pp. 121-152

Fodor, Jerry. (2003 [1987]) "Meaning and the world order," in O'Connor and Robb (Eds) *Philosophy of Mind: Contemporary Readings*, pp. 271-303 (New York: Routledge)

Harman, Gilbert. (1990) "The intrinsic quality of experience," *Philosophical Perspectives* 4, pp. 31-52

Husserl, Edmund. (2001 [1900/1901]) *The Shorter Logical Investigations* (New York: Routledge)

Jackson, Frank. (1982) “Epiphenomenal qualia,” *Philosophical Quarterly* 32, pp. 127–136

Kosslyn, Stephen M. (1975) “Information representation in visual images,” *Cognitive Psychology* 7, pp. 341-370

_____. (1988) “Aspects of a cognitive neuroscience of mental imagery,” *Science* 240, pp. 1621-1626

_____. (1995) “Mental imagery,” in: Kosslyn and Osherson (Eds), *Visual Cognition*, pp. 267-296 (Cambridge: MIT Press)

Kosslyn, Stephen M. and James R. Pomerantz (1977) “Imagery, propositions, and the form of internal representations,” *Cognitive Psychology* 9, pp. 52-76

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

Laurence, Stephen and Eric Margolis. (1999) “Concepts and cognitive science” in: Laurence and Margolis (Eds), *Concepts: Core Readings*, pp. 3-82 (Cambridge: MIT Press)

_____. (2012) “Abstraction and the origin of general ideas,” *Philosophers' Imprint* 12(19), pp. 1-22

McDowell, John. (1994) *Mind and World* (Cambridge: Harvard University Press)

Noë, Alva. (2003) “Causation and perception: the puzzle unraveled” *Analysis* 63(2), pp. 93-100

Peacocke, Christopher. (1995) *A Study of Concepts* (Cambridge: MIT Press)

Prinz, Jesse. (2004) *Furnishing the Mind: Concepts and Their Perceptual Basis* (Cambridge: MIT Press)

Putnam, Hilary. (1973) “Meaning and reference,” *Journal of Philosophy* 70, pp. 699-711

_____. (1975) “The meaning of ‘meaning,’” in: K. Gunderson (Ed), *Minnesota Studies in the Philosophy of Science, Vol. VII: Language, Mind and Knowledge* (Minneapolis: University of Minnesota Press.)

Quine, Willard van Orman. (1960) *Word and Object* (Cambridge: MIT Press)

Rupert, Robert D. (1999) “The best test theory of extension: first principle(s),” *Mind and Language* 14, pp. 321–355

_____. (2001) “Coining terms in the language of thought: innateness, emergence, and the lot of Cummins's argument against the causal theory of mental content,” *Journal of Philosophy* 98, pp. 499–530

_____. (2008) “Causal theories of mental content,” *Philosophy Compass* 3, pp. 353–380.

Ryder, Dan. (2004) "SINBAD neurosemantics: a theory of mental representation," *Mind and Language* 19, pp. 211–240

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

Simmons, W. Kyle and Lawrence W. Barsalou. (2003) "The similarity-in-topography principle: reconciling theories of conceptual deficits," *Cognitive Neuropsychology* 20, pp. 451-486

Speaks, Jeff. (2005) "Is there a problem about nonconceptual content?," *The Philosophical Review* 114(3), pp. 359-398

Stampe, Dennis. (1977) "Toward a causal theory of linguistic representation," *Midwest Studies in Philosophy* 2, pp. 42-63

Treisman, Anne. (1998) "The perception of features and objects," in: Wright (Ed), *Visual Attention*, pp. 26-54 (Oxford: Oxford University Press)

Tye, Michael. (1982) "A causal analysis of seeing," *Philosophy and Phenomenological Research* 42(3), pp. 311-325

_____. (1995) *Ten Problems of Consciousness* (Cambridge: MIT Press)

Usher, Marius. (2001) "A statistical referential theory of content: using information theory to account for misrepresentation," *Mind and Language* 16, pp. 311–334

Drugs and Consciousness: Supervenience, Emergent Dualism and Magic Mushrooms

Victoria Canada Ritenour

Introduction

There are three major chemicals in our brains that are thought to control mood. If a person is diagnosed with depression, a medical doctor will prescribe drugs known to affect mood, emotion, and behavior, from a class of pharmaceuticals called psychotropics (a combination of the Greek words *psycho* and *tropic* meaning "mind turning") commonly known as antidepressants. Current medical theory holds that when a person's brain lacks the ability to produce or absorb one of more of three major neurochemical transmitters (serotonin, dopamine and norepinephrine) their physician will prescribe drugs to alter the mechanical structure that affects those chemicals. These are drugs we have all heard of, and you've probably known people who have taken them, that is, drugs such as Prozac, Zoloft, and numerous others, that are administered to fill the gaps in neurochemical production and aid the individual in their daily life. But psychotropic drugs are only effective to an extent, have a plethora of side effects, need daily administration (becoming costly for individuals needing extended treatment), and can become addictive. The World Health Organization reports that "depression is the leading cause of disability worldwide in terms of total years lost due to disability" and, in a 2012 report, found that "on average about 1 in 20 people reported having an episode of depression in the previous year," with "almost 1 million lives lost yearly due to suicide" translating to "3000 suicide deaths every day" (WHO 2012). Depression is a global issue with a high yearly death toll.

Researchers have been looking for better alternatives for decades, and in the 1950's they began testing drugs created in the laboratory like LSD, as well as naturally occurring psychedelics, to address depression. Currently, MDMA (ecstasy) and psilocybin (magic mushrooms) are being tested again in the neurological community as a means to treat and fight depression. Recently, a Johns Hopkins study on psilocybin has made headlines in tracking the drug's ability to alter consciousness for up to fourteen months after the last dosage, resulting in changes to overall outlook, as well as daily mood, with positive results (Hagerty 2009; Griffiths et. al. 2011, p. 653). Philosophy has a few theories that can possibly address the mind-body problem in regards to the biological structure of the brain, and the mental function of the mind. But with the challenges raised by this new study on psychedelic-induced experiences (as opposed to psychotropic chemical additions) there may need to be a new philosophical theory concerning the mind-body distinction, or, at the very least, amended or updated versions of existing theories.

Emergent Property Dualism and Supervenience Theory

Emergent property dualism is the theory that the underlying chemical structures in our brain create emergent properties, such as consciousness and mood. But these emergent properties are not bound by, nor respond to, the same rules and regulations as the physical structures from which they come. For example, I can alter the chemical structure in my brain by taking a drug to up my production of serotonin, and would therefore expect a change in my mood. So long as the drug is in my system and present in the neurochemical structure of my brain, I can expect the subsequent emotional and mental changes of having more serotonin in my brain. When a doctor administers a psychotropic drug he is addressing the chemical need for a change in the brain while simultaneously addressing the issue of the patient's mental experiences. If we are to take the mind and brain as separate (while still holding to the philosophical theory of emergent dualism, and the medical precedence of treating mental disorders via chemical means) then we need to accept that chemical change begets mental change, even though that mental change is not reducible to the physical mechanism that caused it.

Though separate, the mind and body are linked by causal properties in the physical that alter the phenomenological, but somehow do not explain nor inform the phenomenological. A blueprint is a detailed description of a home, but simply looking at it cannot translate into the experience of being inside a room; you can see the window placement on the page, but that does nothing to induce the feeling of the light or breeze those windows create in the final product. Nor does the actual experience of being in a room reduce down to the information on its blueprint, regardless of how detailed. A way to think about this issue is to consider brain scans. We can see certain points light up on the screen that are associated with mental phenomena, but just because we can map the chemicals in use, or see the physical changes on the screen, this basic knowledge of the neurology of the patient's experience does not allow a researcher to understand happiness, love or a positive mental outlook. Blueprints and rooms are separate substances, which is analogous to the view that brains and minds are separate substances according to emergent property dualism, but the analogy falls short when we think about the sustained and continual connection the mind and brain share, that blueprints and rooms do not.

Something that might inch us closer to understanding where the mind, and mental experiences, come from, is supervenience theory. In supervenience theory, the mental structure of the mind is ontologically reliant on the physical structure of the brain. Its essence is in the brain, an apt analogy being a magnet (the brain) creating a magnetic field (consciousness). For example, we can conceptualize a human brain with consciousness emerging from it (or supervening on it) but we cannot imagine a conscious structure without such physical basis. By analogy, there is no magnetic field without a magnet creating and sustaining that field. In supervenience theory, there is no substance dualism (which posits a mind, or soul, independent of the brain) but rather the mind is physically tethered to the brain via neurochemical activity. Strong supervenience requires that the molecular structure of the brain would need to be altered in order to change the mental properties that supervene on the brain. Often supervenience is described as A-properties (the mind) supervening upon B-properties (the brain), such that any change in the B set of properties changes the A set; that is, there can be no A property change without a B property change (McLaughlin and Bennett 2011). Just as the aesthetic properties of a painting supervene on the physical structure of the paint and canvas, the mind supervenes on the brain, and by analogy, a change in the physical structure of the painting can lead to a change in its aesthetic properties. However, both emergent property dualism and supervenience theory are imperfect in their ability to philosophically explain the repercussions of current neurological research into consciousness, though they both have their place in aiding our understanding of the mind and the scientific possibilities there within.

Emergent Features, Supervenient Structures, and Directional Causality

In his chapter entitled “Emergent Dualism,” William Hasker explains Searle's description of “causally emergent system features,” that is, features of a whole that are not shared by any of its parts (Hasker 2001). The example of a stone is given where *a*, *b*, and *c* are components of a stone, *S*. *S* has a very specific weight, and various other properties, that are not shared with its smaller components. Now, imagine that *S* is a sedimentary rock, and component *b* is its iron content. Following Hasker's description of Searle's “causally emergent system features,” no matter how much we know and understand about iron, that knowledge will never inform us of the depths and lengths of *S* (though by studying *b* we may further our basic understanding of *S*) (Hasker 2001, p. 173). Similarly, though we know and understand many of the brain's physical functions and chemical reactions, there remains a void of understanding as to how such features of the brain translate into mental phenomena, as well as how they make up a mind and where that mind is kept. So, we track the firing of neurons trying to understand the mind and mental phenomena, and, to a point, we can see how advanced neurological phenomena correlate to mental phenomena, but we still cannot understand how or why.

Tack on to this theory the idea that, much like a magnetic field is created by an energized magnetized metallic substance, the general function and operation of the supervening mental structure is not bound by the same rules that the underlying physical structure is bound by. "The existence of consciousness can be explained by the causal interactions between elements of the brain at the micro level, but consciousness itself cannot be deduced or calculated from the sheer physical structure of the neurons without some additional account of the causal relations between them" (Hasker 2001, p. 171). So, our stone *S*, has features that are informed by, but aren't shared with its components *a*, *b*, or *c*. Hasker calls these features, emergent$_1$ properties. That is, the stone's emergent1 properties are bound by a direct correlation between its underlying physical components. What is problematic for Searle, and what he argues against, are emergent$_2$ properties, which are properties that seem unrelated to the physical. That is, properties that, once created by an underlying physical structure, create an autonomous property, no longer bound by the physical structure from whence it came. In my research on this topic, I could not find anyone without some sort of religious affiliation, and especially no one whose writings are based on purely scientific pursuits, that wants to admit that consciousness is possibly emergent$_2$. Hasker quotes Searle as saying, "In fact, I cannot think of anything that is emergent$_2$, and it seems unlikely that we will be able to find any features that are emergent$_2$, because the existence of any such features would seem to violate even the weakest principle of the transitivity of causation" (Hasker 2001, p. 172). But, removed from the heady theoretical pursuit of dualism and consciousness, in the realm of scientific drug studies, there seems to be an emergent2 mental entity created in the psilocybin drug trial done at Johns Hopkins, that is, a consciousness that cannot be filed under emergent$_1$, as the same drug can be applied to the same neurochemical structure, but at different times create distinct conscious experiences.

In 2011, Johns Hopkins neurological researchers conducted a study on the brain and conscious experience, applying doses of psilocybin in carefully tailored environments in order to induce mental and spiritual states that affected the general structure of the study participants' brains (Griffiths et. al. 2011, p. 651). What is at issue in the Johns Hopkins trial are the more restrictive beliefs of philosophers, like John Searle, who hold to a specific version of emergent dualism that is one directional. For Searle, chemical or structural change can only affect the emergent property in a singular direction, as opposed to the view of more open-minded philosophers. William Hasker, for example, entertains emergent$_2$ dualism where the line of communication is open in both directions, so that the brain affects consciousness and consciousness can affect the brain (Hasker 2001, pp. 172-173). Both supervenience theory and emergent property dualism seem to address the same concept: a basic physical structure that begets a metaphysical structure. Traditionally, medicine and philosophy have treated the mind as though this were so (for example, think of using psychotropic drugs as a means to address mental issues, as mentioned in the introduction). In most versions of emergent property dualism, theories of the emergent mind and supervenient consciousness are modeled on a sort of multilevel structure, existing above the baseline neurological functions of our brains. But perhaps the causal interactions between our minds and brains are not so hard lined and one-directional, that is, perhaps Hasker has a point. In causal emergence, the mind is an emergent property of the brain that results from higher-level neurological functions. That is, the mind exists over and above the physical brain as a separate result from the system that creates it, bound by a different set of rules and laws entirely. The relationship between brain and mind is studied tirelessly, but more and more research is showing how non-physical, mental phenomena can affect our physiology. In the Johns Hopkins trial, as a person was dosed with psilocybin, their neurochemical composition was effected, but, according to the study, this is not what changed their consciousness. Rather, when under the effect of psilocybin, pared with a comfortable and safe environment, participants had drug trips that were spiritual in nature, and it was these spiritual experiences that altered their consciousness thereafter, not the addition of psilocybin to their brain chemistry. The Stanford Encyclopedia of Philosophy states that

> *Emergent property dualism* treats conscious properties as arising from complex organizations of physical constituents but as doing so in a radical way such that the emergent result is something over and above its physical causes and is not *a priori* predictable from nor explicable in terms of their strictly physical natures (Van Gulick 2004).

Philosopher Jaegwon Kim explains this further,

> At the core of these ideas was the thought that as systems acquire increasingly higher degrees of organizational complexity they begin to exhibit novel properties that in some sense transcend the properties of their own constituent parts, and behave in ways that cannot be predicted on the basis of the laws governing simpler systems (1999, p. 3).

That is, the properties emerge from (as opposed to being caused by) the underlying physical structure, although they are not necessarily predictable from that underlying structure. Rather than assume that there is an entirely separate substance of "mind," we discuss the mind as a set of properties rooted in, but not directly one-to-one correspondent with, the physical brain.[1] Being careful not to limit this non-physical structure to an epiphenomenal dual substance, we look to supervenience, and emergent property dualism, to see just how the metaphysical entity of mind is connected to, but distinct from, the physical realm.

Repeated introduction of psilocybin to serotonin uptake inhibitors in the brain follows a basic, mechanical framework on the neurochemical level. The general mechanism for absorbing and reacting to the drug in the human brain is predictable and observable, since this is the structure that SSRI's, a popular class of antidepressants, are formulated for. There is the naturally occurring structure to produce and intake serotonin, and adding SSRI drugs (or psychotropics) effect this system with accurate regularity. But psychedelics like psilocybin that (though they produce serotonin and effect SSRI's in the brain) only affect mood in regards to the metaphysical experience of the drug trip (Hagerty 2009). Regardless of the chemical and physical regularity of the brain's reaction to psilocybin, the variety of mental phenomena produced cannot be explained by the same set of rules made for, and applied to, this chemical system. Chemical reactions identical to past usage of the drug may create new and different mental phenomena, and it is this phenomena, the actual "drug trip," that alters the consciousness of the individual, and in turn affects their mind for months thereafter. This effect is significantly longer than the drug was ever present in the brain, which creates an issue in regards to the mental phenomena only being a "property".

Back to our magnetic field, if we could energize a magnet to create a predictable field to, say, start a motor, and that field reacted and performed uniformly every time we flipped the power switch, the magnetic field created would be emergent$_1$. Kim places this inductive connection, or "inductive predictability" as he calls it, in opposition to the theoretical predictability that isn't explained by emergent properties. The transitivity of causation would entail that the magnet, by causing the magnetic field, would be the original cause of anything that field produced. As the brain causes consciousness, then whatever consciousness causes must be reducible to the original brain function. "Even emergent properties are inductively predictable: Having observed that an emergent property, *E*, emerged whenever any system instantiated a microstructural property *M*, we may predict that this particular system will instantiate *E*, at *t*, given our knowledge or belief that it will instantiate *M*, at *t*" (Kim 1999, p. 8). With our drug study we can logically induce that the addition of psilocybin to the brain in a positive and comfortable setting will create the needed mental phenomena, as this is the premise the study was based upon. But, we cannot theoretically predict the same emergent mental phenomena in every instance of psilocybin usage, or corresponding reaction to conscious experience. The addition of the psilocybin chemical to the brain in a comfortable environment is not sufficient to elicit the intense mental phenomena that will create the emotional change seen in some subjects. Kim, in regards to this theoretical predictability states, "we may know all that can be known about *M*—in particular, laws that govern the entities, properties and relations constitutive of *M*—but this knowledge does not suffice to yield a prediction of *E*" (1999, p. 8). Kim speculates that this lack of theoretical predictability inheres in our lack of understanding *E*, and *E*'s inception and inherence in *M*. This is not controversial, as we agree that we have no working theoretical predictability, only inductive, or logical predictability for metaphysical mental events (Kim 1999, pp. 8-9). In regards to psilocybin, past doses in study participants were not always necessary, nor sufficient, to alter the supervening mental structure; conversely, the mental phenomena described in the study altered the underlying neurochemical structure significantly, which means that our A-properties somehow altered our B-properties. This is a phenomenon that Searle believes violates the transitivity of causation, and thus impossible, so it is not allowed for supervenience theories (Hasker 2001, p. 172).

CONSCIOUSLY ALTERING CONSCIOUSNESS

Let's take an example as if it were happening to you. Say, for example, you are a rather bland, unremarkable person. Many people have had religious or spiritual experiences in their past which have had little effect on their daily consciousness, so even if you have a history of attending a church or temple, let us assume you do not consider yourself a very spiritual or particularly altruistic person (and that those around you for any amount of time would agree). Imagine your, and your friend’s surprise, when upon taking a dose of magic mushrooms, those traits that once described you seem drastically changed. Currently, and for months to follow, you are more in-tune with your spiritual self. You've taken up daily prayer and meditation, you donate your time, money and skills to local non-profit organizations, and you are kinder and more jovial in your daily interactions with those of whom you live and work.

Let us say you have dabbled in recreational drug usage before, and the mushrooms themselves were no different chemically from the dozen or so others you have taken in the past. Surely the change in your daily conscious experience was not in direct correspondence to the chemicals you ingested, so supervenience theory does not quite explain how your mind could be altered so drastically with the introduction of familiar chemical compounds. At any rate, the chemical changes necessary for the conscious changes did not happen with the previous doses, which means there is a missing link in the agent causation with regard to the application of the drug to the neurochemical mechanism, as a means to effect (and affect) your brain. The psilocybin, the active ingredient in the mushrooms, is the same chemical that has affected your neurological structure in the past. So, we might assume that the changes in you are not a chemical effect from when you were high. In fact, the last dose wore off weeks ago, so the chemical structure in your brain is no longer responding to the psilocybin. How then is the emergent, conscious structure of your mind so changed?

The answer to the real world psilocybin drug study seems to have little to do with the drug induced high that the chemical gives many of its users, but rather is explained by the intense spiritual experience that certain users have when taking the drug. It seems, as demonstrated by the Johns Hopkins medical research, that though we can track and quantify the chemical changes in a person's brain while they are on a certain drug, it is the non-quantifiable, qualitative, experiences that provide the study participant with the significant shift in consciousness—a shift significant enough to change and inform behavior for up to 14 months. It is also troublesome for emergent property dualism that the spiritual experiences and subsequent personality changes seen in this study are not the direct result of altered brain chemistry. This study also raises questions concerning the current medical treatment of mental disorders through continued chemical means like psychotropic drugs. Some study subjects were dosed with the same drug, and the same neurochemical reactions were tracked on brain scans, yet no spiritual conscious experience was produced, nor was there any change in their behavior. "Psilocybin produced acute perceptual and subjective effects including… extreme anxiety/fear (39% of volunteers) and/or mystical-type experience (72% of volunteers)" (Griffiths 2011, abstract). In the past, researchers were also able to induce, via environment, intense spiritual experiences with varying behavioral changes that had none of the same chemical components as the drug induced ones, yet elicit very similar reactions (Smith 1964). So, it seems that we can alter consciousness with a chemical, not because that chemical directly affects our brains, but because when that chemical is ingested under the right conditions, it will affect our minds. Yet somehow we may not need that chemical, but can attain the desired result through spiritual practice—neither conclusion seems intuitive or scientific.

Conclusion

Sam Harris (a popular thinker, writer, and psychedelic drug advocate) notes that we enter into various consciousness altering environments daily, whether through food, discussion, or other environmental factors, but that the act of taking drugs for such a change is seen as wholly different (2011). Specifically, in our culture there is a negative reaction to altering consciousness via psychedelic compounds. What had once been a major theme in philosophical inquiry and various neurological and behavioral research programs in the mid-20th century has now, due to illicit and recreational drug use, become passé. Now moot in the scientific community, discussion of the legitimate usage of psychedelic drugs (specifically as a means to help people, and to replace a lesser class of psychotropics) is currently relegated to the pseudoscientific ramblings of entertainers like Joe Rogan, and casual conversations at Burning Man. Paired with the empirically minded, markedly un-mystical culture of philosophy in the late 20th century, discussions concerning dualism are rarely concerned with external chemical compounds and their possible effects on the mind. All the while, drug companies and neuroscientists have flooded the market with psychotropic medications (Prozac, Zoloft, Effexor, etc.) to answer the questions being addressed by philosophers in the 1960s and 1970s, namely, "can mood and mentality be affected via chemical means? And if so, what is the best way to go about doing that?" (Smith 1964, p. 518). Instances of altered mood and newfound benevolence were documented in the drug studies of the 1960s. Prior to the discovery of psilocybin as its active chemical compound, magic mushrooms, as well as LSD were found to affect patient's consciousness, perception, and spirituality. Historically, mushrooms have been studied as the catalyst for various intense spiritual experiences, taken by religious leaders and tribesmen alike, to expand knowledge of the metaphysical and, in some instances, interact with one's deity. In the Johns Hopkins study, participants were chosen because of their loose religious affiliations. Currently, new study participants are being sought in regards to their various stages of terminal (or possibly terminal) cancers. The assumption being that if there is such a significant change in one's overall outlook, in one's altruistic behavior, and one's positive mental states post psilocybin doses, then the effect on the quality of life and survival rates of these patients will be significantly changed for the better.

In the psilocybin study, if the test subjects had the same conscious reaction to the drug every time it was applied, during the time it was active in the brain, then we could assume the mental phenomena that emerged from the drug usage was, in essence, like a magnetic field. The drug would be acting as the power supply, and the brain would be the magnet itself. Direct application of this "power" would (and should according to our understanding of chemistry and physics) create a reliable and regular magnetic field with each application; that is, a field that would last the duration of the drug's presence and that should affect the brain for that same amount of time. Unfortunately, there is no such simple and direct correlation between using the drug and having the cognitive change. The correlation is between the spiritual, deeply emotional, experience during the "high," (an experience that is not present in all instances of the drug's usage) and the consciousness of the individual who had the experience (an experience that lasted up to 14 months after a single hour long high). What seems to be exhibited in this study is an emergent2 consciousness. That is, there is mental phenomena that (without direct correlation to the underlying chemical structure in the brain) given the right circumstances impresses upon the user such an intense emotional experience it is, literally, life changing. (In fact, many said that the drug-induced experience was as significant as a child being born.) In regards to our magnet analogy, one special time the power was switched on, created a magnetic field so powerful that it rewired the physical structure of the magnet itself, without any observable or explicable causation. If supervenience is understood as "there cannot be an A difference without a B difference" in regards to A properties supervening on the B structure, we are unable to account for the dynamic quality that consciousness possesses during apparently static neurochemical states. Simply put, the B structure may have created the A supervenient structure, but the change in the B is not always indicative of change in the A; and, seemingly, a change in the A, can actually affect the B structure. In the Johns Hopkins study we saw that the continued application of the psilocybin drug was not the catalyst for the cognitive change (again a change to the B structure did not always beget a change to the A structure), but when the drug induced an intense mental reaction, that mental phenomena changed the study participant's brain chemistry long after the drug (that is the B structural change) had worn off. It is a principle that has moved researchers to now explore psychedelics in lieu of psychotropics as behavioral modifiers, most recently in studies devoted to alcohol and smoking cessation. Specifically, LSD was lab-created for the desired effect of altering human consciousness, and without the administration of psychedelic drugs in all of these consciousness altering and behavior modification trials, there would be no data to speak of. So we know that there is some sort of connection between the physical chemical structure of the test subject's brain and the mental phenomena experienced during the drug trip, which in turn produced the mental phenomena that drastically changed the participant's brain. Now, we just need to create a more descriptive philosophical theory (or further develop Hasker's emergent$_2$ theory) such that it can track all the aspects observed in these studies, which is no simple task.

Notes

[1] Kim explains this in the following quote: "The concept of explanation is invoked in the claim that emergent phenomena or properties, unlike those that are merely 'resultant', are not *explainable*, or *reductively explainable*, on the basis of their 'basal conditions', the lower-level conditions out of which they emerge." He sees this as mereological supervenience, that is, part of the whole system it comes from (Kim 1999, p. 6).

Bibliography

Griffiths, Roland; Johnson, Matthew; Richards, William; Richards, Brian; McCann, Una; Jesse, Robert. (2011) "Psilocybin occasioned mystical-type experiences: immediate and persisting dose-related effects," *Psychopharmacology* 218(4), pp. 649-665

Hagerty, Barbara Bradley. (2009, May 18). The God chemical: brain chemistry and mysticism. *National Public Radio, All Things Considered,* Retrieved from http://www.npr.org/templates/story/story.php?storyId=104240746

Harris, Sam. (2011, July 5) Drugs and the meaning of life. *Sam Harris, The Blog*, Retrieved from http://www.samharris.org/blog/item/drugs-and-the-meaning-of-life

Hasker, William. (2001) "Emergent dualism" in: *The Emergent Self,* pp. 171-203 (Cornell University Press)

Kim, Jaegwon. (1999) "Making sense of emergence," *Philosophical Studies* 95(1-2), pp. 3-36

Mclaughlin, Brian; Bennett, Karen. (2011) "Supervenience," in: *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/entries/supervenience/

O'Connor, Timothy; Wong, Hong Yu. (2012) "Emergent properties," in: *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/entries/properties-emergent/

Smith, Huston. (1964) "Do drugs have religious import?" *The Journal of Philosophy* 61(18), pp. 517-529

Van Gulick, Robert. (2004) "Consciousness," in: *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/entries/consciousness/

World Health Organization. (2012, May 22) "Sixty-fifth World Health Assembly: daily notes on proceedings," Retrieved from http://www.who.int/mediacentre/events/2012/wha65/journal/en/index4.html

Creationism in Fiction

Chuck Dishmon

Introduction

Within my paper I hope to examine creationism in fiction, the view that fictional characters exist by virtue of being created by an author, and to put forth the claims for and against creationism in fiction, so as to gauge the upshots of each. First, I will recreate the argument for creationism in fiction, focusing on Searle's position. Searle argues that there is creationism in fiction; however, in order for the author to bring a character into existence she must first pretend to assert things. Next, I will recreate the argument against creationism in fiction. In doing so, I will focus on Yagisawa's anti-creationist arguments in order to analyze how the anticreationist objections work against the creationist. In turn, I will show how the anticreationist largely hedges their argument against creationism in claiming fictional names do not refer to an entity. And since non-referential sentences are false, authors cannot create fictional individuals. Following this, I will bring in the work of Thomasson to argue that Searle's position is the stronger of the two, showing how I believe the creationist overcomes any relevant objections. Accordingly, I will conclude that fictional individuals exist as a result of being created by their relevant authors, and therefore creationism in fiction is true.

Authors Are Creators

John Searle argues that authors create fictional characters by pretending to refer to real people and places. Accordingly, in talking about fictional individuals, we utilize pretend assertions. However, the foundational tenet of this belief stems from Searle's philosophical position on the systematic relations between written sentences and their meanings. Searle holds that written language is made up of speech acts called "illocutionary acts," such as asking questions, giving orders, making promises, stating facts, et cetera. In turn, the meanings of words and sentences are systematically bound together with the correlated illocutionary acts. "[T]here is a systematic set of relationships between the meanings of the words and sentences we utter and the illocutionary acts we perform in the utterance of those words and sentences" (Searle 1975, p. 320). Yet for Searle this creates a complication. There doesn't seem to be a prima facie explanation for how fictional sentences can utilize the illocutionary framework to their advantage, while at the same time repudiating that framework. "We might put the problem in the form of a paradox: how can it both be the case that words and other elements in a fictional story have their ordinary meanings and yet the rules that attach to those words and other elements and determine their other meanings are not complied with[?]" (Searle, 1975, p. 320).

In order to elucidate the underlying framework comprising our notion of fiction, Searle delineates this work from nonfiction. Works of the latter sort are subject to a lattice of different rules by which authors must adhere in order to maintain credence. In doing so, authors of nonfictional works employ assertions as their fundamental illocutionary act in order to convey meaning to their reader. Insofar as these assertions are subject to the systematic relations between written sentences and their meanings, they must conform to particular semantic and pragmatic rules.

(1) The essential rule: the maker of an assertion commits himself to the truth of the expressed proposition.

(2) The preparatory rule: the speaker must be in a position to provide the evidence or reasons for the truth of the expressed proposition.

(3) The expressed proposition must not be obviously true to both the [writer] and the [reader] in the context of utterance.

(4) The sincerity rule: the speaker commits himself to a belief in the truth of the expressed proposition (Searle 1975, p. 322).

Nonfiction writers must comply with all these rules; however, authors of fictional works make no commitment to the truth of propositions, vis-à-vis the aforementioned rules. For example, authors do not commit themselves to the truth that their characters exist. Nor do they believe that the characters' actions and circumstances are factually true. Authors are not in a position to provide the evidence for much of the novel, since they were not lurking around the characters and furtively recording their conversations and movements. Furthermore, the sentences that are printed for the reader are not pointless. Quite the contrary, for the first read through they are fresh and new. Finally, authors do not engage in self-deception in telling a story; there is no purported belief that the literary work comports with real world history.

However, given that authors do not commit to the truth of their fictional characters, a problem arises for Searle. Authors make assertions throughout novels, and these assertions are governed by the systematic framework of illocutionary acts. But these assertions, cannot be assertions, because they don't comport with the requirements of assertions.

> [The author] is making an assertion, and assertions are defined by the constitutive rules of the activity of asserting; but what kind of illocutionary act can [the author] be performing? In particular, how can it be an assertion since it complies with none of the rules peculiar to assertions? If, as I have claimed, the meaning of the sentence uttered by [the author] is determined by the linguistic rules that attach to the elements of the sentence, and if those rules determine that the literal utterance of the sentence is an assertion, and if, as I have been insisting, she is making a literal utterance of the sentence, then surely it must be an assertion; but it can't be an assertion since it doesn't comply with those rules that are specific to and constitutive of assertions (Searle 1975, p. 323).

In order to maintain the systematic framework that creates the relations between written sentences and their meanings, without committing oneself to the truth of the fictional character, the author simply pretends to make an assertion. In effect, the author begins a game with the reader, and they both play along. "[The author] is pretending, one could say, to make an assertion, or acting as if she were making an assertion, or going through the motions of making an assertion, or imitating the making of an assertion" (Searle 1975, p. 324).

Yet Searle also makes it clear that this pretense is not born out of a desire to deceive the intended audience. On the contrary, the pretense is well known to the audience, and in pretending, the author is engaging in a candid charade known to themselves and the reader. "[The author] is engaging is a nondeceptive pseudo-performance which constitutes pretending to recount to us a series of events. So my first conclusion is this: the author of a work of fiction pretends to perform a series of illocutionary acts" (Searle 1975, p. 325). Therefore, Searle makes some initial conclusions about the nature, and intention, of the author in their work. Searle holds that authors are engaged in pretending to make representative assertions, and in pretending to do so, they are doing this intentionally.

> [M]y first conclusion is this: the author of a work of fiction pretends to perform a series of illocutionary acts, normally of the representative type. [Furthermore] pretend is an intentional verb: that is, it is one of those verbs which contain the concept of intention built into it. One cannot truly be said to have pretended to do something unless one intended to pretend to do it. So our first conclusion leads immediately to our second conclusion: the identifying criterion for whether or not a text is a work of fiction must of necessity lie in the illocutionary intentions of the author (Searle 1975, pp. 324-325).

For Searle, this form of pretense is made possible by extralinguistic, nonsemantic conventions that break the aforementioned rules, by which nonfiction writers must abide. These new conventions do not alter the meanings of words; instead they allow authors to utilize these meanings *without* committing themselves to the truth of the fictional character.

> [A set of extralinguistic, nonsematic conventions does] not alter or change the meanings of any of the words or other elements of the language. What they do rather is enable the speaker to use words with their literal meanings without undertaking the commitments that are normally required by those meanings. My third conclusion then is this: the pretended illocutions which constitute a work of fiction are made possible by the existence of a set of conventions which suspend the normal operation of the rules relating illocutionary acts and the world (Searle 1975, p. 326).

Interestingly, this pretense is often accomplished in a very simplistic manner. Pretending to do something of a higher order actually entails doing something of a lower order, or simulating some of its constituent parts. For example, imagine the amount of pretending that an actor doing a pantomime needs to employ. However, if you consider the movements of a mime, vis-à-vis the normal movements necessary to carry out an according action, the pantomime is more simplistic. Washing a pretend window can be accomplished by movements of the arms in a pantomime, yet when done outside the realm of pretend, it is much more complicated. This same principle applies to the pretenses involved in writing fiction. "The author pretends to perform illocutionary acts by way of actually uttering (writing) sentences…The utterance acts in fiction are indistinguishable from the utterance acts of serious discourse, and it is for that reason that there is no textual property that will identify a stretch of discourse as a work of fiction" (Searle 1975, p. 327). These simplistic means by which the author of a work of fiction is able to escape the conventions of nonfiction writers are then used toward an end. Specifically, the author is able to utilize the pretense of simplicity in order to escape the conventions binding typical utterances. "The pretended performances of illocutionary acts which constitute the writing of a work of fiction consist in actually performing utterance acts with the intention of invoking the… conventions that suspend the normal illocutionary commitments of the utterances" (Searle 1975, p. 327).

Utilizing all of the preceding work, Searle is able to apply this to fictional characters in order to gain some insight. Yet in doing so, he makes sure to distinguish between serious discourse, fictional discourse, and serious discourse *about* fiction. In turn, suppose someone said the following about Dumas's work in *The Count of Monte Cristo*—there never existed a Mme. Mercedes Dantès, because Edmond and Mercedes were never married, but there did exist a Mme. Mercedes Mondego, because Fernand and Mercedes were married. If this were to be taken as a piece of serious discourse it would not be true because Edmond, Mercedes, and Fernand never existed. However, when taken as a serious discourse about a piece of fiction, it is true. This is because it adheres to the marriages that did, and didn't, occur in Dumas's novel. Taken as a statement about the fictional world of that novel, the statement would confirm to the systematic rules.

> I can verify the above statement by reference to the works of [Alexandre Dumas]. But there is no question of [Alexandre Dumas] being able to verify what he says about [Edmond, Mercedes, and Fernand] when he writes the stories, because he does not make any statements about them, he only pretends to. Because the author has created these fictional characters, we on the other hand can make true statements about them as fictional characters (Searle 1975, p. 328).

Most importantly, in order for an author to create fictional characters out of thin air, they must use a proper name (or a definite description or singular personal pronoun). In turn, this proper name pretends to *refer*. Since a referential act must include an object which is referenced, then there must exist an object. Therefore, by pretending to refer to an object, the author is pretending that there is an object to which a reference is being made. Furthermore, insofar as the reader shares the pretense, the reader will also pretend that there is a fictional character.

> By pretending to refer to (and recount the adventures of) a person, [the author] creates a fictional character. Notice that she does not really refer to a fictional character because there was no such antecedently existing character; rather, by pretending to refer to a person she creates a fictional person. Now once the fictional character has been created, we who are standing outside the fictional story can really refer to a fictional person (Searle 1975, p. 328).

This is what makes it not nonsensical to *actually* refer to a fictional character. The reader standing outside of the story doesn't need to pretend to refer to a fictional character; once the character has been created, people can actually refer to him or her.

> By pretending to refer to people and to recount events about them, the author creates fictional characters and events (Searle 1975, p. 329).

> As far as the *possibility* of ontology is concerned, anything goes: the author can create any character or event he likes (Searle 1975, p. 331).

In essence, an author pretends to refer to an individual by use of a proper name. However, a referential speech act is successful only if there exists an object to which the speaker is referring. So, an author pretends that there is an object to which there is a reference, by use of a proper name. Therefore, the author creates the fictional individual.

Denying Creationism in Fiction

The view that authors create fictional characters is contentious. Takashi Yagisawa is one philosopher who disagrees with creationism and criticizes the view, including Searle's version of it. In doing so, his aim is to attack the creationist position, which he believes has gained unwarranted popularity amongst analytic philosophers. "I shall focus on the view that fictional individuals exist as a result of being created by the relevant authors(s). Let us call this view *creationism in fiction*, or *creationism* for short" (Yagisawa 2001, p. 153). Given the overarching aim of discrediting the creationist camp, Yagisawa proceeds by arguing for the common pre-theoretical intuition that fictional individuals aren't actually real. Accordingly, he holds two claims to be true; namely, that fictional characters don't exist, yet they do exist "in the world of" the fiction in which they were created. In denying creationism, Yagisawa simply needs to argue that fictional characters don't exist. "Our pre-theoretical intuition says in general of any fictional individual that it does not actually exist but exists 'in the world of' the relevant fiction. I wish to defend this pre-theoretical intuition. To do so, I need to defend [the claim that] fictional individuals do not actually exist" (Yagisawa 2001, p. 153).

If successful, this would defeat the creationists view, insofar as it would be nonsensical to claim that an author created a nonexistent being. Thus, Yagisawa supports the opposing claim by arguing against creationism. "Creationism is the view that fictional individuals exist (i.e., actually exist) by being created by their author(s), and I shall defend the claim that fictional individuals do not (actually) exist by arguing against creationism" (Yagisawa 2001, p. 153). In attacking creationism, Yagisawa criticizes Searle's position on the genesis of fictional characters. "[M]y criticism will be directed at Searle's account of how an author of fiction creates a fictional individual" (Yagisawa 2001, p. 154).

Yagisawa revisits the way in which Searle argues for his conclusion, claiming that there is a disconnect between his final premise and his conclusion. Searle argues that (1) an author pretends to refer to an individual by use of a proper name; however, (2) a referential speech act is successful only if there exists an object to which the speaker is referring. So, (3) an author pretends that there is an object to which there is a reference, by use of a proper name. Therefore, (4) the author creates the fictional individual. Yagisawa then argues that the move that concludes the author creates the fictional individual is problematic. "The move from 1 and 2 to 3 seems acceptable. But the crucial step from 3 to 4 definitely seems abrupt and unwarranted" (Yagisawa 2001,

p. 155).

Yagisawa thinks that it’s not clear at all that the move creates a fictional individual out of thin air. Additionally, since Searle holds that a referential speech act is successful only if it refers to an object, problems arise. Namely, when saying things like “Dantès found Abbé’s treasure,” we are also saying that Dantès exists.

> There is an additional problem with Searle’s position. As we noted, he says that a speech act of reference is successful only if there exists an object the speaker is referring to. He also says with emphasis that we do really refer to [Edmond Dantès] when we say things like “[Edmond Dantès found treasure].” It follows then that Searle, when we say things like “[Edmond Dantès found treasure],” there exists Edmond Dantès (Yagisawa 2001, p. 155).

Yagisawa then suggests that we might interpret Searle to be saying that authors create fictional characters and bring them into existence in fiction, rather than existence in reality. While this idea seems to square with Searle’s notion of truth in serious discourse *about* fiction, Yagisawa still thinks he is at odds with the creationist position. “The view does not support creationism because creationism asserts the existence of fictional individuals in actuality, not just in fiction” (Yagisawa 2001, p. 156).

CONTINGENT CREATIONS

Yet it seems there is a way that can bring together the pretenses underlying Searle’s theory, with the nonexistence claims of Yagisawa. In doing so, one might argue that fictional characters are contingent, in an attempt to avoid the issues Yagisawa raises with creating necessary existence “out of thin air.”

> Most artifactualists, like Searle, take fictional characters to be created by authors pretending to refer to real people and places, and so take fictionalizing discourse to involve mere pretended assertions (Thomasson 2009, p. 11).

> This has inspired several recent theorists to begin by taking this sort of discourse as the focal case—a view that requires accepting that there are fictional characters and that these are created by authors in the process of writing works of fiction. Since they take fictional characters to be products of the creative activities of authors, call these 'artifactual' views of fiction (Thomasson 2009, p. 11).

Given the pretend assertions, akin to the aforementioned simplistic pantomimed game, there is a solution that allows referential objects to hold in a contingent manner. Thomasson has a view that states that these creations, by authors of fiction, are abstract artifacts that are contingent on physical instantiation or memory. In turn, they do not have a necessary claim to existence once created by an author. They are forever subject to the minds of thinking beings, and if they slip out of that consciousness, they cease to exist.

> [F]ictional characters are abstract artifacts created by authors' activities in writing or telling stories, and dependent for their ongoing existence on those stories (and copies or memories of them). The status of fictional characters as created, dependent, abstracta… is like that of many social and cultural entities such as laws of state, symphonies, and works of literature themselves: none of them may be identified with any concrete entity, none has a definite spatial location, but all come into existence at a particular time given certain types of human activity (Thomasson 2009, p. 14).

One of the most interesting things about this theory is that it can bring together parts of the seemingly contradictory positions of Searle and Yagisawa. Additionally, it does so without making any additional ontological commitments. Yet while it succeeds in doing so, it raises a thorny concern about the fictional characters themselves, given their contingent status. One might claim that it seems odd that an entity may pop in and out of existence, subject to the will of a thinking being.

Yet in response, Thomasson argues that such a contingent notion of creationism in fiction accords within our preconceived notions of the contingency of analogous entities that can serve as vehicles for creation, namely, laws, music, and literature. Thus in doing so, Thomasson's theory seems the strongest candidate for solving the seemingly contradictory implications stemming from the theories of Searle and Yagisawa, while gaining additional credence as an analog for other widely accepted abstract entities.

Bibliography

Crittenden, Charles. (1966) "Fictional Existence," *American Philosophical Quarterly* 3, pp. 317-321

Lamarque, Peter. (2003) "How to Create a Fictional Character," in: Gaut & Berys (Eds), *Creation of Art: New Essays in Philosophical Aesthetics*, pp. 33-52 (Cambridge: Cambridge University Press)

Sauchelli, Andrea. (2012) "Fictional Objects, Non-Existence, and the Principle of Characterization," *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 159(1), pp. 139-146

Searle, John. (1975) "The Logical Status of Fictional Discourse," *New Literary History* 6(2), pp. 319-332

Thomasson, Amie. (2009) "Fictional Entities," in: Kim, Sosa and Rosenkrantz (Eds), *A Companion to Metaphysics*, pp. 10-18 (Oxford: Basil Blackwell)

Yagisawa, Takashi. (2001) "Against Creationism in Fiction," *Nous-Supplement: Philosophical Perspectives* 15, pp. 153-172

The Fact of the Matter: Direct Reference Theory vs. Neo-Fregeanism

Melvin J. Freitas

There is a longstanding debate in the philosophy of language between direct reference theory (DRT) and neo-Fregeanism (N-F). The debate centers on a question concerning the types of propositions we express when we utter sentences containing singular terms, such as proper names or indexicals.[1] Do we ever *directly refer* to the object denoted by a singular term, or is reference always mediated by some *Fregean sense*, or way of thinking about the object? The historical background to this question is a variety of well-known semantic puzzles introduced by Frege (1892), Russell (1905, 1910a), Kaplan (1977, 1978), Kripke (1980), Salmon (1986) and many others. And these puzzles have generally centered on three kinds of propositions that are expressed, respectively, by sentences containing identity statements, vacuous names, and belief reports. Of course, there are many different versions of DRT, just as there are many different versions of N-F, each offering their own solutions to these puzzles. However, we will be considering both DRT and N-F as broadly construed types of semantic theories.

Balaguer (2011) argues that there is "no fact of the matter" between DRT and N-F, which is his shorthand way for saying that DRT and N-F are "tied in terms of factual accuracy" (Balaguer 2011, p. 53). Since, according to him, for any given version of DRT there is a theoretically analogous version of N-F, and vice versa, such that there is no fact of the matter as to which theory is better in terms of their factual accuracy. Furthermore, Balaguer thinks that the relevant facts for this debate, and others like them, are *empirical facts* about "the intentions of ordinary speakers"

(2011, p. 65). So it turns out to be only contingently true that DRT and N-F are tied in terms of explanatory power, since the empirical facts might have been such that either one of them was the correct, or at least better, semantic theory. That is, according to Balaguer, the intentions of ordinary speakers might have been such that some version of DRT was true, or they might have been such that some version of N-F was true, but as it turns out there is actually no fact of the matter between them. Which, if true, is a surprising discovery.

On the other hand, Fiocco (2011) argues that the relevant facts for a semantic theory like DRT or N-F are ultimately *metaphysical facts* about ontology. His main thesis is that there are only two tenable unified theories with respect to *realism* (or *anti-realism*), and *descriptivism* (or *anti-descriptivism*).[2] That is, Fiocco argues that the only coherent unified positions turn out to be *realist anti-descriptivism* and *anti-realist descriptivism*. His self-proclaimed "crucial argument" for this thesis is that *realism* paired with *descriptivism* entails the possibility of some *necessarily elusive entity* (i.e., an entity that cannot even "be merely thought of"), and that the existence of such an entity is strictly incoherent (Fiocco 2011, p.10). More importantly, in terms of semantics, Fiocco derives two corollaries from his main thesis to the effect that any commitment to *realism* should lead one to adopt DRT, while any commitment to *anti-realism* should lead one to adopt N-F (2011, pp. 5-6). That is, according to Fiocco, it turns out that our metaphysical commitments concerning ontology provide answers to at least some of our semantic questions. Which, if true, is also a surprising discovery.

My thesis is that, while I agree with Balaguer that the relevant facts for a semantic theory must always be facts about the intentions of ordinary speakers and that metaphysical facts alone will not do, I agree with Fiocco in thinking that there must be some fact of the matter between direct reference theory and neo-Fregeanism. That is, I think that DRT must be true in an obvious kind of way; otherwise, it's unclear what the whole debate is really about. To make this clear, we'll first look at Fiocco's main metaphysical argument for the relationship between ontology and intentionality, as well as the semantic corollaries that he derives from that argument. I will then argue that the relevant facts for a semantic theory must be facts about speaker intentions, since the whole debate between DRT and N-F rests on puzzles about what we *intend* to mean by what we say. Then, we'll look at Balaguer's argument for non-factualism, which is based on a thought experiment intended to show that theoretically analogous, and empirically tied, pairs of DRT and N-F theories can always be constructed. I will then argue that our ordinary intentions, when we use singular terms, are directly referential in an obvious kind of way.

Fiocco's Metaphysical Argument

Fiocco's argument begins with the following definitions. In terms of ontology, he says,

> *Ontological realism* is a view of the nature of reality according to which there are individual *objects* and many *kinds* that exist independently of thinking beings. Such entities have natures that depend in no way on conscious beings and so would exist as the very entities they are even had there never been any linguistic or mental activity or minds (Fiocco 2011, pp. 2-3).
>
> *Ontological antirealism* [is a view of the nature of reality according to which] what individual objects there are and what kinds of object[s] exist depend on thinking beings. The ontology of the world depends on the classificatory activity of such beings, conducted via thought and language (Fiocco 2011, p. 3).

Furthermore, in terms of intentionality, he says,

> *Descriptivism* [is the view that] the mind can be cognitively related to some feature of reality only by first preparing itself in an appropriate way; it must assume, by some means, a nature that fits uniquely whatever feature

> of reality to which it is to be related…. One way it might establish such a connection—a way to be regarded as paradigmatic for the purposes of the present discussion—is indirectly, by associating with an intermediate abstract entity that captures certain properties or encapsulates specific conditions (Fiocco 2011, p. 4).[3]

> *Anti-descriptivism* [is the view that] it is possible for one's mind to be cognitively related to some feature of reality without first preparing itself, without first assuming a nature by which it fits, in some way, that to which it is to be related. This view permits a *lack of spontaneity* in intentionality: a cognitive connection might be established between a mind and the world *from without*—merely by some object or some feature of the world impressing itself upon the mind (Fiocco 2011, p. 6).

Now to be clear, Fiocco makes no direct argument either for, or against, *ontological realism*; nor for, or against, *descriptivism*.[4] His main argument concerns the possibilities of a *unified* theory in terms of both ontology and intentionality. However, we are primarily concerned with Fiocco's two semantic corollaries for intentionality.[5] In this regard, he adds,

> [Descriptivism] provides a corollary account of the semantics of natural language. A linguistic item is about something in the world in virtue of having associated with it some intermediate abstract entity that captures certain properties or encapsulates specific conditions and thereby describes that thing…. On this account, reference is always mediated and, so, *indirect* (Fiocco 2011, p. 5).

> [Anti-descriptivism] provides a corollary account of the semantics of natural language [such that] some linguistic items can be about things *without being associated with any such intermediate abstract entities*. Thus, anti-descriptivism allows that the reference of a linguistic item might be *direct*, in that the item does, semantically, nothing more

than stand for a thing in the world (Fiocco 2011, p. 6).

Fiocco's self-proclaimed "crucial argument" for his main thesis is that *realist descriptivism* entails the existence of at least one *necessarily elusive entity*. And he argues that the existence of such an entity is strictly incoherent, therefore, by reductio, realist descriptivism is also incoherent. A *necessarily elusive entity* is "an entity that could not be considered via thought or language—a thing that one could not even begin to understand because one could not engage it with one's mind at all" (Fiocco 2011, p.10). That is, it is "an entity that cannot be merely thought of, by some mind or other, [even after] removing any practical impediment that might prevent one from thinking of (or referring to) [it]" (Fiocco 2011, p. 10). Fiocco first argues against the possibility of there being one and only one *necessarily elusive entity*, and then against the possibility of there being more than one such entity. In terms of there being one and only one such entity, he says,

> Suppose that there is a unique entity, *e*, that cannot, in principle, be thought of or referred to. If this is the case, *e* has a particular property that distinguishes it from every other existent thing: it is the sole entity that cannot in principle be thought of or referred to. Given this, though, *e* can be thought of or referred to; one need only bring before one's mind an abstract entity that captures this singular property of *e* and thereby descriptively fits *e*. (The needed abstract entity might be the property or complex expressed by the definite description 'the unique entity that cannot, in principle, be thought of or referred to'.) This is sufficient for a mind to establish a cognitive connection to *e* on either view of intentionality. But then *e* both cannot be thought of (or referred to) and can be thought of (or referred to). This is a contradiction and shows that the initial assumption, that there is a unique necessarily elusive entity, must be false (Fiocco 2011, pp. 11-12).

Then, in terms of there being more than one such entity, Fiocco says,

> Suppose, then, that there is more than one entity that cannot, in principle, be thought of or referred to. There is a particular property, *P*, that each one of these things bears, that distinguishes them from all others, namely, the property of being an entity that cannot in principle be thought of or referred to. Given the initial assumption, it is true that there is something that is *P*. It follows that *e* is *P*, where '*e*' is a name for one of those entities that bears *P*. But then an entity that is supposed to be such that it cannot be thought of or referred to can be thought of—as *e*—and *can* be referred to—by '*e*'. Consequently, *e* both cannot be thought of (or referred to) and can be thought of (or referred to). This is a contradiction and shows that the initial assumption, that there is more than one necessarily elusive entity, must be false (Fiocco 2011, p. 12).

Therefore, Fiocco concludes that "the very supposition that there is a necessarily elusive entity is incoherent and, hence, that such an entity is impossible" (Fiocco 2011, p. 12). This, he says, "yields a momentous conclusion: Everything that exists—universal or particular, substantial or non-substantial, concrete or abstract—is by some means accessible to the mind" (Fiocco 2011, p. 12). Which once again, if true, is a surprising discovery.

Given the impossibility of a *necessarily elusive entity*, Fiocco argues that any version of *realist descriptivism* must entail the existence of at least one such entity. To do this, he starts by assuming that ontological realism is true, so that there are individual objects that exist independently of thinking beings. "It seems, then, that there could be an object that has never been thought of (or referred to)" (Fiocco 2011, p. 13). But, if we assume that descriptivism is true, then it seems that we can just as easily assume that there could be an object that "cannot, in principle, be thought of or referred to [therefore, such an object] could be a necessarily elusive entity" (Fiocco 2011, p. 14). So, by reductio, *realist descriptivism* turns out to be an incoherent position. Fiocco also argues that "ontological antirealism entails descriptivism" and "descriptivism entails ontological antirealism" (Fiocco 2011, pp. 7-9). This, then, precludes the unified position of 'anti-realism anti-descriptivism'. However, we won't go into that part of the argument, as it will have no direct bearing the argument for my thesis. On the face of it, 'anti-realism anti-descriptivism' would seem to entail that there are no objects whatsoever, which is clearly going to be untenable for most everyone.[6] Therefore, according to Fiocco, the only two coherent views are *realist anti-descriptivism* and *anti-realist descriptivism*. And by the semantic corollaries he has given, respectively for descriptivism and anti-descriptivism, the logical result is that realism entails *semantic* anti-descriptivism (DRT) and anti-realism entails *semantic* descriptivism (N-F).

All of this leads Fiocco to the conclusion that "there are real and significant differences between descriptivism and anti-descriptivism"; and, while that "might seem obvious on the face of it," there is a significant group of philosophers who have maintained "that, there are, in the end, no substantive differences between the two views" (Fiocco 2011, p. 25). Of which, he includes Wagner (1986), Forbes (1987), Smith (1988), Caplan (2007), and Balaguer (2011) (Fiocco 2011, p.25 fn. 27-28). More significantly, Fiocco thinks that this shows that in the debate about reference, instead of "considering issues and examples tied to natural languages, one should begin with the underlying metaphysics: one should attempt to ascertain which ontological view of the nature of reality is true" (2011, p. 26). Generally, he says, that the traditional attempts to settle these semantic debates have relied on the wrong kind of examples, ones that disregard one's ontological views in the first place. He says, for instance,

> By focusing exclusively on cases in which cognitive contact has already been established between a thinker (or speaker) and some entity, the real differences between descriptivism and anti-descriptivism can be difficult to discern. This might give the impression that the two views are ultimately interchangeable or, at most, that any real

> difference between them is excessively subtle or esoteric. As I have tried to show, however, this is not so (Fiocco 2011, p. 25).

Fiocco thinks that the traditional puzzles about reference, having to do with identity statements, vacuous names, and belief reports, essentially miss the point. He sees this as coming from the fact that the puzzles generally rely on examples in which reference has already been established.[7] For when we talk about proper names such as "Aristotle," where we already agree on the referent, we already have a well-established cognitive connection to that object, such that the ontology of the object is not in question. But by setting aside the ontological issue, Fiocco argues, the answers to our semantic questions will not be forthcoming. But if we start with the underlying ontology, Fiocco's argument would entail that there must be *some* fact of the matter in the debate between DRT and N-F.

The Relevant Facts

In response to Fiocco's argument, I think we to need to step back and consider just what puzzles got the debate between DRT and N-F going in the first place. Canonically, this begins with Frege's and Russell's early attempts to establish *logicism*, the view that mathematics can be formulated solely in terms of (or reduced to) logic (Frege 1879; Russell 1903, 1910b, 1912, 1913). Of course, their attempts ultimately failed at the hands of Gödel's proof for the first *incompleteness theorem* (Gödel 1931). Nonetheless, it's important to note that Frege was primarily interested in mathematics; however, in his own words, "The logical imperfections of language stood in the way of such investigations" (Frege 1919, p. 253). And this led, for one thing, to his now famous puzzle about identity statements such as "a = a" and "a = b" (Frege 1892). For instance, if proper names have only reference (i.e.,, the refer directly), it seems we cannot account for the obvious difference in cognitive significance between "Samuel Clemens is Mark Twain" and "Samuel Clemens is Samuel Clemens". This is because the referent of “Samuel Clemens” is also the referent of “Mark Twain,” and on the view that names have only reference, only the referent can be contributed to the proposition expressed. That is, the proposition expressed by “Samuel Clemens is Mark Twain” would have to be identical to the proposition expressed by “Samuel Clemens is Samuel Clemens”. And although the latter sentence is obviously *a priori* by the law of self-identity, or a basic understanding of English, the former sentence is clearly *a posteriori* for everyone except possibly Samuel Clemens himself. On this basis, Frege argued that names must have both a *reference* (Bedeutung) and a *sense* (Zinn), where the sense of a name is the *mode of presentation* for, or way of thinking about, its referent (Frege 1892). For instance, someone might correctly think of Samuel Clemens as the publisher of the *Personal Memoirs of Ulysses S. Grant*, and correctly think of Mark Twain as the author of the *Adventures of Huckleberry Finn*, yet not know that Samuel Clemens and Mark Twain are one and the same man.

On the other hand, Kripke famously argued that names like “Samuel Clemens” and “Mark Twain” are *rigid designators*, that is, proper names *necessarily* refer to the same object in every possible world (Kripke 1980). But descriptions, such as “the author of the *Adventures of Huckleberry Finn*,” can only *contingently* refer to the same object. For most everyone would agree that someone else *might have* written the *Adventures of Huckleberry Finn*, but it does not seem right to say that Mark Twain *might not have* been Mark Twain. Though he may have certainly been called something else, the man that we actually call “Mark Twain” is necessarily that very man. Thus, if the sense of a name is something like a description (or even a cluster of descriptions) then it will fail to mediate reference in such a way that it always picks out the same referent. So while Frege was the *de facto* father of Fregeanism, Kripke and the majority of those that have followed him, have argued for some version of DRT. Nonetheless, there are many recent philosophers, the so-called *neo-Fregeans*, who have tried to revive Frege’s original intuition by one means or another. But my point here is not to go through all the puzzles and questions that have been asked in this lengthy debate. Rather, I want to consider the *relevant facts* for such puzzles and such questions as have been proposed for the debate.

What are the relevant facts for the semantic questions that Frege, Kripke, and many others have raised? What are the relevant data for solving the puzzles? For one, I think that it is clear that all of these questions are about our intentions as ordinary speakers of natural languages. We start with certain paradigmatic sentences, such as "Aristotle was the teacher of Alexander," and we reflect on the ordinary intentions that we, and others, have when we use these sentences in everyday communication. That any of these sentences are puzzling when considered under a given semantic theory, results from our ordinary use of them, what we intend to communicate by them, and what we take them to mean when they are spoken to us. It's our semantic theory that must conform to our intentions, not the other way around. Of course, different speakers may have different intuitions about overall speaker intentions on a case-by-case basis, and maybe some semantic questions have no factual answer (ala Balaguer). But when we consider the proposition we intend to express with "Aristotle was the teacher of Alexander," the relevant facts are facts about *what we mean* to say by the use of that sentence. Metaphysical facts can't decide the issue. In fact, if a metaphysical theory ultimately conflicts with *what we mean* by the sentences in our language, then that's a strike against that metaphysical theory.

Consider the very notion of propositions as abstract non-spatiotemporal objects that stand for what we mean, or what we say, by our words. Why would anyone begin to think that such things exist in the first place? In fact, many philosophers have resisted them because they seem to entail the existence of a mysterious "third realm" beyond the physical and mental. But the reason for thinking that propositions exist rests, ultimately, on what ordinary speakers mean by what they say or write along with the fact that we can communicate at all. It's not enough that *I* know what I mean if I am to communicate with you; it must be possible for *you* to know what I mean, at least most of the time. Now certainly, there are all kinds of metaphysical considerations about the nature of propositions. But like the semantic puzzles we've just considered, it's our notion of propositions that must conform to our communicative intentions, not the other way around. Think of it this way: the idea that some metaphysician would dream up propositions without considering the empirical facts about speaker intentions seems incredible. How could one reasonably posit the existence of objects that are purportedly beyond both the physical and mental realms without extra-ontological considerations, such as the phenomenon of human communication, mathematical knowledge, or some other such facts to inspire them?

Balaguer's Non-factualism

Balaguer's argument essentially begins with the claim that for every version of DRT, there is a theoretically *analogous* version of N-F, and that for every version of N-F, there is a theoretically *analogous* version of DRT. He calls these analogs, D-F pairs, and it's important to point out that Balaguer does *not* think that for every version of DRT or N-F, there is a factually *identical* version of the other theory. That is, the D-F pairs are not just notational variants of the very same semantic theory, nor can either member of each pair, simply be reduced to the other (Balaguer 2011, p. 66). Instead, Balaguer argues that the members of each D-F pair are, as it turns out, empirically "tied" in terms of factual accuracy. But that they are tied, is only contingently true, for the world might have been such that either member of each pair was a superior semantic theory. That is, the intentions of ordinary speakers might have been such that some version of DRT was the best theory, or they might have been such that some version of N-F was the best theory. But as it turns out, neither type of theory is better at accounting for the actual facts (Balaguer 2001, p. 66). So, essentially, Balaguer is arguing that the longstanding debate between DRT and N-F is indeterminate.

To draw this idea out, Balaguer has a thought experiment involving two pairs of fictional philosophers, Karl I and Karl II, paired respectively with, Fred I and Fred II. The two Karls share a great affinity with Kaplan and Salmon (as well as Kripke), and the two Freds share a great affinity with Frege and neo-Fregeans generally (such as Burge and Katz) (Balaguer 2011, pp. 54-55; p. 62 fn. 6). Each Karl-Fred pair is meant to represent a D-F pair, which, on Balaguer's view, is deeply analogous, and should lead one to see that such D-F pairs could be created *ad infinitum* (2011, p. 66). Balaguer develops the views of each Karl, and each Fred, in multiple stages, but then primarily focuses on the views of Karl II and Fred II, which we will do as well. The end result is a version of DRT (from Karl II), and a version of N-F (from Fred II), that have been subjected to various challenges in terms of the analysis of belief reports and vacuous names. For brevity, we will simply focus on the ending, where Karl II and Fred II offer their most refined versions, respectively, of DRT and N-F.[8] Specifically, they each offer an analysis of propositions involving belief reports, expressed by sentences of the form, '*S believes that a is F'*. However, it is important to keep in mind, that Balaguer is not claiming that either Karl II, or Fred II, has the *best* semantic theory in their respective camp (2011, p. 66). Rather, he's demonstrating how such theoretically analogous D-F pairs can always be constructed.

Karl II offers the following DRT analysis of belief reports.

> A sentence of the form 'S believes that a is F,' where 'a' is a name or indexical, is true iff S believes the singular proposition <a, Fness> under some (contextually appropriate) mental representation or other that, in the relevant context, is either coreferential with 'a' or covacuous with 'a' (Balaguer 2011, p. 74).[9]

To explain, we are considering propositions expressed by such sentences as,

(K) "Kayley believes that Stephen King is an author"; and

(D) "David believes that Santa Claus is nice" (Balaguer 2011, pp. 71-73).

In each case, there is a *believer*, ‘S’, and a *thing believed*, ‘a is F’, which is a proposition in-itself. There is also some property ‘F’ (i.e.,, ‘being an author’ or ‘niceness’) that is being predicated of ‘a’ by ‘S’. Notice also, that the name “Santa Claus” is vacuous, in that it has no referent, though four-year-old David thinks that it does. However, for brevity we will solely focus on the case of proper names that have referents, leaving aside both the case of vacuous names and indexicals.

The proposition expressed by (K) is true, according to Karl II, if and only if, Kayley believes the *singular proposition* <Stephen King, being an author>, where Kayley has in mind some contextually appropriate mental representation of, or way of thinking about, Stephen King. That is, if I say (K) to you, I may not know how Kayley is thinking of Stephen King, in fact, she may not even know that his name is “Stephen King.” Someone may have just introduced Kayley to Stephen, and simply said that he’s the author of *The Running Man*, whom Kayley knows to be “Richard Bachman.” In which case, Kayley may be representing Stephen King in a *Richard-Bachman-sort-of-way* (e.g.,, a Hollywood type, involved in science fiction movies), while I represent him in a *Stephen-King-sort-of-way* (a bookish fellow from Maine, who writes horror stories). However, what Kayley doesn’t know is that “Richard Bachman” is simply a penname for Stephen King. More importantly, Karl II thinks that in this context, the names “Stephen King” and “Richard Bachman” are directly coreferential to a particular man, and it’s *that man* who is a *direct* constituent of the proposition expressed by (K).

Fred II offers the following N-F analysis of belief reports.[10]

> A sentence of the form ‘S believes that a is F,’ where ‘a’ is a name or indexical, is true iff S believes some (contextually appropriate) neo-Fregean proposition of the form <S(a), the sense of ‘is F’> where S(a) is the sense of some expression ‘b’ that, in the relevant context, is either coreferential with ‘a’ or covacuous with ‘a’ (Balaguer 2011, p. 73).

To explain, let’s continue to consider the sentences (K) and (D). In each case, there is once again a *believer*, 'S', and a *thing believed*, 'a is F', which is a proposition in itself. There is also some property 'F' (i.e., 'being an author' or 'niceness') that is being predicated of 'a' by 'S'. However, Fred II's theory is different than Karl II's theory. For Fred II, the use of the proper name 'a' involves a mediating N-F *sense*, S(a), which is the sense of some expression called 'b' that is coreferential (or covacuous) with 'a'.[11]

So the proposition expressed by (K) is true, according to Fred II, if and only if, Kayley believes some *neo-Fregean proposition* of the form <S(Stephen King), the sense of 'is an author'>, where in the relevant context, S(Stephen King) is the sense of some expression that is coreferential with "Stephen King." That is, if I say (K) to you, I may have no idea what mediating sense Kayley has in mind in order to refer to Stephen King, even if she knows his name. Someone may have just introduced Kayley to Stephen, using the name "Stephen," however, she may know little if anything else about the man. So that, for Kayley, S(Stephen King) might be synonymous with the sense of "the man I've just now met," while for me, S(Stephen King) might be synonymous with the sense of "the most famous living horror novelist." More importantly, Fred II thinks that my sense for "Stephen King," in the context given, is coreferential with whatever sense Kayley has in mind, so that both of our *sense(s)* co-mediate reference to Stephen King, who is thus an *indirect* constituent of the proposition expressed by (K).

Now Balaguer's claim is that the views of Karl II and Fred II "are deeply parallel in pretty transparent ways" (2011, p. 54). In fact, one may initially suspect that the two views are identical. This is because the debate between DRT and N-F, roughly comes down to accounting for both the *denotation* and *connotation* of singular terms. Karl II emphasizes denotation, and accounts for connotation in virtue of the contextually appropriate *mental representation* we have in mind, for the object denoted. Fred II emphasizes connotation, and accounts for denotation in virtue of the mediating *sense* for the singular term, that occurs in the contextually appropriate proposition. But Balaguer sees these as two different views, since while Karl II thinks that the *object* denoted is a direct constituent of the proposition expressed, Fred II thinks that only the *sense* of the denoting term can be a direct constituent of the proposition expressed. Nonetheless, Balaguer has intentionally developed the views of both Karl II and Fred II in ways that are clearly parallel to each other, and thus theoretically analogous.

Balaguer's intuition is that for every semantic challenge one might come up with, for either Karl II or Fred II, the same thing is going to happen. He thinks there's always going to be a way to tweak either semantic theory in such a way that any challenge can be met (or unmet) in an analogous way. It's an argument by cases, and to my mind it works well enough so far. However, Balaguer also claims that "we don't have any *evidence* for thinking that ordinary speakers have the kinds of intentions they would need to have for there to be a fact of the matter in the debate between Karl II and Fred II" (2011, p. 66). And he thinks this is true no matter what version of DRT, or N-F, you care to propose. That is, Balaguer thinks that if you ask an ordinary speaker whether they mean to be expressing a Karl II-type proposition, or a Fred II-type proposition, when they utter either (K) or (D), they're liable to be confused by the question. For although we certainly *mean* something when we utter either (K) or (D), what we intend to say is explained equally well (or poorly) by either Karl II's or Fred II's analysis of belief reports.

The Fact of the Matter

I think that, though Balaguer has made some important observations about the debate between DRT and N-F, there is *some* fact of the matter between them. That is, I agree that the relevant facts for a semantic theory are facts about the intentions of ordinary speakers, but I also think that those facts support DRT in an obvious kind of way. For it seems to me, that the intentions of ordinary speakers are plainly directly referential in terms of what we *intend* to be saying when we use singular terms. When I say, "Aristotle taught Alexander," I'm talking about two particular men, and saying that the first man (Aristotle) was a teacher of the second man (Alexander). Such that the proposition expressed by that sentence is an ordered triple, including two men, and the two-place relationship of 'teaching' in the past tense, <Aristotle, Alexander, 'taught'>. In fact, I believe that proposition is true, come what may, no matter how I may be mentally representing Aristotle to myself. For I might even temporarily forget that Aristotle is named "Aristotle" and only remember him as the most famous ancient Greek philosopher who was born in Stagira. And I might tell you that, though I've forgotten his name, "The most famous ancient Greek philosopher born in Stagira was the teacher of Alexander." In which case, I will have expressed the very same proposition to you. Namely, that one man (Aristotle) taught another man (Alexander), regardless of how I may be thinking of either one of those men.

Nonetheless, Balaguer has provided an important analysis of the debate between DRT and N-F. Namely, that, for every version of DRT, or N-F, one can construct a theoretically analogous version of the other theory. That is, for every challenge that we might come up with for either Karl II, or Fred II, we can construct another theoretically analogous D-F pair, that potentially answers that challenge. However, I think that the various D-F pairs that can be constructed will not be tied in terms of factual accuracy. For there is one overriding speaker intention that stands in favor of DRT, namely, we obviously *intend* to be talking about, and predicating over, an individual object in our ordinary use of singular terms. And as semantic intuitions go, I don't think any other intuition is likely to be stronger than the intuition that DRT must be true in this obvious kind of way. Of course, there have been many challenges to this view, beginning with Frege's observation about identity statements. But it seems to me that any answer to this semantic question, is going to have to square up with our DRT intuitions, more so than with any other semantic intuitions we might have. For simply ask yourself, when making a typical referential-use of the name "Aristotle," of whom are you talking? At the end of the day, it's just obvious that speakers *intend* to be directly referring to an individual man, namely Aristotle.

Now one might simply object to this by saying that her own intuitions are plainly neo-Fregean, when it comes to the use of singular terms. That is, to simply say that one's intuitions are plainly of the DRT variety, is to make no argument at all. That's certainly a plausible viewpoint from the perspective of the various challenges that have been made against DRT. Nonetheless, I think that the onus must fall on the advocates of N-F to show that their intuitions are stronger. For it makes perfect sense to ask for the definition of a word like "philosopher," but it's clearly a category mistake to ask for the definition of "Aristotle." For all we can do, in the latter case, is point at the man himself by way of some sort of demonstration. In this regard, I agree with both Kaplan's (1977) and Salmon's (1986) intuitions that names are, at least in some ways, very similar to indexicals. For I might say to you that "Aristotle is the most famous ancient Greek philosopher born in Stagira," but that's not to say that "Aristotle" means the same as "the most famous ancient Greek philosopher born in Stagira." It's simply a way of my pointing out to you what my use of "Aristotle" refers to, that is, a way of showing you that I'm directly referring to a specific individual who happens to be (or so I think) the most famous ancient Greek philosopher born in Stagira. On the other hand, the advocate of N-F must come up with a sense for "Aristotle" that's always going to pick out the correct individual in all relevant contexts. And even if they do come up with such a sense, which seems unlikely, it's incredible to think that that's what an ordinary speaker has in mind when they use the name "Aristotle." Just ask them. Their apt to tell you that they simply meant to refer to Aristotle, and had no such N-F sense in mind, whatever that might be.

Here's another way to look at it. Balaguer is arguing that the non-factualism between DRT and N-F is only contingently true when it comes to our ordinary speaker intentions. That is, there is a possible world where DRT is the best semantic theory, and there is a possible world where N-F is the best semantic theory, but the actual world is one in which neither is better. So let's imagine two possible worlds, one in which DRT is the better semantic theory, call it W_{DRT}, and another where N-F is the better semantic theory, call it $W_{N\text{-}F}$. Each of these possible worlds is just like our own, but purportedly different in terms of the intentions of ordinary speakers when it comes to our use of singular terms. Now to my way of thinking, it's obvious that the actual world is just like W_{DRT} for the reasons just given, but also because it's hard to imagine what it would be like to live in a world like $W_{N\text{-}F}$. Balaguer claims that it would simply be like living in a world where the neo-Fregeans turn out to be correct about the nature of propositions. But, what exactly would that be like? For instance, say some ordinary speaker says "Aristotle taught Alexander" in $W_{N\text{-}F}$ and we ask them to explain exactly what they meant by that utterance. Presumably, they would say that their intentions are plainly neo-Fregean, since by "Aristotle" they always mean to be talking about, let's say, "the most famous ancient Greek philosopher born in Stagira." Could we possibly be living in a world like that? I think that the answer is plainly "no." For we plainly mean to be saying that a particular man taught Alexander, come what may, since for one thing Kripke has shown that "Aristotle" rigidly designates that man, in a way that the "the most famous ancient Greek philosopher born in Stagira" cannot. For although the most famous ancient Greek philosopher born in Stagira *might not have been* Aristotle, Aristotle is *necessarily* Aristotle. And our use of the name "Aristotle" is thus plainly directly referential.

CONCLUSION

Fiocco claims that our ontological commitments to metaphysical realism (or anti-realism) ultimately ground our commitments to semantic descriptivism (or anti-descriptivism). However, I have argued that the relevant facts for a semantic theory are always facts about ordinary speaker intentions, since the puzzles that drive these debates are all grounded in questions about speaker intentions. Metaphysical facts, alone, cannot answer questions about what speakers intend to say when they use singular terms. Balaguer claims that, as such, there is no fact of the matter between DRT and N-F, since either theory can be constructed in a theoretically analogous way, such that they are both tied in terms of factual accuracy. However, I have argued that speaker intentions are plainly directly referential in an obvious kind of way. Speakers obviously intend to directly refer to individual objects in their ordinary use of singular terms. This might very well be unsatisfying to the reader, since the debate seems to come down to a case of dueling intuitions, of the DRT variety on the one hand, and the N-F variety on the other. But certainly some intuitions are stronger than others, and it's counter-intuitive to think that ordinary language speakers have full-fledged N-F senses in mind whenever they use singular terms. Especially, since there's no clear consensus as to the exact nature of N-F senses. Thus, the onus is on the advocate of N-F to show that our ordinary speaker intentions always involve mediating *senses* when using singular terms. The more basic intuition is that singular terms directly refer to the object we intend to denote. That constitutes compelling evidence, in my mind, that DRT is the better semantic theory.

Notes

1. This debate (and this paper) assumes that one accepts the existence of propositions in the first place, which are canonically: (i) the *meanings* or *thoughts* expressed by well-formed sentences of the natural languages in which they occur, (ii) *abstract* (non-spatiotemporal) objects that can be shared inter-subjectively, (iii) the bearers of *truth-value*, and (iv) the objects of the various *propositional attitudes*, such as belief.

2. Note that in this context "*descriptivism*" is <u>not</u> synonymous with neo-Fregeanism.

3. Fiocco is here keying into the well-known view that propositions are "intermediate abstract entities." However, he does acknowledge that one could just as easily adopt some other view on the nature of intentionality. For instance, he mentions that one could have an "adverbial variation of descriptivism" such that a mind could instead prepare itself in an appropriate way by "standing Φ-ly," with respect to some feature of reality (Fiocco 2011, p. 4 fn. 4).

4. However, at the very end of the article, Fiocco does admit that he thinks that "ontological antirealism is demonstrably false and so, consequently, is descriptivism" (2011, p. 27). So, as we will see, it's safe to assume that

Fiocco also advocates some version of DRT, although he makes no direct statement to that effect.

5. Fiocco is making a distinction between *cognitive* intentionality, in terms of a mind's being "cognitively related" to an object, and *semantic* intentionality, in terms of a linguistic item's being about an object. So, while the former is strictly a metaphysical concern, the later can be seen as strictly semantic.

6. Fiocco also points out that the argument against 'realist descriptivism' is the more interesting since, as Fiocco puts it, "many prominent philosophers, including Frege, David Lewis and Frank Jackson, have assumed, without question, that the two can and do go together" (2011, p. 1).

7. A similar point is made by Almog (2012).

8. By so doing, it's important to note that I'm skipping over a large part of Balaguer's argument, in terms of his demonstrating in detail how these D-F pairs are theoretically analogous. However, I do not dispute that such D-F pairs are always possible. But, if the reader wants the full picture, they should consult Balaguer's (2011) article directly on this point.

9. Balaguer defines 'covacuity' as follows,

   > [It is a] relation that does for vacuous terms what the relation of corferentiality does for non-vacuous terms. In other words, the idea here is that the relation will hold between, e.g.,, 'Santa Claus' and 'Kris Kringle', and 'Romeo' and 'Juliet's boyfriend', but not between 'Pegasus' and 'Oliver Twist', or 'Sinbad' and 'Mrs. Dalloway' (2011, p. 73).

10. Fred II's view is developed extensively in Balaguer (2005).

11. Again, we will set aside the case of vacuous names and indexicals. Similarly, we won't discuss the *sense* of predicates, since for one thing, Karl II need not disagree with Fred II in their analysis, since many advocates of DRT freely acknowledge that predicates have *senses*.

Bibliography

Almog, Joseph. (2012) "The puzzle that never was—referential mechanics," in: R. Schantz (Ed), *Current Issues in Theoretical Philosophy*, *Prospects for Meaning* 3, pp. 21-34 (Germany: Walter de Gruyter)

Balaguer, Mark. (2005) "Indexical propositions and de re belief ascriptions," *Synthese* 146, pp. 325-355

_____. (2011) "Is there a fact of the matter between direct reference theory and (neo-)Fregeanism?," *Philosophical Studies* 154, pp. 53-78

Caplan, Ben. (2006) "On sense and direct reference," *Philosophy Compass* 1, pp. 171-185

Fiocco, M. Oreste. (2011) “Descriptivism and ontological realism,” Retrieved from http://www.umass.edu/philosophy/events/papers/Fiocco%20-%20Descriptivism%20and%20Ontological%20Realism.pdf

Forbes, Graeme. (1987) “Review of Nathan Salmon’s *Frege Puzzle*,” *The Philosophical Review* 96, pp. 455-458

Frege, Gottlob. (1879) “Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought,” S. Bauer-Mengelberg trans., in: J. van Heijenoort (Ed), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931* (1967), pp. 1-82 (Massachusetts: Harvard University Press)

_____. (1892) “On sense and meaning,” Max Black trans., in: A.P. Martinich (Ed), *The Philosophy of Language* (1985), pp. 200-212 (New York: Oxford University Press)

_____. (1919) “Notes for Ludwig Darmstaeder,” P. Long & R. White trans., in: H. Hermes, F. Kambartel, & F. Kaulbach (Eds), *Posthumous Writings* (1979), pp. 253-262 (Oxford: Basil Blackwell)

Gödel, Kurt. (1931) “On formally undecidable propositions of Principia Mathematica and related systems I,” J. van Heijenoort trans., in: J. van Heijenoort (Ed), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931* (1967), pp. 592-617 (Massachusetts: Harvard University Press)

Kaplan, David. (1977) “Demonstratives,” in: J. Almog, J. Perry, & H. Wettstein (Eds), *Themes from Kaplan* (1989), pp. 481-563 (New York: Oxford University Press)

_____. (1978) “Dthat,” in: A.P. Martinich (Ed), *The Philosophy of Language* (1985), pp. 315-328 (New York: Oxford University Press)

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

Russell, Bertrand. (1903) *The Principles of Mathematics* (Cambridge: At the University Press)

_____. (1905) “On denoting,” *Mind* 14, pp. 479-493

_____. (1910a) “Knowledge by acquaintance and knowledge by description,” *PAS New Series* 11, pp. 108-128.

_____. (1910b, 1912, 1913) (with Alfred North Whitehead) *Principia Mathematica,* 3 vols. (Cambridge: Cambridge University Press)

Salmon, Nathan. (1986) *Frege’s Puzzle* (Cambridge: MIT Press)

Smith, David. (1988) “Review of Nathan Salmon’s *Frege’s Puzzle*,” *Mind* 97, pp. 136-137

Wagner, Steven. (1986) “California semantics meets the great fact,” *Notre Dame Journal of Formal Logic* 27, pp. 430-455

PROMISING: AN INTUITIVE RATIONAL CONVENTION

Nigel Aitchison

I. INTRODUCTION

A convention is a social practice that a society prefers to be some way rather than some other way. Just as traditions over time become commonplace, a convention does not need a rational foundation, it simply requires that the population both (a) prefer things to be one way rather than some other way, and more importantly, (b) that everyone generally conforms by acting one way, rather than some other way. Hume, for example, thinks that the social practice of promising is just this sort of thing. Promising, to Hume, is a social practice that is conformed to due to preference and happenstance. While promising is a social practice, it is not the case, I think, that promising can be reduced to a practice that is conformed to, due to mere coincidence and preference. Though reason often has nothing to do with a regular everyday convention—such as the convention to not wear sandals with slacks—it is not the case that all conventions can be accounted for due to happenstance. Rather, some conventions, such as the social practice of promising, are created and conformed to due to reasons that transcend mere coincidental preference.

Throughout this paper, I will be discussing the social practice of promising. I will first discuss the intuitive notion of promising, as well as the purpose and importance of promising. I will then discuss a conventional account of promising, and then move on to discuss why it is rational to keep a promise based upon self-interest. I will conclude with a brief exposition of Scanlon's expectationalist promissory theory, how that theory can account for important intuitive notions about promising, and I will then suggest a type of hybrid view between conventional and expectational accounts of promising.

II. The Intuitive Notion of Promising

The most intuitively obvious characteristic of a promise is that a promise made between two people obligates the one making the promise—i.e., the promisor. The person to whom the promise is made—the promisee—is conferred a special standing in relationship to the promisor, in which case the act promised is, in one sense, his *right*. This is to say that when a promisor makes a promise to perform some act 'X', the promisee is granted a right to the promisor's performance of 'X'. The promisee, then, has (a) a power to release the promisor from his obligation to perform 'X', (b) the right to demand performance of 'X', or (c) the right to rebuke the promisor for non-performance of 'X'. This is to say that the promisee has both a right with respect to the promised action, and a power, in a Hohfeldian sense, in respect to his ability to release the promisor from his obligation to perform 'X' (Hohfeld 1964, p. 36). For example, Al promises Bert that he will pay him five dollars. When the promise is made between these two people, the standing between them changes along with it. Where previously Al and Bert were on equal footing, neither owing the other anything, the promise intuitively changes that relationship. The moment Al makes the promise, he (a) obligates himself to pay Bert five dollars, (b) confers a right to Bert—i.e., it confers Bert's right to Al's paying him five dollars, and (c) grants Bert a power over him with respect to the promised action. Al has an obligation to pay Bert five dollars, and Bert has a *right* to Al's payment—i.e., the ability to demand payment, and the ability to rebuke Bert for non-payment—as well as a *power* to release Al from his obligation. Al can either pay Bert the five dollars, or renege on his promise and risk physical, legal, or moral rebuke. Bert, as the promisee (and subsequently, the right and power holder), can (a) demand payment or rebuke Al for non-payment by invoking his *right* to the promised action, or (b) release Al from his obligation (and thus even their standing once again) by invoking his *power* with respect to Al and the promised action.

A promise does something special to the relation between people: it alters the power relations between those involved. A promisor gives up a power he has over himself, and grants that power to the promisee. By doing so the promisee gains some amount of control over the promisor with respect to the promised performance of some action 'X', and the promisor loses his power in exchange for an obligation as to the performance of the promised action 'X' (Hohfeld 1964, p. 36). This is to say that prior to the promise being made the promisor had a choice to do as he wished in respect to the performance of some action 'X', however, after making a promise he no longer has that power, but instead has a responsibility to perform 'X' as owed to the promisee.

III. The Purpose of Promising

The act of promising is in some way, shape, or form an everyday occurrence. People use promises/agreements/covenants/contracts etc., regularly in order to streamline cooperation with other people. Trustful social coordination is one of the main purposes of promising, and as such, being able to make a promise, or exchange promises, in order to obtain some mutually beneficial end is a societal cornerstone. In the absence of promising—that is, essentially, the absence of trusting cooperation—it is hard to conceive of how we could engage in any type of trusting social coordination.

Promises tend to be a little more complex than the aforementioned example with Al and Bert. Most promises usually fall into the bi-directional category. For example, let's say that Al asks Bert to meet him at Starbucks at 5pm, and Bert agrees to do so. What ensues is a classic example of a mutual promise. Both parties are, in some sense, promising to do some act—i.e., meet at Starbucks at 5pm—so, unlike the first one-directional account, by the mutual-promise account Al and Bert are both simultaneously a promisee and a promisor in that they are both obligated to meet the other person at Starbucks, and they both have a right to the other person's meeting them at Starbucks (with the standing to rebuke, power to release, etc.). This is a paradigmatic example of how promising aids us in trustful coordination.

A promise does not come in a 'one size fits all' sort of package. A promise can be spelled out clearly, as in the 'I promise you that I will do X' variety, however, most of the time promises are disguised in everyday language. For example, if Bert says 'meet me at X', and Al says 'Ok', then it seems as though that is, for all intents and purposes, a binding promise. A promise is not a magical incantation that, when uttered correctly, conjures an obligation from beyond. It is rather an everyday means that enables us to engage in trustful coordination with other people, and as such, promises can come in all shapes and sizes and can be established in all sorts of ways.

IV. Conventionalist Promissory Theory:

Now that we have discussed what the intuitive notion of a promise is, we must discuss why, or how it is that a promise obligates. On the one side, there are the natural law theorists, such as Aristotle and Aquinas, who believe that one should keep a promise because it is, more or less, the moral/virtuous/right thing to do. There are others, like Scanlon, who believe that promises obligate because assurance, intent, and consequently trust is established between the promisee and the promisor. Then there are the conventionalists, like Hume and Lewis. The conventionalists hold that promising is a social convention, and that the obligating power that a promise has is bestowed upon it by the convention of promising itself. The convention of promising, by this view, is rule-based, and is governed by the very society in which it is established. As such, promising is limited to that society in which it is an established convention. I am using 'convention' as defined by David Lewis:

> A regularity R in the behavior of the members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common

> knowledge in P that, in almost any instance of S among members of P,
>
> (1) Almost everyone conforms to R;
>
> (2) Almost everyone expects almost everyone else to conform to R;
>
> (3) Almost everyone has approximately the same preferences regarding all possible combinations of actions;
>
> (4) Almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R;
>
> (5) Almost everyone would prefer that any one more conform to R′, on condition that almost everyone conform to R′.
>
> Where R′ is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R′ and to R (Amadae 2011, p. 330).

According to this view, promising, like any other convention, is a social practice, R, to which some population, P, conforms because they both expect and prefer that practice R be conformed to. The wrong in breaking a promise here is not specifically moral, nor is it directed at the promisee, but, rather, it is the general wrong of acting contrary to the social convention (and against social preference), and the free riding on other people's conforming to that convention. If, for example, Al makes a promise to Bert to buy him a drink, and then Al refuses to purchase said libation for Bert, it is not the case—by the conventionalist view—that Al is wronging Bert *directly*. It is rather the case that Al is committing some general wrong against all those who uphold the convention of promising, and by both expecting other people to keep their promises to him, while not upholding his own promises, he is free riding upon that convention.

The bare bones conventionalist, like Hume, would argue that there is no strictly moral wrong in breaking a promise. Breaking a promise is, rather, a wrong that is akin to wearing sandals with slacks, boys wearing pink, and not holding the door open for an elderly woman. The wrong committed in breaking a promise is less like a ‘wrong’, in a moral sense, and is more accurately described by the general term used when one goes against a convention, that is, it is better described as being *rude*. Since by this view promising is—like boys wearing blue—based upon mere happenstance and preference, the only wrong committed is the wrong of going against a particular population’s preferences and sensibilities. This is not to say that conventionalism does not allow for varying degrees of wrong—i.e., rudeness. For example, a conventionalist does not have to say that the wrong committed in breaking one convention—like breaking a promise—is the exact same as the wrong of breaking another—like a boy wearing pink. However, seeing as in both cases a convention is being broken, the wrong committed—while potentially differing greatly in severity—is still just the wrong of going against a society’s preference. Thus the two wrongs committed might not be equal, but they are in the same realm of wrong. This view does not capture some key intuitive notions of promising, that is, it does not capture either (a) the intuition notion that one should keep his promises, or (b) the intuitive directional wrong done toward the promisee in breaking a promise. In the following sections, I will be discussing a self-interested justification for why it is rational to keep one’s promises, as well as a Rawlsian rule-utilitarian take on promises. I will then move on to discuss an expectationalist view that accounts for the intuitive directional wrong we feel in breaking a promise.

The motivation to keep one’s promises under a conventionalist view is a self-interested utilitarian one (Rawls 1955, p. 16). By conforming to the convention of promising, one is able to benefit from trustful coordination with others via the making or exchanging of promises. While it is not the case that by breaking a promise one is committing a specifically *moral* wrong, it is the case that by breaking a promise one harms his own reputation as part of the population P who conform to the convention R—i.e., a promise breaker is trusted less, and will eventually be unable to participate in trustful coordination whatsoever. The motivation to keep promises, then, is a self-interested motivation that is based upon one's interest (and indeed the whole society's preference) to benefit from the practice of promising by allowing for trustful coordination.

In his paper 'Two Concepts of Rules', Rawls discusses promising as a utilitarian practice. While conforming to the convention of promising can be seen as being based upon utilitarian principles—i.e., bringing about the most overall benefit, etc.—it is not the case, according to Rawls, that when considering a particular promise, one can apply the same utilitarian principles used to justify the practice as a whole. It could be reasonably argued, for example, that one should renege on a promise because—as utilitarian principles dictate—the most overall utility would be gained by reneging on that promise. For Rawls, it is not the case that one should use utilitarian calculations to deliberate whether to keep this or that promise, but, rather, one must use them only in deliberating whether (a) upholding the practice of promising will maximize utility, or (b) tearing down the practice of promising will maximize utility. In the following section I will discuss two accounts of promising. First I will discuss a self-interest account, as well as why it is rational under this doctrine to keep promises, and then I will move on to discuss a Rawlsian rule utilitarian account of why one should keep promises.

V. Why Keep a Promise? A Rational Account

If one is under the impression that some sort of objective morality holds the power of a promise—i.e., natural law, etc.—then the problem of why one should keep a promise is, in one way, solved by that belief. However, if one does not believe in an objective morality (natural law or otherwise), then one must inquire as to why it is we should keep our promises, or at least why it would be *rational* to keep our promises. The problem at the heart of the matter is trustful coordination. The coordination problem is common in game theory, and it deals with the question of why should we cooperate—especially in situations where the action of the other party is uncertain. This is illustrated by such thought experiments as the 'prisoner's dilemma', which generally involves a payoff matrix with two parties, and four possible outcomes. For the sake of illustration I will construct a payoff matrix of a mutual promise—where each party makes a promise to the other to do some action—and I will analyze the various possible outcomes based upon the potential benefit gained or lost. Let's say that there are two fellows, Jake and Ben. Jake is an expert window washer, but he hates to mow his lawn. Ben, on the other hand, has a knack for mowing lawns, but he hates to wash windows. Seeing as they recognize each other's strengths and their own weakness, they each make a promise to each other; Jake promises to wash Ben's windows, and Ben promises to mow Jakes lawn. The resulting payoff matrix is as follows:

	Ben Mows Jake's Lawn	**Ben doesn't mow Jake's lawn**
Jake washes Ben's Windows	1 , 1	–1 , 2
Jake doesn't wash Ben's windows	2 , –1	0 , 0

The unit of measurement is—for the purposes of this paper—a 'utile'. A utile represents a unit of benefit, or utility, where a positive number of utiles represents an amount of utility gained, and a negative number of utiles is a loss of utility. So, from this matrix we can see four possible outcomes. If both Jake and Ben fulfill their promises—as seen in the top left box—then they both stand to gain an equal benefit of 1 from their respective promises. If one fulfills the promise, and the other reneges—as seen in the top right or bottom left box—then the reneging party stands to benefit whilst not having to do any work—thus potentially gaining the maximal benefit of 2—and the other person would do some work without gaining any benefit, thus resulting in a –1 loss of utiles. The last possibility is where they both renege, and neither gains nor loses anything.

The first type of approach we will discuss here is that of a ‘self-interested agent’ (henceforth SIA), that is, an account of an agent who only considers his own benefit and loss in his utilitarian calculations. In the aforementioned case, one who acts solely on self-interest would look at the possible outcomes, and would choose the option that would result in the most personal utility in this instance. Thus, it seems the SIA would renege on the promise, because reneging in that particular case can have the most personal benefit. If, however, both parties were SIAs, then presumably both would choose to renege, and thus no benefit would be had on either side. This is precisely the type of case that leads to the practice of promising being rendered moot. If everyone acted as a strict SIA, then it seems as though no one would keep a promise if reneging would produce more personal benefit, and hence promises would be rendered useless as a social practice.

When considering the example as a ‘one-shot’ sort of deal where both parties are SIAs, it seems as though the SIA would be making the rational choice by reneging on his promise, seeing as both parties stand to potentially gain more and lose less by not fulfilling their promise (since by not fulfilling the promise one is not risking the possible negative outcome). However, when considering promising as a social practice that one partakes in regularly within a society, the outcome is much different. By upholding the practice of promising, the SIA stands to gain from the maintaining of trusting promissory relations, as well as by maintaining his reputation as a member of the promising community. Even on a small scale, it is clear that more benefit is to be had by maintaining the practice of promising—even if the motivation for doing so is based solely upon self-interest. If, for example, Jake and Ben were to consider the potential benefit they could gain by exchanging promises every week for a year to do the aforementioned chores, then, as it is seen in the following diagram, the benefit is clear:

(Ex) A one-time benefit of mutual cooperation as per the previous diagram: 1, multiplied by 52.

Once a week for 52 Weeks	Ben Mows Jake's Lawn
Jake washes Ben's Windows	52 , 52

Thus, as per the example, when considering potential future benefit, the SIA would be rational in keeping his promises, prolonging trusting relations, and maintaining his reputation as a trustworthy practitioner of the social convention of promising rather than sacrificing all future benefit for a paltry one time gain.

Unlike the SIA account, the rule utilitarian does not consider the *personal* utility of a particular case, but rather the overarching utility gained for *everyone* by having promising as a social practice. Where the SIA only considers his own utility in his decision, the rule utilitarian considers the utility of all those who participate in the practice of promising. The rule utilitarian applies utilitarian principles to decide whether to uphold the overarching practice of promising, rather than using those principles to decide on a particular case. The rule utilitarian asks the two following questions:

(a) Is it more beneficial to have promising as a social practice? If so, then I should keep my promises in order to uphold that practice; or

(b) Is it more beneficial to not have promising as a social practice? If so I should renege on all of my promises.

Instead of considering a single case based on personal utility—as the SIA would—the rule utilitarian considers the utility of the overarching social practice of promising in order to justify the keeping or breaking of promises in general (Habib 2008). In the case of Jake, Ben, and the exchange of promises, if either party were to consider the potential benefit that promising as a practice provides, rather than a single case, then it seems as though they would uphold their promises—even if they do so only to benefit personally by upholding the practice as a whole. Unlike the SIA, a utilitarian would have no problem deciding what to do in the aforementioned cases. If, for example, we consider the *combined* potential utility gained, then it is clear that both parties keeping their promises would result in maximal utility. For example, if one party reneges, and one keeps his promise, then the combined total utiles gained would be 1—since one side gains 2 utiles, and one side loses 1, resulting in a total of 1 utile gained. If both parties renege, then 0 utiles are collectively gained. If, however, both parties fulfill their side of the promise, then a combined total of 2 utiles are gained, thus the most utility is gained if both sides fulfill their promises. Utilitarianism is, however, a moral theory that adheres to the basic tenet that one should act in such a way as to bring about the most benefit and least loss. As with many other moral theories—like natural law, etc.—accepting such a moral theory solves the previously pressing dilemma as to whether one should or should not keep a promise. As such, the introduction of rule-utilitarianism is only expository, and is only meant to offer another angle on the problem. A SIA would do well to approach his promissory relations with a rule-utilitarian leaning, seeing as a rule-utilitarian approach would undoubtedly lead to more personal benefit in the long run from prolonging those promissory relationships, even if it were under the rouse of utilitarianism. After all, if the SIA would apply his own principles to all his promissory relationships, he would subsequently renege when it benefitted him most, and eventually he would be unable to participate in the practice of promising for being known as a promise-breaker.

VI. Problems with the Self-Interest Account and Conventionalism

The elephant in the room, so to speak, with the SIA account is the dreaded 'one-shot' deal (death bed cases, state of nature cases, etc.) where there is no potential for future benefit or loss of reputation. Even though a SIA would be rational in *generally* keeping promises and maintaining the practice of promising, when considering 'one-shot' promises it seems as though it actually *is (or can be)* rational to renege, and to thus gain the maximal one time benefit in cases where there are no consequences nor potential for future benefit. For example, let's say Jake is visiting Ben on his deathbed, and Jake and Ben exchange promises. Ben promises Jake to leave him his Model-T in his will and Jake promises Ben in exchange that he will shoot Ben's ashes out of a cannon. Ben fulfills his side of the promise by leaving his Model-T to Jake in his will, and he dies directly afterwards. Jake is now in a position where he has exchanged promises with Ben, he has gotten his side of the promise paid in full, and he hasn't yet fulfilled his side of the bargain. Knowing that a) the promise was just between him and Ben, b) that no one would be the wiser if he reneged—since the only person with knowledge of the arrangement is now deceased—and c) that there is no potential for future benefit, it seems as though the SIA version of Jake would subsequently renege and not go through the trouble of commissioning the ash cannon for the dear old Ben. The rule-utilitarian version of Jake, on the other hand, would uphold his promise, since he would not be considering the personal utility of the particular promise, but rather the overarching utility of the practice of promising when deciding whether to renege or fulfill his promise. These cases pose a problem for promissory theory, since in such cases it is hard to imagine a scenario—in the absence of an objective morality—where it would be strictly rational for Jake fulfill the 'one-shot' promise he made to Ben. One could make the argument that one can never foresee the consequences of his actions, and as such the SIA should act in such a way as to minimize potential negative consequences. As in the case of Jake and Ben, Jake cannot be absolutely certain that Ben has not been recording their conversations, nor could he be aware that there is a clause in Ben's will that states that if Jake fails to fire his ashes out of a cannon—and thus fails to fulfill his promise—then Jake would no longer get the Model-T. In the set of potential negative consequences for Jakes reneging lies a laundry list of negative possibilities, and to fulfill his promise would result in more utility than many of them. So, by this rationale, even Jake should not renege on his promise, because there could be all sorts of unforeseen consequences to his reneging including, but not limited to, his not getting the prized Model-T. The argument for unforeseen consequences is not terribly strong. People tend to act based on probability, and even if there is a small chance of a negative consequence, given a high enough probability for success the risk will be generally be taken. So the SIA, given a high enough probability for getting away with it, would renege on a promise despite the possibility for unforeseen consequences.

Another intuitively important problem with a self interest conventionalist approach is that while it rationalizes why one should generally keep a promise, it does not account for the intuitive directional wrong of breaking a promise. When a promise is broken, intuitively the promisee is wronged in some way by the promisor. By the conventionalist account, however, the wrong of breaking a promise is not directed toward the promisee, but rather it is directed at all those who uphold the practice, and in the promisors 'free riding' on the convention of promising. This suggests that a conventionalist approach does not provide a robust enough account of promising. While the self interest conventionalist account does an adequate job in explaining why it is rational to generally uphold promises—and the general convention of promising—it does not account for the intuitive directed wrong involved in breaking a promise.

VII. Scanlonian Expectationalist Promissory Theory

By breaking a promise, one intuitively wrongs the promisee in some way. The conventionalist account does an inadequate job in accounting for this intuition. To the expectationalist, on the other hand, a promise is not based upon a convention, nor is it a specific speech act, but, rather, the expectationalist believes that it is based upon intention and assurance (Kolodny 2003, p. 127). The wrong in breaking a promise is not, as the conventionalist suggests, in the breaking of a coincidental convention, nor is it in the free riding upon other people's upholding such a convention, but, rather, the wrong in breaking a promise is done to the promisee by the promisor by his intentional misleading of the promisee to be assured of the promisor's performing some action 'X'. This notion is put forth by Scanlon in 'Promises and Practices' as Principle F:

> If (1) A voluntarily and intentionally leads B to expect that A will do X (unless B consents to A's not doing so); (2) A knows that B wants to be assured of this; (3) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (4) B knows that A has the beliefs and intentions just described; (5) A intends for B to know this, and knows that B does know it; and (6) B knows that A has this knowledge and intent; then, in the absence of special justification, A must do X unless B consents to Xs not being done (1990, p. 208).

This non-conventionalist approach allows for the absence of a social convention of promising, and makes it so that the mutual understanding of intentions and assuredness accounts for the keeping of, and the directed wrong in the breaking of, a promise. For example, lets say that Al makes Bert believe that Al will do some act 'X' according to principle F. If Al were to then *not* do act 'X', the wrong would be Al's intentional misleading of Bert to be assured of Al's performing act 'X', thus the wrong is done by the promisor directly to the promisee (Shockley 2008, p. 389). This directionality is, according to Scanlon, part and parcel to promising, and this essential part of promising is one that is not captured by a conventional account. By the conventionalist view, the wrong in breaking a promise is done to all those who uphold the social convention of promising—instead of directly to the promisee—and the wrong is done by the promisor's 'free riding' on other people's contributions to the practice (Kolodny 2003, p. 122). The wrong, by Scanlon's view, is not—as the conventionalist theory indicates—dependent upon mutual knowledge of a rule or convention, instead the wrong can be seen independently of a convention, via the promisor's establishing and breaking the trust of the promisee. This view fits well with our natural intuitions about promising in that it accounts for the wrong of breaking a promise being directed at the promisee.

VIII. Criticism of Expectationalism

It does not at first seem like intention and assurance account for, or equate to, the wrong of breaking an explicit promise. Assurance, it seems, is not equivalent to a promise, in fact, it does not seem wrong to say that assurance can be given to a very high degree without creating any obligation whatsoever. For example, Al wants Bob to meet him at the bar at 9pm. Bob tells Al that he will most likely be able to meet him there, however, there is a small chance that he might get called in for a night shift. Bob intends to go, and he makes it clear to Al that he intends to go, and let's just assume that the rest of the exchange meets all other conditions of principle F. Al goes to the bar at 9pm, and Bob is not there; at 8pm, Bob got called into work the night shift. Even though Bob provided a high degree of assurance to Al, and all parameters of Principle F obtained, it does not appear as though Bob has committed a wrong, or at least if there was a wrong committed, it is not equivalent to the wrong of breaking a promise. If, on the other hand, Bob were to explicitly *promise* Al that he will be at the bar at 9pm—regardless of whether he might get called into work or not—and Bob then doesn't show up, then Al could rebuke Bob for committing the wrong of breaking a promise. But in the prior example, where only assurance and intention were given, Al has no standing to rebuke or demand anything, and thus the wrong committed by violating principle F seems to be a weaker type of wrong than the wrong of breaking a promise.

This criticism is easily put to rest. When Bob made the promise by meeting principle F, he inserted another condition, which, if that condition obtained, would excuse his obligation to meet Al (Kolodny 2003, p. 149). In that same situation, if Bob were to have not met Al because he simply didn't want to—or some other reason that was not contained in the initial promise that met the conditions of principle F—then Bob would be committing the wrong of breaking a promise. So, when looked at as a special condition, the initial criticism fades in significance.

IX. Defense of Expectationalism

Scanlon argues that we do not need to have any prior knowledge of the practice of promising in order for a promise to obligate. This, especially to the conventionalist, is not clear. If promising is a convention with rules that are inherent to its successful operation, it does not appear as though someone without knowledge of the game could meaningfully participate.

In Scanlon's state of nature example, there are two fellows on either side of a river, Al and Bert. It just so happens that both Al and Bert have tossed their weapons on the opposite side of the riverbank. By Scanlon's view, despite the fact that these two blokes have no conventional common ground, they could—by principle F—establish a mutual promise where if one throws the other's weapon to the other side, then the other would be obligated to throw the corresponding weapon back to his side. Let us assume that promising is a convention on Al's side, and on Bert's side it has never been heard of. The conventionalist would deny that Al could engage in a promissory relationship with Bert. However, assuming the two sides could communicate linguistically, it seems intuitive to say that if Al were to say "if you throw me my spear then I will throw you yours," and Bert—understanding what Al has said—acknowledged with an "ok"—and for simplicity all other conditions of Principle F obtained—that a promissory relationship has, without a conventional common ground, been established. If, after meeting all conditions of principle F, Al were to throw Bert's spear back, and Bert then refused to throw the corresponding spear back to Al, it seems as though Bert would be doing a wrong towards Al. Since Bert understood what Al was saying, and Bert acknowledged that he understood, it seems apparent that

Bert would know that by not throwing Al's weapon back to him he would be doing a wrong towards Al. Bert's reasoning would likely be something along the lines of 'I know I said I would throw his spear back, but now I am not going to', but even in saying that, it seems as though he is *aware* of the obligation he entered into—along with the wrong involved in reneging—and he is simply deciding to renege regardless.

If we were to think of it as a game, it could even be the case that Al is conveying the conditions and rules of the game to Bert—via the 'I'll throw your spear back if you throw mine back' exchange—and that by doing so Al allows for Bert to meaningfully participate in the practice. When Bert reneges, it is not the case that Bert is ignorant of the rules. In other words, it is not the case that Bert is *unaware* that to renege is to wrong Al directly, but, rather, he is aware of the wrong that he is doing to Al in virtue of knowing the rules to the game, and, as a cheater will do whenever possible, he does it anyway. The expectationalist promissory theory could be used along side a conventionalist self-interest approach where, as is the case in the previous example, the communication of individuals could suffice in the establishing of a conventional common ground for promising (or at least the rules of the game), principle F could account for the directional wrong felt by breaking a promise towards one of the practitioners of that convention, and a self-interest conventionalist theory could serve as a second check on the players inclined toward cheating as a rational justification for keeping a promise.

X. Conclusion

The intuitive notion of promising cannot be fully accounted for using either conventionalist or expectationalist promissory theory. While the self-interest and rule-utilitarian conventionalist theory does well in justifying why it is rational to generally keep promises, and to uphold the practice of promising, they fail in accounting for the intuitive directed wrong we feel is associated with breaking a promise, and the aforementioned 'one shot' cases pose serious problems as well. The expectationalist promissory theory (à la Scanlon) does well in accounting for the intuitive wrong we feel is done towards the promisee when breaking a promise, and it seems as though the two can be used in tandem to account for both sides of the problem. Scanlon's Principle F can be used to establish a promise through intention and assurance, as well as account for the directional wrong felt in breaking a promise, and a utilitarian or self-interest conventional rationale can serve as a rational second justification on why one should generally keep promises and uphold the social institution of promising. Promising is a convention, yes, but is it a mere convention? No. Prior knowledge of the convention of promising does not seem necessary. It does seem, however, that communication and reason can suffice as a foundation for promises. Thus, intent and assurance, via the expectationalist view, are sufficient for the creation of a promissory relationship, and if one's intuitive notion that breaking a promise is wrong does not serve as sufficient motivation to keep promises, then the potential utility—self interested or otherwise—gained by having the social convention of promising should.

Bibliography

Amadae, Sonja. (2011) "Normativity and Instrumentalism in David Lewis' Convention," *History of European Ideas* 37, pp. 325-335

Habib, Allen. (2008) "Promises," in *Stanford Encyclopedia of Philosophy*, Retrieved from http://plato.stanford.edu/archives/win2008/entries/promises/

Hohfeld, Wesley (1964). *Fundamental Legal Conceptions as Applied in Judicial Reasoning* (New Haven: Yale University Press)

Kolodny, Niko; Wallace, R. (2003) "Promises and Practices Revisited," *Philosophy and Public Affairs* 31(2), pp. 119-154

Rawls, John. (1955) "Two Concepts of Rules," *Philosophical Review* 64, pp. 2-32

Scanlon, Thomas. (1990) "Promises and Practices," *Philosophy and Public Affairs* 19, pp. 199-226

Shockley, Kenneth. (2008) "On That Peculiar Practice of Promising," *Philosophical Studies* 140(3), pp. 385-399

MORAL RESPONSIBILITY, ALTERNATIVE POSSIBILITIES, AND HARD CASES

Jose L. Guzman Jr.

Generally, in situations where we attribute moral responsibility to an agent, we are inclined to believe that he or she could have done otherwise. For example, if an agent is under hypnotic control and he or she commits a crime under that hypnotic control, we are generally inclined to attribute the responsibility of that crime to the hypnotizer. Or if, for example, a neuroscientist manipulates an agent's brainwaves in a particular way that results in him or her committing a murder, we are generally inclined to attribute the responsibility of that murder to the neuroscientist. In fact, after such events occur, we sometimes say things, to the manipulated agent, like: "It's okay… There was nothing else you could have done." It is apparent, at least intuitively, then, that we tend to think that, if an agent committed an action and he or she could not have done otherwise, then he or she is not responsible for that action.

Harry Frankfurt calls the principle underlying this intuition "the principle of alternative possibilities," which I characterize as follows: if an agent is morally responsible for an action, then he or she could have done otherwise; or, if an agent could not have done otherwise, then he or she is morally non-responsible for his or her action (1969, p. 829). In that paper, Frankfurt proposed a counterexample to the principle of alternative possibilities (henceforth "PAP"). For PAP to be rendered false, Frankfurt had to show that there could be a case in which there is an agent who is morally responsible for an action, even though he or she could not have done otherwise. If we say, in the end, that Frankfurt was successful, then we must concede that moral responsibility does not require alternative possibilities. On the other hand, if we say that Frankfurt is unsuccessful, then one way to characterize our position is as PAP defenders. My interest in this paper is primarily in exploring the implications of the latter view. I contend that if we are PAP defenders, then we must answer to some problems with PAP.

I begin by putting forth my own brief characterization of Frankfurt's example. I say "brief," because in putting forth the characterization, I aim for nothing more than the reader's basic understanding of what is at issue. After having put forth that characterization, I present what I think to be an interesting and possible response to Frankfurt's example by Kadri Vihvelin. It is not my intention to explore whether Vihvelin successfully defeats Frankfurt here. I leave that matter unsettled. The portion of the paper following Vihvelin's response to Frankfurt is solely dedicated to exploring the implications of her conclusion, which I lead into by presenting some problems to PAP. Having presented these problems, I offer PAP defenders some solutions. Furthermore, I aim to show that as it stands, PAP does not account for all the cases we would normally think that it would, which leads me to formulate a new principle, namely PAP*.

Frankfurt aimed to render PAP false with the following case. Suppose Jones has decided that he is going to murder Smith. Suppose, further, that Jupiter is a neuroscientist who has the power to manipulate Jones' brainwaves in a particular way, such that Jones does whatever she wills. If, for instance, Jones is about to make a left turn while driving his car, then Jupiter can stop Jones with the use of her own volition. Jupiter is in agreement with Jones' decision to kill Smith. In other words, if Jones is going to change his mind about killing Smith, then Jupiter will intervene and manipulate Jones' brainwaves in a particular way, such that Jones does kill Smith. Conversely, if Jones is not going to change his mind about killing Smith, then Jupiter will not intervene.

Suppose that Jones goes on to kill Smith on his own, without Jupiter's intervention. In that case, Jones could not have done otherwise, because if he were about to do otherwise, then Jupiter would have intervened. Nevertheless, Jones is still responsible for killing Smith because he acted from his own will without Jupiter's intervention. We have a clear case, then, where there is an agent who is still responsible for an action, even though he could not have done otherwise, which renders PAP false.

At first glance, Frankfurt's argument seems to deliver a fatal blow to PAP. But Kadri Vihvelin (2008) argues that Frankfurt's example does not render PAP false. According to Vihvelin, although it *looks* as if Jones could not have done otherwise, he actually could have. In Frankfurt's example, if Jones is going to change his mind about killing Smith, then Jupiter intervenes, so as to make sure Jones kills Smith. If Jupiter *does* intervene, then she renders Jones unfree, since Jones cannot do other than what Jupiter has manipulated him to do. On the other hand, if Jones is not going to change his mind about killing Smith, then Jupiter does nothing, because that is exactly what she wants Jones to do. In that case, if Jupiter does not intervene, then Jones remains free, since in this case Jones could do otherwise, even though in the end he chooses not to. In other words, the mere fact that Jupiter is present does not render Jones unfree on its own. It is not until Jupiter intervenes, if she does, that Jones is rendered unfree. This is apparent in the fact that if, for example, Jupiter falls asleep, and Jones is going to change his mind about killing Smith, then Jones will successfully change his mind about killing Smith without any intervention. In Frankfurt's example, Jones goes on to kill Smith on his own without Jupiter's intervention, which thereby entails he did so freely and is morally responsible for doing so. Frankfurt's example does not render PAP false, then, because Jones could still have done otherwise.

Suppose Vihvelin is right and Jones had alternative possibilities.[1] PAP defenders seem to think that Jones' responsibility for having killed Smith is at least partly grounded in his having had alternative possibilities. But there is reason to think that one's responsibility for an action is sometimes not grounded in one's having had alternative possibilities. Suppose, for example, that Jones killed Smith, Jones is morally responsible for having killed Smith, and Jones had alternative possibilities to killing Smith. And suppose further that all Jones' alternative possibilities would have led to his being morally responsible for having killed Smith.[2]

According to PAP, Jones' being morally responsible for having killed Smith in this case is at least partly grounded in his having had alternative possibilities. But this seems to go against our intuitions. Intuitively, we want to say that Jones' being morally responsible for having killed Smith in this case is *not* grounded in his having had alternative possibilities. Indeed, Jones had alternative possibilities, but none of them would have exempted him from the responsibility he has, so they are irrelevant to his being responsible for having killed Smith (this is a similar argument to the *Irrelevance Argument* in McKenna 2008, p. 772). It seems logical to conclude, then, that PAP, as it stands, is too weak to ground moral responsibility.

In order for PAP to sufficiently ground moral responsibility, we are going to need to add something that accounts, not only for alternative possibilities, but also for alternative possibilities of a specific kind. One way to do this is to say that the relevant alternative possibilities necessary to ground moral responsibility are the preclusive kind (this is the route Pereboom takes for his account of robustness in 2009, p. 110 and 2012, p. 299). Call this version of PAP the *Preclusion Version*, which I characterize as follows: an agent is morally responsible for an action only if he or she had alternative possibilities that would have precluded his or her being morally responsible for that action.

The *Preclusion Version*, although it seems right and important, does not fully capture the scope of our intuitions regarding responsibility and action. Generally, when we attribute moral responsibility to an agent, we do seem to believe that he or she could have done otherwise in a way that satisfies the *Preclusion Version*. But what we do not seem to believe is that those alternative possibilities preclude an agent from bearing responsibility in whatever manner possible. Let us imagine the following scenario. Suppose Jones killed Smith. And suppose that Jones had alternative opportunities that would have precluded his being responsible for killing Smith. One of Jones' alternative possibilities that would have precluded his being responsible for killing Smith is that Jones could have gone out for a cup of coffee instead of killing Smith.

Surely, if Jones had gone out for a cup of coffee instead of killing Smith, that would have precluded his being held responsible for having killed Smith. Hence, Jones' having had alternative possibilities that would have precluded his bearing responsibility for killing Smith sufficiently grounds his being morally responsible for killing Smith.

This result seems right in theory, but it yields some odd results if we think about it in the context of our everyday lives. When we attribute moral responsibility to an agent, it does not seem to be the case that we think his or her being responsible for that action is grounded in his or her having had alternative possibilities, which included going out for a cup of coffee. Rather, when we attribute moral responsibility for an action to an agent, it seems to be the case that we think that he or she had alternative possibilities, which included refraining from performing that action. If, for example, Jim harmed Steve, and Jim is responsible for having harmed Steve, then my intuition is not that Jim had alternative possibilities that would have precluded his being responsible for harming Steve, but that Jim had alternative possibilities that included not harming Steve. In that case, Jim's having had alternative possibilities, which included not harming Steve, is the fact that sufficiently grounds his being responsible for harming Steve, not merely the fact that he had alternative possibilities that would have precluded his being responsible for harming Steve. Indeed, Jim could have gone out for coffee instead of harming Steve, which would have precluded his being responsible for harming Steve, but that fact seems irrelevant to the situation. It seems logical to conclude, then, that although the *Preclusion Version* captures something right and important, it ultimately gets things wrong.

Instead of attempting to amend the *Preclusion Version*, I propose that we do away with it altogether and introduce a replacement version which goes as follows: an agent is morally responsible for an action only if he or she had alternative possibilities which included refraining from committing the action he or she committed. Call this version of PAP the *Refrainment Version*.

But, like the *Preclusion Version*, there is reason to believe that the *Refrainment Version* does not do all the work we would like it to do either.

As it stands, the *Refrainment Version* (or versions similar to it) sometimes fail by rendering an agent morally non-responsible in cases where we are inclined to think otherwise (e.g., see Pereboom 2012, p. 299). We could easily think of a case, for example, where an agent gets drunk (but not blacked-out drunk) and does something he or she would normally think wrong. It is not a stretch to imagine the agent citing his or her drunkenness as the direct cause for his or her wrongdoing, suggesting that he or she did not have his or her normal possibilities of action available to him or her at that time. If what the agent says is true, and he or she did not have his or her normal possibilities of action available at that time, then is he or she morally responsible? If we invoke the *Refrainment Version* in this case, we would have to say that the agent is not morally responsible, since he or she did not have alternative possibilities that included refrainment. But this answer goes against our intuitions. Intuitively, we want to say that the agent is morally responsible for his or her wrongdoing, even though he or she was drunk at the time.

We are going to have to say something about what is meant when we say an agent "had alternative possibilities." If, for instance, a person commits a murder, and we establish that he or she had alternative possibilities to committing that murder, in what way did he or she have them? If the person who commits a murder is a psychopath, then from the outside looking in, it might look like he or she had alternative possibilities. But from the perspective of the psychopath, it is at least not clear that he or she had the particular type of alternative possibilities we normally take an agent to have. This leads us to think that there is an epistemic element at work in cases of moral responsibility (Pereboom 2012, p. 299).

In order for us to account for the drunkenness and psychopath cases, we are going to have to add something that accounts for the understanding of the agent, to the extent that he or she understood that he or she had particular alternative possibilities. But there is reason to think that "understood" is too strong (Pereboom 2012, p. 300), for the laymen recognizes a moral difference between killing and not killing his or her neighbor, but unless he or she is a moral philosopher, it is unexpected that he or she possesses the competence to adequately explain that difference. It is, then, better to say that the agent merely have had some "cognitive sensitivity" to his or her having had particular alternative possibilities (Pereboom 2012, p. 300). This alone handles the psychopath cases. The psychopath is cognitively insensitive to him or her having had particular alternative possibilities. This is enough to render the psychopath morally non-responsible for committing a murder.

But the drunkenness cases are a bit trickier. In the drunkenness cases, the drunks are in a way oblivious to their having alternative possibilities to actions they would have normally thought wrong at a particular time. If we have evidence that the drunks were cognitively sensitive at a previous time—in this case, at the time before they got drunk—then we could say that their having particular alternative possibilities were at least within the scope of their deliberative frame. If alternative possibilities of action are "within the scope of one's deliberative frame," then I take it to mean that there are possible actions that one is aware of, but not necessarily taken into consideration, at a given time. On the other hand, if an agent has a series of possibilities of action he or she is considering at a particular time, then I take those possibilities to be in his or her *active* deliberative frame, as opposed to being merely within its scope. In response to the drunkenness cases, we could say, for example, that the drunk had particular alternative possibilities within the scope of his or her deliberative frame, if we have evidence of him or her at a previous time at which he or she was cognitively sensitive to those possibilities of action. Call this version of PAP the *Deliberation Version*, which I characterize in the following way: an agent is morally responsible for an action only if he or she had alternative possibilities and was cognitively sensitive to these alternative possibilities, where "having alternative possibilities" means the alternative possibilities were, at least, within the scope of his or her deliberative frame (because any alternative in one's active deliberative frame is, necessarily, within the scope of the deliberative frame).

If we combine the *Refrainment Version* and the *Deliberation Version*, then we give rise to the following new version of the PAP.

> PAP*: an agent is morally responsible for an action only if he or she had alternative possibilities that included refrainment and was cognitively sensitive to his or her having had those alternative possibilities, where "having had alternative possibilities that included refrainment" means those alternative possibilities are, at least, within the scope of his or her deliberative frame.

This principle, I argue, gets things right in theory, in the context of our daily lives, and also in hard cases as well, like those of drunkenness and psychopaths. It involves looking at moral responsibility in a different light, which sufficiently grounds it in much more than the mere presence of alternative possibilities. I do not think PAP* is the ultimate solution to the problems of PAP defenders, nor do I think it is the best solution. But I do think that it is at least one plausible way to get around some of the aforementioned problems.

I began by arguing that as it stands, PAP has some problems, since sometimes one's responsibility for an action is not grounded in one's having had alternative possibilities. I then argued that in order for PAP to sufficiently ground moral responsibility, we have to make it stronger, namely by continually revising it and testing it in particular cases. In testing PAP in particular cases, I introduced the *Refrainment Version* and the *Deliberation Version* as alternatives in response to hard cases, like the drunkenness and psychopath cases. In the end, I articulated PAP* by combining the *Refrainment Version* and the *Deliberation Version*. PAP* not only gets things right in theory and in the context of our daily lives, but also in hard cases as well. In putting forth PAP*, I aimed for nothing more than to offer it to PAP defenders as a possible solution to some of their problems.

Notes

[1] It is worth noting that, in PAP, alternative possibilities are a necessary condition for moral responsibility, not a sufficient condition. In other words, if a person commits an action, and he or she had alternative possibilities to that action, then it would not necessarily be the case that he or she is morally responsible for that action. On the other hand, if a person commits an action, and he or she is responsible for that action, then, according to PAP, it would necessarily be the case that he or she had alternative possibilities.

[2] One might be prompted to ask the following question here. If an alternative possibility leads to the same outcome as the original action, then in what way is it an alternative? The answer, I think, lies in that there are at least two ways to think of the notion of "alternative possibilities." One way to think of the notion of alternative possibilities is to think of them as alternative actions to an original action, even in the event that they would have had the same outcome. We could think of a case, for example, where a person X killed a person Y. In this case, we could think of an alternative possibility, for example, where X coerces a person Z to kill Y. The second action, then, is an alternative action to the original action, even though it would have also led to X's being responsible for having killed Y. A second way to think of the notion of alternative possibilities is to think of them as alternative actions to an original action only if they would have led to a different outcome. Taking from the previous case, then, the second action would not be a legitimate alternative action, since it would have led to the same outcome—i.e., X's being responsible for having killed Y. In my mind, it is clear that it is the first of these that is the relevant sense, for we are interested now in agents, their actions, and the connection between those actions and the outcomes of those actions, not merely the outcomes themselves.

Bibliography

Fischer, John M. (2008) "Freedom, foreknowledge, and Frankfurt: a reply to Vihvelin," *Canadian Journal of Philosophy* 38, pp. 327-342

Frankfurt, Harry G. (1969) "Alternative possibilities and moral responsibility," *The Journal of Philosophy* 66, pp. 829-839

Levy, N.; McKenna, M. (2009) "Recent work on free will and moral responsibility," *Philosophy Compass* 4, pp. 96-133

McKenna, Michael. (2008) "Frankfurt's argument against alternative possibilities: looking beyond the examples," *Noûs* 42, pp. 270-793

Pereboom, Derk. (2012) “Frankfurt examples, derivative responsibility, and the timing objection,” *Philosophical Issues* 22, pp. 298-315

_____. (2009) “Further thoughts about a Frankfurt-style argument,” *Philosophical Explorations* 12, pp. 109-118

Sher, George. (2009) *Who Knew? Responsibility Without Awareness* (New York: Oxford University Press)

Vihvelin, Kadri. (2012) “Foreknowledge, Frankfurt, and ability to do otherwise: a reply to Fischer,” *Canadian Journal of Philosophy* 38, pp. 343-372

The Gap Between “Is” and “Ought”

Jianli Wang

The Is-Ought Problem

Bob is a student, and one day his friend John asks him whether he wants to play video games that night. “I probably shouldn’t,” Bob says. “I ought to study tonight because there is going to be a final exam tomorrow.”

This response seems rational. But John is a philosophy student, and he notices that there is a problem in Bob’s reasoning. Bob’s argument goes roughly like this:

(a) If there is an exam tomorrow, then Bob ought to study tonight.

(b) There is an exam tomorrow.

(c) Therefore Bob ought to study tonight.

This argument is valid, so now we need to prove (a) and (b). Suppose that (b) is a matter of fact, thus the question would be: can we derive the claim that “Bob ought to study tonight” from “there is an exam tomorrow?” In this case, the sentence “There is going to be a final exam tomorrow” is a descriptive statement (about what is) while “I should study tonight” is a normative claim (about what ought to be). But there is a significant difference between descriptive statements and normative statements, and some philosophers believe that it is difficult or even impossible to derive normative statements from descriptive statements. Therefore, it would be impossible to derive the claim that “Bob ought to study tonight” from “there is an exam tomorrow” and the premise (a) is false.

This problem is usually called the is-ought problem, since it is about the gap between the claims about “what ought to be” and the statements about "what is." David Hume, in *A Treatise of Human Nature*, first discussed this problem:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary ways of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when all of a sudden I am surprised to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it should be observed and explained; and at the same time that a reason should be given; for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. But as authors do not commonly use this precaution, I shall presume to recommend it to the readers; and am persuaded, that this small attention would subvert all the vulgar systems of morality, and let us see, that the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceived by reason (Hume 1739, pp. 244-245).

In his paper, Hume pointed out that many authors make claims about "what ought to be" based on statements about "what is," but there is a significant difference between these two kinds of statements, and it is not obvious how we can derive "what ought to be" from "what is."

G. E. Moore discusses a similar problem in *Principia Ethica*. According to Moore, the moral "good" is simple and unanalyzable, so we cannot define "good." He argues that it would be fallacious to explain that which is good reductively, in terms of natural properties such as "pleasant" or "desirable" (Moore 1903, pp. 1-36). This problem is close to the is-ought problem, since in many cases the things that we ought to do are the morally good things. So if we can define "good" in terms of natural properties, then it would not be difficult to derive what we out to do from natural properties. Thus, the is-ought problem will be solved, and vice versa.

In contrast, naturalists are the philosophers who believe that we can explain moral "good" in terms of natural properties, and try to solve the is-ought problem by proving that we can derive "ought" from "is" (Searle 1964, pp. 43-58). John Searle, for example, tries to show that making a promise places one under an obligation by definition, and this obligation involves an "ought." For example, if John promised Bob that he would play video games with Bob tonight, then that promise placed John under an obligation of playing video games with Bob, because that's what the words "promise" and "obligation" mean. If John is under an obligation of playing video games with Bob tonight, then John ought to play video games with Bob tonight because the notion of obligation involves an "ought."

So there are mainly three different views concerning the is-ought problem. Hume holds the first one, and it is the view that since we cannot derive an "ought" from what is, there is no such thing as a true "ought" claim. The second one is held by G. E. Moore—that true "ought" claims exist, but they cannot be derived from what is. The last one is naturalism—that we can derive "ought" claims from what is. In this paper I will examine a naturalistic view and make an attempt to derive "ought" claims from a goal of the action and desire of the goal.

Non-moral "Ought"

A common naturalistic solution to the is-ought problem is that we can derive "ought" claims from goal-directed behavior. On this view, ought claims can be derived in a way like the following:

> In order for agent X to achieve goal Y, X reasonably ought to do Z.

For example, it seems true to say that in order to win a race, one ought to run faster than other players. But the problem with this solution is that the "ought" derived here is a non-moral "ought." Thus, although we may be able to derive an "ought" in this way, we still need to figure out how to derive a moral "ought." So there are two different kinds of "ought" claims, and we can define ought$_1$ as the non-moral, goal-directed "ought," and ought$_2$ as the moral "ought." To make the problem easier, we can try to determine the nature of ought$_1$ first. Consider the following statements:

(1) If Bob doesn't study tonight, Bob will not get a good grade on tomorrow's exam.

(2) Bob ought$_1$ to study tonight if Bob wants to get a good grade on tomorrow's exam.

(3) Bob wants to get a good grade on tomorrow's exam.

(4) Therefore, Bob ought$_1$ to study tonight.

We can derive (4) from (2) and (3), so the problem is whether we can derive (2) from (1). Logically, sentence (1) is equivalent to the following sentence:

(1a) Bob will get a good grade on tomorrow's exam only if Bob studies tonight.

How is (1a) related to (2)? If event X can only happen if event Y happens, then, other things being equal, in order to make event X happen, one should make event Y happen. Of course, it is possible for all sorts of things to happen which would make sentence (2) false when sentence (1) is true. For example, suppose Bob didn't sleep very well last night so he is tired, and consider the following sentences:

(1*) If Bob stays up late tonight, he will be very tired.

(2*) Bob ought$_1$ to go to bed early tonight if he doesn't want to be very tired.

(3*) Bob doesn't want to be very tired.

(4*) Bob ought$_1$ to go to bed early tonight.

It is possible that statement (1)–(3) and (1*)–(3*) are all true, but it seems that (4) and (4*) cannot both be true. If Bob ought$_1$ to study tonight, then he ought$_1$ not go to bed early tonight, and vice versa.

There are two ways to solve this problem. The first one is to deny that statement (4) is in conflict with statement (4*). So, if Bob wants a good grade on tomorrow's exam and doesn't want to get very tired, then Bob ought$_1$ to go to bed early tonight and he also ought$_1$ to study tonight. On this view, both of these "ought" claims can exist at the same time, just as the two claims about Bob's desires can. In this case, Bob is confronted with a dilemma, and he needs to choose which action to take, or figure out which thing he wants more. If he chooses to go to bed early, it doesn't necessarily mean that he ought$_1$ not study tonight. By this view, many ought$_1$ claims can exist at the same time, and a non-moral ought is not necessarily motivational. Although Bob can only do one thing at a time, he ought$_1$ to do many things at the same time.

Another solution is to add a *ceteris paribus* clause. That means to get an entailment between (1) and (2), we need a qualifying statement like the following:

(1b) Other things are equal.

This qualifying statement combined with (2) and (4) would result in:

(2a) If Bob wants to get a good grade on tomorrow's exam, then all else being equal, Bob ought$_1$ to study tonight.

(4a) Therefore, other things being equal, Bob ought$_1$ to study tonight.

The *ceteris paribus* clause in sentence (2a) can rule out the possibility of other factors that could override the relationship between the antecedent and the consequent, so in this case, it rules out the situation that sentences (1*), (2*) and (3*) are all true. Or we can say that sentence (4a) no longer contradicts with sentence (4*). Therefore, the problem is solved. This solution will make the claim of the argument weaker because with the ceteris paribus clause, the argument is just making a general claim that may not be true in a specific situation. For example, if (1*), (2*), and (3*) are true, and Bob wants his health more than a good grade on tomorrow's exam, then, we will not be able to derive the sentence that "Bob ought$_1$ to study tonight" from sentences (1), (2) and (3).

Both of these two solutions can solve the problem. But the first one may have a problem. Although we are discussing "ought$_1$" here, my goal in this paper is to derive an "ought$_2$." If we endorse the first solution to derive moral "ought$_2$," that means we endorse the view that more than one ought$_2$ claims can be true at the same time. In that case, how to choose between them would be a problem, since one can only do one thing at a time. If we cannot provide a good way to choose one ought$_2$ claim over another, or there is no good reason for choosing one over another, we are facing a moral dilemma. Therefore, if we choose the first solution, we care committed to saying that moral dilemmas are real, which can be controversial. Ethicists as diverse as Kant (1768), Mill (1843), and Ross (1939) have assumed that an adequate moral theory should not allow for the possibility of genuine moral dilemmas. There are also philosophers who challenge that assumption by arguing that it is not possible to preclude genuine moral dilemmas, or that it is not desirable to do so. While the first solution is introducing a controversial theory, the second solution is making the conclusion of the argument weaker, by adding the *ceteris paribus* clause. Since I am not going to discuss the problem of moral dilemmas in this paper, I will endorse the second solution.

Therefore, the new argument will be:

(1a) Bob will get a good grade on tomorrow's exam, only if Bob studies tonight.

(2a) Other things being equal, Bob ought$_1$ to study tonight if Bob wants to get a good grade on tomorrow's exam.

(3) Bob wants to get a good grade on tomorrow's exam.

(4a) Therefore, other things being equal, Bob ought$_1$ to study tonight.

Sentence (1a) makes a claim about an empirical fact concerning a causal relation between two events; this causal relation is objective, and as such it is derived from logical reasoning. It has the form:

(1') X, only if Y.

Statement (2a) is a crucial part of the argument, and it has the form that:

(2') Other things being equal, a person P ought to do Y if P wants X.

Here, we may ask that whether X and Y can be identical. If X and Y are identical, then (1') is necessarily true because it will have the form:

(1") X, only if X.

And (2") will be like:

(2") Other things being equal, a person P ought$_1$ to do X if P wants X.

If (1") is true, then (2") should be true, since we derive (2') from (1') and (1")–(2") are just (1')–(2') when X and Y are identical. It may seem awkward to say that if a person P wants to do X then P ought$_1$ to do X, since it feels that there is something more in the claim that "what I ought to do" than the claim that "what I want to do".

To solve that problem, we should keep in mind that "what I ought to do" does mean something more than "what I want to do," but it will reduce to "what I want to do" in some cases. There is nothing wrong in saying, "If you want to do X, then do X." It seems people say that quite often. Furthermore, we should notice the *ceteris paribus* clause in (2"). If we rule out all other factors, and the only thing P wants to do X, then it is plausible to say that P ought$_1$ to do X. (2") seems awkward because it is redundant, just like (1"). But a sentence can be redundant and still be true.

Sentence (3) is about the desire, or we can say, the will, of a person. It has the form:

(3') A person P wants X.

This "ought$_1$" is derived from reasoning and desire. If we look at sentence (1a), or any other if-then sentences, we will notice that we can only get this kind of claim by reasoning. An if-then relation is a logical relation, so we cannot see or feel an if-then relation between two events; we can derive an if-then sentence deductively from other if-then sentences, or inductively from empirical facts. For example, we can inductively derive sentence (1a) from the fact that every time Bob didn't study the night before an exam, he got a bad grade. Or we can derive sentence (1a) deductively from the sentences such as "if Bob doesn't study tonight, he will not be able to answer the questions of the exam tomorrow," and "if Bob is not able to answer the questions of the exam tomorrow, he will not get a good grade." Since sentence (2a) is derived from (1a), and sentence (3) is a claim about a desire of Bob, which can be considered as a mental fact, we can see that claims about what ought$_1$ to be can be derived from the statements about what is.

From Non-moral "Ought" to Moral "Ought"

Now that we have discussed the non-moral "ought$_1$" and how to derive the claims about what ought$_1$ to be from statements about what is, it is time to move on to the moral "ought$_2$."

Since we already have the formula to derive ought$_1$ claims, we can try to derive ought$_2$ claims in a similar way. If we examine the derivation of the ought$_1$ claim that we discussed above, we will find that there are three main factors of the derivation: the action, the goal, and the desire. What a person X ought to do is an action Z, and this action is directed by a goal Y, which means there is a conditional relation between Y and the Z, such that to achieve Y, X must take the action Z. Since X desires Y, X ought$_1$ to do Z. If we assume that the derivation of the ought$_2$ claim has the same structure, then we just need to find out the three factors of the derivation. Since the action is already included in the ought$_2$ claim, the problem is to find out goal of the action and desire of the goal.

First of all, we need make it clear that to derive an ought$_1$ claim, we just need to find *a* goal of the action, not *the* goal of the action. For example, having a good grade is *a* goal of studying tonight, but not *the* goal of studying tonight. Having more knowledge can also be a goal of studying tonight, and fulfilling curiosity can be another goal of the same action. Therefore, an action may have more than one goal, and one can take more than one actions to achieve one goal, since an action can have conditional relations with multiple goals, and vice versa. So, if we assume that the derivation of ought$_2$ is similar to the derivation of ought$_1$, then what we are looking for is a goal of being moral.

Is there a goal of being moral? At least there can be a goal of being moral in some cases. For example, we can take producing a good consequence, as a goal of being moral. That means, in such cases, one must take moral actions in order to producing the greatest good, because there is a conditional relation between producing a good consequence and a moral action. For instance, we can assume that "other things being equal, a good consequence will be produced only if Bob doesn't steal John's car," and in that case, producing a good consequence is a goal of Bob not stealing John's car.

This is similar to the view of Utilitarianism, which is the view that the proper course of action is the one that maximizes utility, or happiness. But the difference is that the Utilitarian believes that producing the greatest utility is the only goal of all moral actions, while here in the derivation we just assume that producing a good consequence is one goal of some moral actions.

There are some problems about Utilitarianism or Consequentialism. One problem is that the notion of good consequence is vague, and it can be difficult to measure which action produces the maximum utility, because almost everything that happens after an action can be considered as a consequence of the action on some level. But this problem would not be a problem for us. Since we will need the third factor—desire of the goal, for our derivation of ought$_2$ claims, we can define a good consequence as the one that is desirable.

Now we need to consider the third factor—desire of the goal. Let's consider the following statements:

(5) Well-being will be maximized only if Bob doesn't steal John's car

(6) Other things being equal, to maximize well-being, Bob ought not to steal John's car.

(7) Bob wants to maximize well-being.

(8) Therefore, Bob ought$_2$ not to steal John's car.

Here, we treat "not taking an action Z," as also an action, and we can define this action as "~Z." To make it simple, we assume that the actions of the kind "~Z," have the same nature of other actions.

There is an obvious problem in this derivation. In this case, what Bob wants is irrelevant to the ought$_2$ claim, because even if Bob doesn't care about maximizing well-being, he still ought$_2$ not to steal John's car. Can we just get rid of the desire of the goal, and try to derive the ought$_2$ claim directly from the goal of producing maximum well-being? In that case, the derivation would be:

(5) Well-being will be maximized only if Bob doesn't steal John's car

(6) Other things being equal, to maximize well-being, Bob ought not to steal John's car.

(7*) Well-being should be maximized.

(8) Therefore, Bob ought$_2$ not to steal John's car.

This argument is valid, but it cannot solve the is-ought problem, because sentence (7') is already a normative claim, and we just derived a normative claim from another normative claim. Therefore, it should be clear that the desire of the goal is important in the derivation and it cannot be replaced. Since what Bob wants is irrelevant to what Bob ought$_2$ to do, we need to find other people's desire that is relevant to what Bob ought$_2$ to do.

One possible answer is that in this case, we derive the "ought$_2$" claim from John's desire. So we have:

(5a) John's well-being will be maximized only if Bob doesn't steal John's car.

(6a) Other things being equal, to maximize John's well-being, Bob ought not to steal John's car.

(7a) John wants to maximize his well-being.

(8) Therefore, Bob ought$_2$ not to steal John's car.

But there are problems with this answer. First of all, a counterexample would be that if John were a terrorist, and he is using his car as a bomb to kill some civilians, then even if John wants to maximize his well-being, we should say that Bob ought$_2$ to steal John's car. So we have:

(9) Civilians' well-being will be maximized only if Bob steals John's car.

(10) Other things being equal, Bob ought$_2$ to steal John's car if civilians want to maximize their well-being.

(11) Civilians want to maximize their well-being.

(12) Therefore, Bob ought$_2$ to steal John's car.

Sentence (8) and (12) cannot both be true. This problem can be considered as caused by the conflict between John's desire and some other civilian's desire, since sentence (8) is derived from John's desire, and sentence (12) is derived from the other civilian's desires. In this case, it seems as though there is no moral dilemma, because if John is a terrorist and he tries to use his car as a bomb, then it is plausible to say that Bob ought$_2$ to stop him.

So when sentences (9)–(11) are true, sentence (9) stops being true, and that means in this case, the civilians' well-being is more important than John's well-being, or the desire of the civilian's is stronger than the desire of John's.

There is also another problem about this solution. Consider the following sentences:

(5b) Bob's well-being will be maximized if John let Bob steals his car.

(6b) Other things being equal, John ought$_2$ to let Bob steals his car if Bob wants to maximize his well-being.

(7b) Bob wants to maximize his well-being.

(8b) Therefore, John ought$_2$ to let Bob steal his car.

Sentences (5b)-(8b) have the same structure as the sentences (5a)-(8), but it seems wrong to say that John ought$_2$ to let Bob steal his car.

Since there are many serious problems with these two solutions, we may not be able to derive ought$_2$ claims directly from what is. When we say "Bob ought$_2$ not to steal John's car," or "John ought$_2$ not to steal Bob's car," it can be considered that we are claiming that "we ought not steal other people's car," or "stealing is something that we ought not to do." When we make ought$_2$ claims, are actually making moral judgments about actions. To say a person X ought$_2$ to do Z is to say that action Z is morally good. To say that action Z is morally good, is to say that we ought$_2$ to do Z. In most cases, "what I ought$_2$ to do" can be derived from "what we ought$_2$ to do." One difference between ought$_1$ and ought$_2$ is that what ought$_2$ to be is something objective, and it is not affected by who takes the action, or what the person wants. Here, "we" is used to refer to people in general, or human society. Therefore, if we can derive "what we ought$_2$ to do" from "what is," we can solve the is-ought problem.

Consider the following sentences:

(5c) Our overall well-being can be maximized only if we do not steal from each other.

(6c) Other things being equal, we ought$_2$ not to steal from each other if we want to maximize our overall well-being.

(7c) We want to maximize our overall well-being.

(8c) So, we ought$_2$ not to steal from each other.

(9) Therefore, Bob ought$_2$ not to steal John's car.

There are also problems with this solution. One problem is that we cannot measure our overall well-being, because there is no way to calculate all the people's well being. One reply to this problem can be that we can roughly calculate the overall well-being, and sometimes it is not difficult to compare one action with another. For example, it is obvious that other things being equal, the overall well-being of a society without genocide is greater than a society in which genocides take places. Also, since it is not easy to calculate the overall well-being of all human, we don't have many moral rules, it is also the reason that we made many mistakes in moral judgments, especially in history.

Another problem is that sentence (7c) is not obvious, because it seems that not many people actually care about the overall well-being of all humans. But one can reply that even if it seems that most people as individuals do not care about the overall well-being of all human beings, we as a unity do care about the well-being of all humans. Also, a person's well-being is always connected with the overall-well being of human society. For example, the overall well-being today should be greater than the overall well-being 3000 years ago, and the well-being of most individuals is greater than the well-being of a person from 1000 BC.

Conclusion

Now we finally derive the "ought$_2$" from goal of the action and desire of the goal:

(5c) Our overall well-being can be maximized only if we do not steal from each other.

(6c) Other things being equal, we ought_2 not to steal from each other if we want to maximize our overall well-being.

(7c) We want to maximize our overall well-being.

(8c) So, we ought_2 not to steal from each other.

(9) Therefore, Bob ought_2 not to steal John's car.

Sentence (5c) can be considered as an empirical fact, and sentence (7c) is a mental fact about all human beings. David Hume believes that morality is not an object of reason but an object of passions (1751, pp. 197-212). But I think that although we cannot find morality in reasoning, we can find reasoning in morality. At some level, we can reason about morality, because in many cases, it is the combination of reason and desire.

In this paper, I tried to derive the moral "ought" in a different way from how most naturalists did. The view I've sketched in this paper is still very rough, and it may not be as clear as it should be, but I think it can give us a new way to solve the is-ought problem.

Bibliography

Hume, David. (1751) *An Enquiry Concerning the Principles of Morals* (London: A. Millar)

_____. (1739) "Moral Distinctions not derived from Reason," in: *A Treatise of Human Nature*, pp. 244-245 (Oxford: Clarendon Press)

Kant, Immanuel. (1768) *The Metaphysics of Morals* (Philadelphia: University of Pennsylvania Press)

Mill, John Stuart. (1843) *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and The Methods of Scientific Investigation* (New York: Harper & Brothers)

Moore, G.E. (1903) *Principia Ethica* (Cambridge: Cambridge University Press)

Ross, W.D. (1939) *The Foundation of Ethics* (Oxford: Clarendon Press)

Searle, John. (1964) "How to Derive 'Ought' From 'Is'," *Philosophical Review* 73(1), pp. 43-58

The Egotistical Reason to be Moral and the Problem of Mere Appearance

Samuel Chen

Discussions about what is the right or wrong thing to do have demanded the attention of academic philosophers since the inception of philosophy. More so, debates about morality have long been a familiar mode of discourse for even those wholly unfamiliar with philosophy (such as an analysis of a politician's moral character). This is understandable, given the incredible importance ethics commands in our lives. However, while talk of whether abortion is morally permissible is abundant, I believe a deeper meta-ethical question is often forgotten in recent moral discourse: why be moral? This seemingly simple question serves as a foundation for any normative ethical theory—after all, if the answer to the question is that there's no good reason to be moral, then why care about what's morally permissible or impermissible? Why care how to live the good life?

Of course, I'm not the first to raise this question, nor am I the first to attempt to solve it. The why-be-moral question (WBM) has been a fundamental question in moral philosophy since its birth—the famous myth of the Ring of Gyges is a common introduction. However, it hasn't received nearly as much limelight in the past century as the questions concerned with how to live a morally virtuous life (and more often than not, discussions about what actions are consistent with the virtuous life). Regardless, my purpose isn't to whine about the unpopularity of WBM, but to attempt to answer it. Above all else, a proper solution to this problem must first be preceded with the sufficient conditions for solving it. That is, we must know what the problem is, and exactly what must make up a satisfactory answer. Thus, I will first detail what kind of problem WBM is, and argue that it is essentially a question of psychological motivation concerning prudential reasons. I will then visit three prominent historical philosophers, Plato, Thomas Hobbes, and David Hume, to highlight their proposed solutions that we will still see argued for in modern times. Finally, by considering further insights from business ethics and Neo-Confucianism, I will argue that while the solution is likely to be found by relying on prudential reasons, there still has yet to be a satisfying answer to why I should be moral due to the potency of an ancient loophole.

WBM is essentially a meta-ethical question; it's an attempt to understand the nature of ethical statements, properties, etc. As stated before, this underpins any normative ethical theory. Whether one is a virtue ethicist, consequentialist, deontologist, etc., isn't of direct relevance to WBM as they are meant to prescribe what one ought to consider when acting morally. The WBM-skeptic isn't concerned with what actions are moral, such are only secondary concerns. What he is asking is why he should even embark on a virtuous life. And by virtuous life I don't mean something grandiose akin to the pure life of a monk, I simply mean the attempt to live as morally virtuous of a life as reasonably possible. This goal shouldn't be unfamiliar with anyone—most folk want to be thought of as "good people," whether in the eyes of others or not. Granted, almost everyone has a different conception of what it means to be "good people," but the fundamental desire to want to be moral is shared by many people in this world. WBM concerns that very desire. Furthermore, it should be emphasized that the type of cause for moral motivation ought to be a justificatory reason, not an explanatory reason. The explanations for why many people want to be, and are, moral are plentiful and relatively uncontroversial: given that humans are social animals, many people are conditioned to be moral, to like the moral, and to pursue moral courses of action in the future. This is not what I'm concerned with; I'm not concerned with the empirical causes of moral behavior. Rather, WBM is asking for a justificatory reason, or in other words, what good reasons are there for being moral?

If this is the case, if WBM is outside the purview of normative ethics, then how are we to answer the problem? There are two primary potential ways of figuring this out: moral reasons and prudential (i.e., self-interested) reasons. I'll first analyze the moral-reasons strategy. To talk about moral reasons is to come to an important juncture in the history of WBM. As stated before, I observed the fact that WBM pales in popularity to normative ethics. A large reason for this is the large amount of dismissive replies to WBM—dismissive in regards to the legitimacy of WBM as a philosophical question. Many philosophers, some notably being F. H. Bradley, H. A. Prichard, John Hospers and J.C. Thornton, accuse this problem of not being a real one at all. They have said that the question is essentially logically impossible, an obvious truth, or incoherent (or often some mix of the three). These spurious responses are plentiful, and consequently many regard WBM as a "pseudo-question."

One popular tactic is to rephrase the question to the following: "why should I do what I should?" In that light, it seems uncontroversial that the question is flaccid and redundant. But this simply confuses the various meanings of 'should.' Moral talk isn't the only way of using the word, and in fact it often isn't. When I recommend that you "should wear that blue shirt with those jeans," this is a recommendation not on the basis of ethical principles, but of one built on aesthetic criteria relevant to the current culture. If I'm asking a bystander whether I should take a right on this street or the next, this isn't a question of ethics but one of maximizing travelling efficiency. Thus, accusations of circularity are empty, for WBM is phrased as "what nonmoral reasons do I have for acting morally?" This is a clearly meaningful sentence, and thus in no sense a pseudo-question.

Another strategy is to regard it as incoherent, as they would phrase the question as the following: "why is it in my interest to not be moral when it is not in my interest not to be moral?" This rephrasing grants the admission of the previous criticisms (that 'should' has multiple meanings), and charges the question while asking for the impossible. If someone is demanding a reason to do

X so long as it pays, in reference to something that does not pay (morality), then this seems logically contradictory. The problem with this strategy is that it confuses the point of view of the person asking the question. He isn't asking for a reason to conform to his self-interest, but is rather asking what reason he has for choosing a self-interested perspective or a moral perspective. Another way of conceiving of this is realizing that someone who knows what is right and wrong can still ostensibly ask himself the WBM question in hopes of figuring out a non-circular reason for it.

> [E]ither I act from the moral point of view, where logically speaking I must try to do what is right, or I act from the point of view of rational self-interest, where I must seek to act according to my rational self-interest. But is there any reason for me always to act from one point of view rather than another when I am a member in good standing in a moral community? (Nielsen 1989, p. 181).

Given that at the very least WBM is a legitimate question, and that moral reasons are wholly inadequate, prudential reasons remain the only probable explanation. The myth of the Ring of Gyges is one of the foremost and earliest formations of the WBM question. Mentioned by Plato in book 2 of *The Republic*, the interlocutor Glaucon, who happens to be the older brother of Plato, relates the story of Gyges of Lydia, who was a shepherd that stumbled upon a ring that granted its owner the power of invisibility. With this newfound power, Gyges goes on to seduce the queen and murder the king, and thus relishes in wealth and poverty. Glaucon asks us whether we would be moral if we also had this power of invisibility, which conceptually is meant to be the ability to avoid punishment and blame for any moral transgression we commit (Plato 2004, p. 36). If we could become immune from not only the law, but from anyone knowing of our wrongdoings, why would we act morally all the time?

> Suppose now that there were two such magic rings, and the just put on one of them and the unjust the other; no man

> can be imagined to be of such an iron nature that he would stand fast in justice. No man would keep his hands off what was not his own when he could safely take what he liked out of the market, or go into houses and lie with any one at his pleasure, or kill or release from prison whom he would, and in all respects be like a god among men. Then the actions of the just would be as the actions of the unjust; they would both come at last to the same point. And this we may truly affirm to be a great proof that a man is just, not willingly or because he thinks that justice is any good to him individually, but of necessity, for wherever anyone thinks that he can safely be unjust, there he is unjust. For all men believe in their hearts that injustice is far more profitable to the individual than justice, and he who argues as I have been supposing, will say that they are right. If you could imagine any one obtaining this power of becoming invisible, and never doing any wrong or touching what was another's, he would be thought by the lookers-on to be a most wretched idiot, although they would praise him to one another's faces, and keep up appearances with one another from a fear that they too might suffer injustice (Plato 2004, pp. 38-39).

In hindsight, Glaucon's accusation is too harsh; he argues that we would never do just things if *any* opportunity presented itself to do evil and escape unharmed. He paints an unrealistically dark portrayal of humans as preying upon any opportunity that the ring of invisibility/injustice would grant. However, his core point that morality is tied with self-interest is crucial, for it seems apparent that there would be rational reasons to do some immoral things with the invisibility ring. It is important to note that the WBM question, as informed by Glaucon's myth, is asking "why should *I* be moral?" not "why should *we* be moral?" The latter question has been dealt with by the likes of social contractarians, such as Thomas Hobbes' discussion of the state of nature. To act immorally against each other would bring about great risk to one's own self-interests—we would be in a perpetual state of war with each other. The slimmer version of WBM, the one that asks why *I* should be moral, invites the possibility that one could act immorally and still benefit prudentially (i.e., acting immorally without being detected). The difference is that while having morality as a collective group is justified on the basis of protecting ourselves from the viciousness of an anarchic state of war, and thus preserving our self-interest, having morality instilled on an individual basis is not necessary. This is because the free rider problem poses potential issues, in which people could simulate moral behavior for the most part, but also commit immoral acts when the opportunity arises for escaping capture.

Plato's argument for why we should be moral becomes a precedent that many others, such as Hobbes, have reiterated. He basically says it "pays" to be moral. We will benefit prudentially if we were to adopt the just life, and thus this is the reason we should live the just life. He conceives of humans having a tripartite soul: the appetitive, the spirited, and the rational part. Maintaining and balancing these three parts is crucial for individual justice: what is important is which part of the soul controls the soul, in relation to pleasure. When the appetitive part controls the soul, there is the pleasure of profit. When the spirited part controls the soul, there is the pleasure of honor. When the rational part controls the soul, there is the knowledge of reality/pleasures of necessity. Plato considers the rationally ruled man as enjoying true pleasure. And as such, even if we were to evade detection for injustice, Plato argues, "the one who remains undetected becomes even worse, while in the one who is discovered and punished, the bestial element is calmed and tamed and the gentle one freed" (2004, p. 295).

There are some obvious, glaring problems with Plato's tripartite theory of the soul (and what it entails) that render it an inadequate solution. To be charitable, we can come to understand his theory of the soul with purely physicalist concepts, as the philosophical baggage of supernaturalism is much too heavy and unfounded. Regardless, even if we were to reduce the soul to psychology, Plato's argument about the corruption of one's character is empirically outdated, due in many ways to being based on an incomplete view of human nature. The mind is much more complex than Plato's conception, and thus he fails to account for the repeated instances in which people can commit injustice and yet be happy. Even by taking the most extreme examples of corrupt leaders in power, we can see that their disposition has not only led them to great happiness, but seemingly at a minimal cost of "wretchedness." There are a myriad of psychological explanations to which I could refer, but it is an obvious, folk-psychological fact that humans are capable of justifying their actions regardless of justice. Given the potential for human depravity, Plato's theory of the soul fails on basic scientific grounds. Furthermore, Plato's theory of the soul is too restrictive when it comes to achieving a harmonious soul. It demands the attitude and willingness to engage in philosophical inquiry, which Plato notes as being a rare feat for men to achieve. In fact, to be a philosopher partially demands that one also be well born (Plato 2004, p. 190). The problem with this is that WBM is a question that applies to anyone who is motivated to ask such a question, and ostensibly this can be a legitimate question even for the non-philosopher. If Plato's theory demands that only the philosophers can truly solve the problem, then it comes with a heavy price.

Thomas Hobbes raises the question again, this time in reference to the irresponsible fool. As mentioned before, Hobbes is part of the social contractarian tradition that provides a compelling answer to the question of "why should *we* be moral?" He convincingly paints a picture of the state of nature in which everyone is at risk, because of one another, and thus life in general is much more perilous and unhappy. By joining in pacts and agreements, and creating laws and systems of government, we are able to maximize everyone's self-interest by relinquishing some of our freedoms. The irresponsible fool questions why he should always uphold his contracts (i.e., be moral) when there arises opportunities in which injustice pays. Hobbes argues that the irresponsible fool is unreasonable, for the success of the fool's goals (to stealthily benefit from injustice) will inevitably lead to failure, due to his overestimation of his own abilities and underestimation of others' abilities to detect deception; their hopes are based on "a false presumption of their own wisdom" (1998, p. 311).

Hobbes' argument is quite weak, as it depends on a strong claim about human psychology. He argues that any devious attempt by the irresponsible fool will inevitably backfire on him, due to the weakness of his skills. It's an obvious statistical fact that crimes go unpunished and some criminals never get caught. His point about the irresponsible fool failing is important however, as it homes in on an important centerpiece of many prudentially oriented solutions to WBM. Often stated, the problem arises by there being unjust opportunities arising, that one could benefit from safely; but, responders would want to say that the mentality of engaging in these opportunities will eventually be their downfall. For Hobbes, it was the mental constitution of the irresponsible fool that was too weak, but his conclusion was too strong for the empirical evidence to support it.

To highlight the error of Hobbes' argument, imagine the perfect immoral person: he successfully portrays himself as a pure and righteous individual to his peers, but in secret is dastardly immoral and self-focused. In fact, he's so dastardly and clever that to others he is a shining example of moral virtue; he committed, and continues to commit, wrongdoings of all variance in both type and degree. He makes promises and fulfills most, but sometimes breaks them when he knows he can get away with it. He steals when no one is looking. He cheats on his taxes, and has so for years. He has been serially cheating on his wife since their wedding night. The hypothetical list can be continually generated, but the point is that he's a perfect criminal in the sense of fulfilling all of his self-interests, even the ones of the endangering type. What do we say about the existence and continued activity of such a person? According to a prudential reason solution, he seems to slips through the cracks of the system. He plays the game so perfectly that he is exempt from rational criticism, for what more can we say if in the end he is benefitting?

But aha, says a confident reply! Surely such a kind of person can't exist. No one is that devious, that consistent and slick that he is able to sustain such an incredibly split dual moral life—eventually his immorality will become exposed, and thus the benefits he once reaped will not only be gone, but all of the other benefits he normally would have enjoyed (a loving marriage, a comfortable home, the company of his peers, etc.) will be jeopardized. But while the existence of a perfect immoral person may be straining to locate, especially when coming from a purely statistical point of view, we can refer to lesser corollaries of such a person. In other words, while the perfect immoral person is likely impossible to find, we can easily bring up examples of people who have led somewhat-immoral lives and yet have grossly benefited from it during their lifetime. In fact, the key is to note that with such people, often taking their *somewhat-immoral* lifestyle has led to more pleasure and happiness than if they were to otherwise have lived a moral life. All throughout history we have prime examples of people in power such as Genghis Kahn who have committed atrocious crimes in incredible magnitude and yet enjoy most of life's pleasures (abundant food, sexual partners, success, respect, wealth, etc.). Even if we were to move our examination toward people in the modern age, we will still easily find somewhat-immoral people successfully benefiting from their livelihood. Powerful leaders of nation such as North Korea enjoy bountiful pleasures in gross disproportion to the rest of the population. Gangsters of all kinds, but most notably ones in the Mafia, have had great success in leading a life of crime. And perhaps most notably, many businesses managers/owners have engaged in a modern Gyges-esque activity. The common thread between these examples is that they focus on the proportion of success between evading and being caught via external forces.

A conjecture could be raised that though criminals with a relatively perfect track record do exist, there still exists the problem, or "opportunity cost," that they may ultimately be caught and thus suffer the consequences of their immorality. I don't see this as a significant problem in the slightest. It assumes an empirical premise that is essentially unfounded. Either this criticism refers to the criminal being plagued with the horrors of the possibility of capture (not only are many criminals unperturbed by this horror, but I'd be willing to say most aren't), or more likely it refers simply to the possibility of capture in and of itself. Frankly, it's quite simple to imagine situations and people who have rational reasons to suppose that undergoing the risk of possibility of capture is heavy enough of a burden to carry as a criminal, and to have such risk-taking pan out in their favor. During the heyday of the Mafia, policemen actually hesitated to interfere with Mafia members, as their presence often instilled low crime rates (due to their domination and suppression of the local gangs), and in fact this led to many law-abiding townsfolk praising and glamorizing Mafia members. Crime is as eclectic and amorphous as any type of human activity, and in particular, given the abundance of crime in both number and type, it shouldn't be difficult to be able to refer to actual people who have been successful criminals. We can't rely on the uneasiness of the conscience, as it does not haunt everyone. In fact, I would argue that most often the type of person I am referring to is a criminal who habitually commits crime, and thus is much more likely to have been psychologically accustomed to immoral acts. And while risks are inherent, there are plenty of situations in which a gamble can be made between a high-risk, high-rewards lifestyle over a mundane suburban life. The final quote of the movie *Goodfellas*, a quintessential film about the Mafia by Martin Scorsese, comes from the main character (who used to be an active Mafia member) before being forced into witness protection. He confined himself to living a law-abiding life in suburban America, and sums up his attitude about risk versus reward perfectly:

> Anything I wanted was a phone call away. Free cars. The keys to a dozen hideout flats all over the city. I bet twenty, thirty grand over a weekend and then I'd either blow the winnings in a week or go to the sharks to pay back the bookies. Didn't matter. It didn't mean anything. When I

> was broke, I'd go out and rob some more. We ran everything. We paid off cops. We paid off lawyers. We paid off judges. Everybody had their hands out. Everything was for the taking. And now it's all over. And that's the hardest part. Today everything is different, there's no action...I have to wait around like everyone else. Can't even get decent food...right after I got here, I ordered some spaghetti with marinara sauce, and I got egg noodles with ketchup. I'm an average nobody...get to live the rest of my life like a schnook (Goodfellas 1990).

David Hume has a different answer, in ways similar to Plato. He puts forth the example of a sensible knave, who has the exact same qualms as the irresponsible fool. Cognizant of the rationality of generally being moral, the sensible knave will only commit injustice when the opportunity permits. Hume makes several arguments. One is identical to the one Hobbes made, namely about the risk of failure and thus forfeiture of the trust of others—as we've talked about before, this falls flat for empirical reasons. He makes another argument in which he criticizes the value prioritization of the sensible knave, arguing that the treasures and happiness from having a healthy conscience and just character is much more valuable than the material goods that injustice would bring about. He makes further suggestions that having an inward peace of mind and morally consistent life is of much more worthy, which gives his argument distinction from Hobbes as these are things the sensible knave cannot avoid. Whereas the external punishments of crime can be escaped by anyone given the right circumstances, avoiding the torments of an unhealthy inner life is something Hume uses as a blockade against the sensible knave, assumedly because no one can escape their own mind. The problem with this argument is once again empirical—nothing in history has been shown to suggest that people can't live mentally healthy lives while living a life of injustice. The harmonious soul and the healthy conscience both fall under the same category of mistakes. And thus, it seems that if an external, prudential reason is the only strategy for WBM, there hasn't been a convincing case.

Business ethics intersects comfortably in the discussion of whether morality pays (or rather, whether we should conceive of morality from a profit driven aspect). While it would be wrong to say a model for businesses to successfully operate is identical to how humans beings should live their lives, in society with other similar people (i.e., people who also are self-interested with their own goals and desires), it would be seriously wrong to deny a connection between the two. As such, the WBM problem naturally arises in business ethics: how do we handle the natural incentive for businesses to maximize profits and minimize costs while at the same time preserve ethical concepts such as rightness and fairness? Beyond the obvious business grievances that transgress the law (e.g., the Enron scandal), there are many "harming other people in ways outside their own control [that are not] covered by laws or influenced by markets" (Hosmer 1994, p. 192). In the face of this problem, LaRue Tone Hosmer argues that businesses should act morally in order to build trust, commitment, and effort, and that in the long term this will bring about increased profits and corporate stability. His argument is outlined as follows:

> (1) treating people in ways that can be considered to be "right" and "just" and "fair" creates trust; (2) trust builds commitment; (3) commitment ensures effort; and (4) effort is essential or success (Hosmer 1994, p. 199).

There is something evidently correct about this. Despite the criticism I directed towards the three previous attempts, they all are building on the approach that self-interest can be seen as the best strategy for answering WBM. There is truth to the fact that committing crimes, even if they seem safe, comes with the severe risk of jeopardizing our pleasure. And although saying our soul will become wretched, or our mind unhealthy, is exaggerated, having a healthy conscience is positive, in some sense; being moral as a general guideline is often an optimal outlook. And yet, there still remains a problem. In response to Hosmer, Bill Shaw and John Corvino agree with Hosmer's causal reasoning, but remind us of the ancient loophole, namely, cases in which mere appearance of moral behavior replaces moral behavior. If this can achieve the same results, why practice actual moral behavior?

Shaw and Corvino suggest a virtue ethics approach, in which personal development of moral character is stressed. By consistently practicing moral behavior, one can eventually condition a desire for the good, and in turn shun the bad.

> Through character formation, virtue ethics seeks to cultivate right desire. In the same way that good eating habits over a period of years make it increasingly easier to choose healthful foods over fattening ones, good life habits (i.e. the virtues) are self-reinforcing (Shaw 1996, p. 379).

More importantly, Shaw and Corvino believe the appealing aspect of virtue ethics is that it blends self-interest and morality more closely than the strict calculations of utilitarianism or the formalism of Kantianism. Instead of a list of do's and don'ts, virtue ethics is contrived as an effort toward a life of excellence. They take a step further, arguing managers can learn that profit is not only in terms of money, and in a virtue ethicist framework they will come to learn the value of concepts such as the health and the well being of their community. It pushes the recognition of self-interest past mere profit and material consideration. "It suggests that to constantly ask, 'How much can I get away with?' is not merely the wrong way to approach business; it is the wrong way (ultimately, an unfulfilling way) to approach life" (Shaw 1996, p. 381). I think this strategy has great merit. However, although it improves on Hosmer's ideas and thus mitigates the problem of mere appearance of morality, it can't account for all cases. A virtue ethicist can practice and regulate himself from most instances, but surely not all. And importantly, when one asks the question, they are in a situation in which the odds are in favor of injustice. Corvino himself admits this misstep ten years later, noting his approach failed.

> 'Why be moral?' question does not involve a global skepti-

> cism about moral reasons. Rather, it grants that there are moral reasons but wonders whether they are sufficient in the present case. So we can imagine someone saying, "Yes, I understand that character is important. But I have the opportunity to make millions—millions!—and maybe that's worth putting up with a little character tarnish." On this point, our answer seems to have been no improvement over (Corvino 2006, p. 6).

There emerges another approach to the WBM, coming from the surprising source of the often forgotten philosopher Confucius. This approach entails not preferring moral reasons or prudential reasons, but rather somewhat a combination of the two. Yong Huang argues from the writings of the Neo-Confucian philosopher brothers, Cheng Hao and Cheng Yi, for a synthesis explanation. That is, the reason to be moral is that "[to be moral is] something joyful. Since we are all inclined to do things that are joyful, we should be moral" (Huang 2008, p. 335). To obtain this happiness is not immediately easy and Huang would argue that many people will realize they not only will not find happiness in moral acts, but might find the very opposite. This puzzling paradox is to be explained by the necessity of "genuine knowledge." To acquire this knowledge is not by learning it in school or through analytic apprehension, but by habitual practice; one must pursue an "affective function of the heart… nothing is as important in learning as to get it yourself" (Huang 2008, p. 336). It is through this self-learning exercise of the heart that one can learn to gain genuine knowledge, and thus be in joy when being moral.

While one is tempted to draw an immediate connection to Plato's and Hume's theories, and indeed there is some similarity, Huang's position is unique in that it emphasizes the pleasure from morality by self-cultivation through effort and determination. The final component of the theory is that the motivation to want to be just is that to be moral is characteristic of being human. They argue that the essential difference between humans and animals is our moral heart. Therefore, the reason to be moral is because we are human, and since we are inclined to find joy we should aim to be moral. This bears striking resemblance to Aristotle's conception of pleasure. He too believes we can't find the pinnacle of human happiness in mere pleasures shared by beasts. For Aristotle, "the pure pleasure proper to human beings must be related to virtuous activities" (Huang 2008, p. 343).

Is the moral part of humanity actually an essential quality? This is an underlying assumption for which I think there needs to be a convincing argument, but Huang, or any of the Neo-Confucian brothers, fail to do so. Notwithstanding the heavy literature surrounding what is the essential nature of humanity, at the very least we must recognize that a good definition must both be something unique to humans and representative. Morality is arguably unique, but it isn't the only unique part of mankind. I could be facetious and reference our DNA as being unique, but we would hardly think DNA as being the essential characteristic of humanity. But in Huang's favor, morality plays a much more significant part in the life of mankind, more so than DNA ever does (that is, in the sense of how we conduct our lives). However, couldn't I say our higher functioning psychology is what makes us unique? To reduce our differences to the essential quality of morality seems arbitrary, especially since higher functioning psychology is what allows for morality and other features. Calling morality the characteristic of humanity is inappropriate when higher functioning is both more fundamental and more pervasive. And without the moral heart, being a necessary characteristic of humanity, Huang's theory loses its footing and relegates itself to a similar status as Plato's theory. Notwithstanding the fact that morality is not necessarily the only essential characteristic of humanity, to say of people that they have a proper "function," or role, is somewhat curious; there is a scent of a mistake when talking about Aristotelian/Augustinian-esque conceptions of humanity with which Neo-Confucianism is familiar. It's perfectly coherent, and in some sense essential to the concept, to ask whether a car is a "good car" or whether a doctor is a "good doctor," for being a good X in this sense is to perform its function well. When we ask what a hammer is for, we appeal to a functional definition, and thus we are able to understand how to answer the qualities of a good hammer (e.g., effectiveness in hammering in objects, ease of use, safety, cost, etc.). But to ask the question of what people are for becomes problematic—is there really something humans are for? Are we not just prima facie *as such*?

Some may intuitively feel as though this conception of morality is wrong. To expand on this intuition, it can be argued that prioritizing self-interest over morality is in some sense an illegitimate process of creating a moral framework. As Theodore Drange correctly notes, "many of us are inclined to say that people who are moral only for the sake of expediency have not internalized the moral rules that they follow, and so they are not moral at heart: their so-called 'morality' is shallow and superficial" (Drange 1998). And indeed, there arises a quite peculiar image of a person who is constantly shifting through daily life by examining moral acts under the analysis of whether or not it will benefit him. It seems to portray a deeply cynical view of morality; a reality in which people are appealing to moral behavior simply to profit.

While this intuition may be bothersome, a revisionary understanding of how a prudential approach to morality can perhaps satisfy some of the intuitive qualms. Attempts to internalize prudence-reasoning as a general rule may bring about a scenario in which not much is different from normal value systems of morality. By gradually making injustice unprofitable and undesirable, we could perhaps shift the desires and mentality of otherwise would-be offenders from attempting to profit from opportune injustice. I'm not advocating some behavioral therapy program in the sense of "shifting desires;" rather, I'm referring to the long-term goals of a large social/political entity, such as a state or nation. Granted this revisionary approach is quite broad and vague, but imagine a society in which the police have a remarkable success rate in solving crimes, economic inequality is negligible, the economy is booming, etc. A healthier society is likely to have many less reasons to be immoral, and in effect, the endeavor of pursuing even opportune injustice would become an increasingly unattractive option. As stated before, when considering Hume's and Hobbes' objections, while they haven't provided an airtight proof for the risk a somewhat-immoral person undergoes, a society in which such actions become increasingly unfavorable can at least solve a substantial amount of somewhat-immoral cases. And while the somewhat-immoral person can easily become a reality in countries with terrible living conditions, a country with vibrant living conditions will be less disposed to the occurrence of somewhat-criminals. When the list of avoidable crimes becomes smaller and less worthwhile, the simple empirical fact is that crime lowers. If the prudential reasoning model were to be applied in a utopia, and the normal model of having morality as the ultimate value were to be applied in the same utopia, it would seem to me that nothing would be different insofar as there would be no substantial difference in crime. The only worry that remains is perhaps the mentality of the prudential reasoning model would still be an uneasy condition—but as I've stated before, internalizing the general rule does not lead to a crude analysis of everything in cost-benefit terms. Having in mind that a virtue ethics is advantageous, we can develop a personal character that is accustomed to actual moral behavior, even if the original purpose is to benefit. I may not commit a murder in part because I realize the effects the social contract would admonish upon me, but this isn't a dominating mentality that pervades my mind every time I have moral thoughts.

I grant this response may be weak, but it's not a worrying concern. This is because even if the intuitions held are too strong, for my revisionary understanding to overcome, it doesn't overrule the assumption that morality must have ultimacy. Those who claim a prioritization of self-interest over morality reduce the power of morality, are inexplicitly assuming that morality must be the ultimate value (for only assuming this would it be coherent to say prioritizing self-interest is defacing morality). But this is merely an assumption that, unless it has some reasoning behind it, serves as merely a point of disagreement that may be dismissed. They are essentially begging the question when saying self-interest can't be held over morality.

I remain convinced that a self-interest oriented approach is the way to solve WBM, if there truly exists a solution. However, given the struggle of many philosophers, the constant threat of the ancient loophole of mere behavior in morality is ever present. With Plato and Hume we had the proposal that happiness is to be found inward, and thus maintained moral integrity is innately connected to happiness. This proved to be wanting, as empirical evidence suggests moral viciousness and internal happiness can very well co-exist. Hobbes and Hormer argued that external risks would mitigate our morality, yet even with Shaw and Corvino's virtue-ethics suggestion, the problem of mere appearance is not defeated. Morality is often seen to be founded on a pervasive attitude of disinterest, but if the only reinforcement for WBM is a self-interested account, it's exceedingly difficult to see in what way morality can be consistently upheld. Even if we allow the utmost periodic moments in which it is rational for a crime to occur, then this significantly weakens morality. No moral system has the conditional clause of waning when the subject wants to. It is almost universal among ethical theories that when self-interest and morality collide, that morality remains triumphant—or else what is the point of morality? Despite all these failures from proposed solutions, collectively they demonstrate the power of the prudential approach. For most events and most people, acting morally for ultimately self-interested reasons is both coherent and motivating. It is the perpetual issue of mere appearance that reinvigorates the problem.

Bibliography

Corvino, John. (2006) "Reframing morality pays toward a better answer to why be moral in business," *Journal of Business Ethics* 67(1), pp. 1-14

Drange, Theodore M. (1998) "Why Be Moral?" retrieved from http://www.infidels.org/library/modern/theodore_drange/whymoral.html

Goodfellas. (1990) Dir. Scorsese, Martin. Warner Brothers, Film.

Hosmer, LaRue Tone. (1994) "Why be moral? A different rationale for managers," *Business Ethics Quartely* 4(2), pp. 191-204

Huang, Yong. (2008) "'Why be moral?' The Cheng brothers' neo-Confucian answer" *Journal of Religious Ethics* 36(2), pp. 321-353

Hume, David. (1983) *An Enquiry Concerning The Principles of Morals* (Indianapolis: Hackett Publishing Company)

Hobbes, Thomas. (1998) *Leviathan* (New York: Oxford University Press)

Nielsen, Kai. (1989) *Why Be Moral?* (Buffalo: Prometheus Books)

Plato. (2004) *Republic* (Indianapolis: Hackett Publishing Company)

Shaw, Bill; Corvino, John. (1996) "Hosmer and the Why Be Moral Question," *Business Ethics Quarterly* 6(3), pp. 373-383

CONTRIBUTORS

Nigel Aitchison. B.A. Philosophy, University of California, Irvine. Logic, Social Phenomena, Moral Philosophy. Nigel intends to teach at the community college level or apply to law school.

Samuel Chen. B.A. Philosophy, California State University, Los Angeles. Ethics, Metaphysics, Philosophy of Religion. Sam intends to apply to law school.

Chuck Dishmon. B.A. Philosophy, James Madison University. Applied Ethics, Logic, History of Philosophy. Chuck plans to apply to Ph.D. programs.

Melvin J. Freitas. B.A. Philosophy, University of California, Los Angeles. Philosophy of Language, Logic, Philosophy of Mind, Metaphysics, Epistemology, History of Analytic Philosophy, History of Modern Philosophy. Melvin intends to apply to Ph.D. programs in the Fall of 2013.

Nathaniel Greely. B.A. Communications, University of Southern California. Philosophy of Mind, Philosophy of Language, Phenomenology. Nathaniel plans to apply to Ph.D. programs in the Fall of 2014.

Jose L. Guzman Jr. B.A. Philosophy, University of California, Los Angeles. Plato, Aristotle, Kant, Ethics, Moral Responsibility, Free Will, Philosophy of Action. Jose plans to teach philosophy at the community college level or higher.

Victoria Canada Ritenour. B.A. Philosophy, California State University, Long Beach. Cognitive Science, Medical and Applied Ethics, Social and Political Philosophy. Victoria plans to apply to Ph.D. programs.

Adam Sanders. B.A. Philosophy, California State University, Los Angeles. Metaphysics, Philosophy of Science, Philosophy of Mind, Logic. Adam intends to apply to Ph.D. programs in the Fall of 2013.

Douglas C. Wadle. B.A. Comparative Literature, New York University. M.A. Ethnomusicology, University of California, Los Angeles. M.F.A. Music Composition, California Institute of the Arts. Epistemology, Logic, Philosophy of Language, Philosophy of Mind, Philosophy of Perception. Douglas intends to apply to Ph.D. programs in the Fall of 2013.

Jianli Wang. B.E. Mechanical Engineering, Shanghai Jiaotong University. Logic, Meta-ethics, Metaphysics. Jianli plans to attend graduate school.

Master of the Arts in Philosophy
California State University, Los Angeles

The Department of Philosophy at California State University, Los Angeles offers a program of study leading to the Master of Arts degree in Philosophy. The program aims at the acquisition of a broad background in philosophy. It is designed for those preparing for further graduate study or community college teaching, and for self-enrichment. Although the department is analytically oriented, it encourages work in other areas, for example Asian philosophy, feminist philosophy, and the interaction between European and Anglo-American thought. The Department includes faculty members with diverse backgrounds and interests actively working in a wide range of philosophical specialties. Classes and seminars are small with a friendly, informal atmosphere that facilitates student-faculty interaction.

The academic programs in philosophy at California State University, Los Angeles are intended to engage students in philosophical inquiry. They aim to acquaint students with noteworthy contributions by philosophers to the tradition; to explore various philosophical issues, problems, and questions; to provide students with principles of inquiry and evaluation relevant to the many areas of human activity, such as science, law, religion, education, government, art, and the humanities; to develop in them skills of analysis, criticism, and synthesis needed for advanced work in various scholarly fields; to encourage the development of skills and attitudes leading to self-reflection and life-long learning.

Philosophy in Practice
Submission Information

Each of the student contributors was specially selected to submit a paper for this issue of *Philosophy in Practice* by one or more faculty members in the Department of Philosophy at California State University, Los Angeles. All writers are currently either students in the master's program of philosophy or undergraduate majors in philosophy. All philosophy students at California State University, Los Angeles are eligible for nomination, and those who were chosen to contribute have demonstrated a superior ability to develop and compose works of advanced philosophical writing.

For more information on *Philosophy in Practice*, please contact:
philosophyinpractice@gmail.com