Table of Contents

Olivia Y. Lee
Dr. Richard Gipps
Philosophy Directed Reading
17 October 2022

Might your new philosophy tutor be a non-conscious 'zombie' for all you know?

This paper argues that it is impossible to know if another agent is a non-conscious 'zombie'. The phenomenological zombie is a widely debated philosophical concept, namely an agent that is "functionally and psychologically identical" to a conscious agent, but whose functioning is not accompanied by any phenomenal experience, namely what it is like to subjectively undergo the experience (Chalmers, 1996). The impossibility of knowing with certainty if another agent is a non-conscious zombie can be attributed to only the suggestion, with no guarantee, of the existence of conscious experience through behavior or language of other agents. In addition, the intrinsic lack of perspective-neutral sensational experience makes it difficult to reconcile first-person and third-person conceptions of sensations. The unreliable nature of the conceivability argument for non-conscious zombies further complicates one's ability to know about the consciousness of another agent. Overall, this suggests that it is impossible to assess from a third-person perspective, through language inference or behavior, whether another agent is conscious, given that one is paradigmatically limited to first-person conceptions of consciousness.

The impossibility of ascertaining whether another agent is conscious lies in the assertion that the behavior (both physical and verbal) of other agents can only suggest the existence of conscious experience but cannot guarantee it. Wittgenstein (1953 p. 244) asserts this by probing the question of how we use language to refer to sensations. He posits that "words are connected with the primitive, natural, expressions of sensation and used in their place", highlighting the example wherein a child has hurt himself and cries, and over time an adult teaches him new pain behaviors such as exclamations and sentences. If one were to imagine a world where no behavioral expressions of pain exist, it would be impossible to teach the child what pain is, even if the child is himself experiencing pain. Therefore, verbal and behavioral expressions associated with sensations like pain are distinct from the sensations themselves. Wittgenstein (1953 p. 246) then explores the privacy of sensations. He asserts that "only I can know whether I am really in pain; another person can only surmise it" through one's behavior. It is conceivable that agents can learn to simulate behaviors (Wittgenstein, 1953 p. 250) that third parties surmise underlying sensations from – consider a non-conscious agent that observes and mimics how other agents respond to touching a hot stove, from immediately retracting their hand to an exclamation of pain. There also exist other experiences, such as the experience of seeing red, in which the corresponding mental states do not manifest in certain behaviors that can be used to infer the presence of such experiences, in the way that the experience of pain does. Any inference of those experiences in other agents would almost solely lie in their articulation of those experiences, which are again detached from whether the underlying sensation is actually present. Ultimately, Wittgenstein's analysis of the

language associated with sensational experience suggests that we cannot know that what other agents say, or more broadly how they behave, have any underlying sensations tied to them, since the language and behavior associated with sensations are distinct from the underlying sensations. The only thing one can be certain of is whether one is undergoing sensational experience, but one cannot be certain of this in other agents.

Beyond the impossibility of knowing whether another agent genuinely undergoing phenomenal experiences through behavioral inference, a crucial element of the problem of other minds is there is no notion of sensational experience independent of perspective. Both Wittgenstein (1953 p. 350) and McGinn (1984) put forward the idea of using one's own experiences as a model from which one extrapolates the experiences of others, or an "ostensive paradigm for getting the concept of another's experience" (McGinn, 1984). Wittgenstein posits the imaginative extension of one's first-person model of sensation on another agent: "if I suppose that someone has a pain, then I am simply supposing that he has just the same as I have so often had" (Wittgenstein, 1953 p. 350). It is hard to justify the soundness of such a projection since the "sameness" of these experiences is questionable, given the logical leap required to conceive of the sensation of "pain" learned from the paradigm of one's own field of consciousness being true at another field of consciousness as one imagines "rain" to be true at spatial region different from one's current spatial region (McGinn, 1984). To better understand the logical leap required in such a projection, consider Molyneux's question as an analogy where "first-person uses of pain are to third-person uses what tactually based uses of "square" are to visually based uses" (McGinn, 1984). In the same way a blind man cannot imaginatively extend the concept of "square" that was acquired through the tactile sensory modality to visual concept – because any image formed of the "square" property is fundamentally shaped by the tactile modality through which the concept was presented to him – one cannot imaginatively extend the concept of sensations from first-person to third-person, since the acquisition of those sensations are constrained by the first-person perspective. Tanney (2004) also alludes to this notion of the divergence between first-person and third-person criteria for the application of consciousness. To abstract such sensations from the first-person perspective to form a conception of another agent's sensations from a third-person perspective, one would need to determine an abstractable, objective entity which can provide a conception of sensations independently of how they are presented in the first- or third-person, or as McGinn (1984) puts it, "a conception that is somehow neutral between these two epistemic routes onto the state of pain". However, this is impossible – any way for one to conceive of sensations incorporates either the first-person perspective (i.e., the subjective experience of oneself undergoing pain) or the third-person perspective (i.e., the perceptual experience of observing another agent supposedly experiencing certain sensations). As such, there is no perspective-neutral notion of sensations, making it impossible to conceive or know of another agent's experience.

It is also worth considering whether the concept of a non-conscious zombie is even conceivable. Wittgenstein (1953 p. 353) states that "[a]sking whether and how a proposition can

be verified is a special form of the question “How do you mean?” The answer is a contribution to the grammar of the proposition.”. Through his discussion in subsequent paragraphs about the meaning of sentences, he shows that the determining the truth of such propositions regarding sensations is not straightforward. His arguments have a verificationistic slant, in which words are meaningful only if they are first understandable by others, and then publicly verifiable. For instance, Wittgenstein (1953 p. 351) states “when you ask me, “Don’t you know then, what I mean when I say that the stove is in pain?”, I can reply: “These words may lead me to imagine all sorts of things; but their usefulness goes no further.”” Beyond objects like stoves, the same line of questioning could be applied to an arbitrary third person, potentially a non-conscious zombie, and the response would be similar. Crucially, this line of reasoning is related to the above argument about the limits of language in penetrating other minds. Because a non-conscious zombie would say the same things as a conscious agent about their sensations and experiences, it would be very likely that a non-conscious zombie is convinced that it has experiences, and it would be difficult to convince it otherwise. This means that it is conceivable that we are all zombies, at which point the idea of distinguishing non-conscious zombies from conscious agents breaks down.

An alternative view is because there is no logical contradiction in conceiving the notion of a phenomenological zombie, they are not logically impossible. This view, held by Chalmers (1996), puts forward that there is no internal incoherence in the notion of a zombie twin that is functionally and psychologically identical, but whose functioning is not accompanied by any phenomenal experience, hence the notion that other agents are non-conscious zombies is intelligible and thus possible, even if it is not plausible. To this view, the conceivability argument for non-conscious zombies is begging the question. Presenting the supposed conceivability of a functionally and psychologically identical agent that lacks conscious awareness as proof that there is more to a conscious agent than its functional and behavioral composition assumes the subject one has set out to prove, specifically the existence of phenomenal aspects (or “qualia”) in sensory experiences. Furthermore, the conclusion that non-conscious zombies are possible rests on the premises that first, zombies are conceivable, and second, anything that is conceivable is possible. While this argument is valid, its soundness is less certain. If the second premise is true and conceivability entails possibility, any conceivable fact can be both possible and impossible, which nullifies the usefulness of conceivability as a concept for evaluating the potential of non-conscious zombies. However, tighter constraints on the definition of conceivability make it more difficult for the first premise to be true, which means a very specific definition of conceivability is required to satisfy both premises. The unreliable nature of the conceivability argument makes it difficult to successfully argue for the possibility of non-conscious zombies.

Acknowledging the difficulty of conceivability arguments, Chalmers (1996) also puts forward an indirect argument for phenomenological zombies by pointing to nonstandard realizations of functional organization. Citing Block’s (1980) Homunculi-headed system as an example, Chalmers demonstrates that patterns of causal organization can be composed of

constituent elements that are not neurons. Such organizations are functionally equivalent but lack phenomenal experience, thus consciousness fails to logically supervene on the physical. To this view, if there is indeed no conceptual entailment from biochemistry to consciousness, it leaves open the question what is unique about the functional organization of conscious agents that gives rise to conscious experience. This open question may suggest that consciousness resists functional analysis, which causes the discussion of evaluating another agent's consciousness to break down. If the type of consciousness experienced by an agent is indeed substrate-dependent, another agent may be conscious but experience a different type of consciousness from oneself or may lack consciousness entirely. This concept is related to Wittgenstein's (1953 p. 293) beetle-in-a-box argument: other organisms may possess a different form of consciousness or be non-conscious, despite being functionally and psychologically identical, but there is no way to tell from language or behavior alone what experience they undergo, if any.

Another view, held by Tanney (2004), is that behaviorally identical agents must be conscious. In response to Kirk's question about zombies as physical duplicates of humans with no conscious experience, Tanney claims that the behavioral zombie twin is necessarily conscious and must have a mind. One of Tanney's arguments for the consciousness of zombie twins is that they "meet all the second- and third-person criteria for the ascription of "experiential" mental predicates in exactly the same situations as their counterparts do". She highlights that we already extend sensation concepts like pain, as well as feelings and emotions, to animals, despite their lack of behavioral similarity to humans, but unlike animals, one must also consider more complex notions like self-consciousness. Tanney argues that because of the behavioral similarity of the zombie twin, it will be able to articulate "what it feels like" to undergo certain experiences as coherently as a human and apply these concepts to itself, using this as justification for the necessity of consciousness in behaviorally identical agents. A rebuttal to this view is similar to the circularity argument made above: indeed, Tanney has pointed out that a non-conscious zombie would say the same things as a conscious agent about their sensations and experiences. However, rather than implying that behaviorally identical agents must be conscious, it instead implies that a non-conscious zombie is likely to be convinced that it has experiences. It is therefore conceivable that we are all zombies, at which point the idea of distinguishing non-conscious zombies from conscious agents breaks down.

More broadly, it is important to determine the nature of consciousness that is being contested when it comes to the potential existence of non-conscious zombies. Tanney (2004) demonstrates conflation of several definitions of consciousness or is at least arguing based on a different definition of consciousness from Wittgenstein (1953) or Chalmers (1996) above. In another argument for the necessity of consciousness in behaviorally identical agents, Tanney asks the reader to assume, for the sake of argument, that the zombie twin lacks something only privately accessible and cannot be described in the 'public' or 'common' language. However, the zombie twin will pass all tests for sight, smell, and hearing, no matter what the zombie twin claims about

whether “all is dark inside”, thus one would not conclude that the zombie twin is not conscious or not undergoing sense experiences. In a later argument, she highlights the difficulty of describing how an unconscious zombie twin falls asleep or becomes unconscious, which it would presumably be able to, given that it is behaviorally identical to a human. This suggests that Tanney is using the definition of consciousness that Hacker (2013) would refer to as perceptual and intransitive consciousness, while Wittgenstein and Chalmers are arguing what Hacker (2013) would define to be consciousness of feelings and sensations like pain. Tanney’s ambiguous definition of consciousness as “whether they have sense experiences or are conscious in the normal senses of those terms” demonstrates a definitional problem in this argument. Behaviorally identical zombie twins would clearly be deemed conscious in the perceptual and intransitive sense since it would be absurd to hold these zombie twins to a higher standard that one would a human in trying to ascertain those same forms of consciousness. However, as demonstrated above in arguments from Wittgenstein (1953) and Chalmers (1996), the argument for consciousness of feelings and sensations is less clear.

In conclusion, this paper argues for the impossibility of knowing if another agent is a non-conscious ‘zombie’. Observing the behavior or language of other agents can only suggest, though may not guarantee, the existence of conscious experience in those agents. In addition, reconciling first- and third-person conceptions of sensations is complicated by the lack of perspective-neutral sensational experience. Alternative views include conceivability and indirect arguments for consciousness in other agents, however the unreliable nature of the conceivability argument, as well as the resistance of consciousness to functional analysis, makes the soundness of these arguments questionable. Arguments relying on the behavioral similarity of non-conscious zombies often encounter a circularity problem in that behaviorally identical zombies would be convinced it has experiences, in which case the conceivable case that all agents are zombies makes distinguishing zombies from non-zombies irrelevant. Together, analysis of these arguments suggests that it is impossible to assess from a third-person perspective whether another agent is conscious, given that one is paradigmatically limited to first-person conceptions of consciousness.

Bibliography

Colin McGinn (1984) What is the problem of other minds?, *Proceedings of the Aristotelian Society*, 1984, (supplementary vol.) 58, 119-37

David Chalmers (1996) *The Conscious Mind*, New York/Oxford: Oxford University Press

Julia Tanney (2004) On the conceptual, psychological, and moral status of zombies, swamp-beings, and other 'behaviorally indistinguishable' creatures, *Philosophy and Phenomenological Research*, 69, 1, 173-186

Ludwig Wittgenstein (1953) *Philosophical Investigations*, 244-250, 310-317, 344-363; Part II, ix-x

Ned Block (1980) Readings in the Philosophy of Psychology, Volumes 1 and 2, Cambridge, MA: Harvard University Press, 268–305

P M S Hacker (2013) *The Intellectual Powers: A Study of Human Nature*, Oxford: Wiley Blackwell.

Olivia Y. Lee
Dr. Richard Gipps
Philosophy Directed Reading
24 October 2022

Might you see red where I see blue?

The inverted spectrum scenario was first proposed by Locke (1975): "There are the ways things look to me, and sound to me, and smell to me, and so forth. That much is obvious. I wonder, though, if the way things appear to me are the same as the way things appear to other people." More concretely, the simple inverted spectrum scenario posits that since people learn color terms by being shown publicly available colored objects, collective verbal behavior will corroborate even if individuals experience entirely different subjective colors. For instance, even if the way blue things look to me is the way red things look to you, we would say the same thing when we look at the sky despite our private experiences being different. This paper begins by exploring the phenomenon of engaging in public discussion about private experiences, noting that these discussions can continue even though said experiences are inaccessible to others. There is general agreement that it is not possible to publicly verify whether interpersonal subjective experiences differ. However, because discourse involving these sensations accompanying the use of the public words is possible, this renders the qualitative contents of the experience semantically irrelevant. This paper then tests the soundness of the assumption that sensation concepts are learned and mastered through the first-person paradigm before they are applied to ascribe psychological predicates to others, and attempts to dismantle this assumption by demonstrating the impossibility of defining sensations with a private language. Having pushed back on this assumption, this paper proposes that personal understanding of sensation concepts and consequent attribution of psychological states to others is conceptually linked to the behaviors displayed when undergoing sensations, which are part of the public language. Counterarguments include the potential generation of mappings between individuals' dispositions and knowledge and using these mappings to imagine the sensations another individual experiences, as well as the intuition that the "qualitative feels" accompanying one's experiences are undeniable. Overall, I conclude that the question itself is incoherent as it presupposes a private interpretation of sensation experiences divorced from public language, which is impossible.

Although it is not possible to publicly verify whether another agent's experiences of sensation are identical to one's own, the fact that those corresponding sensations have the same agreed-upon denotation in public language renders the qualitative contents of the experience semantically irrelevant. Wittgenstein introduces the inverted spectrum scenario as follows: "The essential thing about private experience is not that each person possesses his own specimen, but that nobody knows whether other people also have *this* or something else. The assumption would thus be possible – though unverifiable – that one section of mankind had one visual impression of red, and another section another." (Wittgenstein, 1953, §272). Indeed, if the inverted spectrum

scenario were true, interacting agents would still agree on calling the same colors by the same names even though their subjective visual experiences were different. Even if the qualitative contents of your visual experience viewing the sky is the same as my experience viewing an apple, we would still agree that the sky is blue, since your learned associations between this visual experience you undergo and the word "blue" has been consistent throughout your life, as are mine.

It may be tempting to settle at this conclusion that what is beyond public language or verification is beyond knowledge, meaning I could never know if you might see red where I see blue, which makes it a conceivable notion. That being said, the conceivability of this notion has important implications for the relevance of the qualitative character of sensational experiences. Since private ostensive color experiences could conceivably differ while their public meanings remain intact and agreed upon, it follows that these private color experiences have no bearing on the meanings of color terms in public conversation. Wittgenstein asserts this through the beetle-in-a-box analogy: "it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But what if these people's word "beetle" had a use nonetheless? – If so, it would not be as the name of a thing. The thing in the box doesn't belong to the language-game at all … for the box might even be empty. … That is to say, if we construe the grammar of the expression of sensation on the model of 'object and name', the object drops out of consideration as irrelevant." (Wittgenstein, 1953, §293). In essence, in such a situation where everyone has a box – which may contain a beetle, a squirrel, some shapeshifting animal, or nothing at all – and no one has access to the contents of others' boxes, the nature of each box's contents is irrelevant to the meaning of the label 'beetle' describing those contents. Crucially, the reason for this semantic irrelevance is not because the private object is beyond knowledge or verification, but because public discussion can continue *despite this fact*. Therefore, even though the qualitative contents of subjective visual experiences may differ, the fact that individuals can publicly agree on and converse about these private experiences accompanying the use of the public word renders the contents of said experiences semantically meaningless.

Fundamentally, the incoherence of the notion that you might see red where I see blue lies in the questionable presupposition that supposedly infallible knowledge of sensation concepts is first acquired through the first-person and then projected onto the third parties. Questioning if another "sees red" where you might "see blue" presupposes that psychological phenomena (e.g., of "seeing blue" or experiencing "pain") can only be learned from one's own case and then applied to others. The notion is incoherent because it suggests sensation experiences can be learned or discovered by turning inward and analyzing one's own mental states, colloquially termed "introspection". It cannot be said that one comes to "know that [one is] in pain", one simply *is* in pain (Wittgenstein, 1953, §246). This incorrect notion regarding the supposedly infallible knowledge of one's own sensations stems from the false analogy that identifying or pointing inwardly to a sensation involves concentrating one's attention on the sensation of a headache or looking at the sky, and then knowing what pain or seeing blue is respectively (Wittgenstein, 1953,

§258, 275). Wittgenstein (1953, §305) thus posits the idea of an "inner sense" through which one can perceive one's own psychological states, just as we have an "outer sense" to perceive our environment, is illusory. Not only is this "knowledge" of one's own sensations questionable in truth, but it may not be considered knowledge one can acquire at all. Even if one were to persist with this idea of privately identifying sensations (e.g., seeing blue or pain) through introspection – if this is even possible – and projecting these concepts onto others, it is unclear if this generalization is sound (Wittgenstein, 1953, §293). The question becomes a transition from "There exists a concept of blue" to "You see blue", as opposed to the transition from "I see blue" to "You see blue" implied in the question. This transition would involve taking concepts acquired through introspection and applying it to another agent that cannot be observed analytically through introspection, which is a logical misstep (Wittgenstein, 1953, §253, 302). It would be equivalent to saying, "If I suppose that someone has a pain, then I am simply supposing that he has just the same as *this*" (Wittgenstein, 1953, §350). However, relying on introspection alone would not logically permit the ascription of sensation concepts to another agent outside of oneself. Dismantling the problematic fundamental assumption that sensation concepts are acquired through first-person experience and then applied to others undermines the coherence of this question.

To push back further on the idea that sensation concepts are mastered through private introspection and then applied to others, one must consider whether the concept of a private language or ruleset for sensations presupposed by this idea is even tenable. Wittgenstein proposes the following thought experiment: "I want to keep a diary about the recurrence of a certain sensation. To this end I associate it with the sign "S" and write this sign in a calendar for every day on which I have the sensation. … that is done precisely by concentrating my attention; for in this way I commit to memory the connection between the sign and the sensation. – But "I commit it to memory" can only mean: this process brings it about that I remember the connection *correctly* in the future. But in the present case, I have no criterion of correctness. One would like to say: whatever is going to seem correct to me is correct. And that only means that here we can't talk about 'correct'." (Wittgenstein, 1953, §258). This thought experiment, where an individual creates a private mental dictionary mapping words (or more broadly, signs) to the sensation concepts they name, prompts reflection on how one learns whether one is categorically experiencing a certain sensation. For if a sign were to become a name of a sensation, one must determine how it is to be used, even if only for private use. Ultimately, attempts to determine "private" definitions of sensation experience inevitably venture into the public language, even the word "sensation" itself (Wittgenstein, 1953, §261). Relying solely on concentrating one's attention on the sensation "S" cannot establish criteria of identity for subsequent uses of "S"; this is analogous to using a ruler to measure its own length. It makes no sense to say, "You don't see my blue" when looking at the sky, since there cannot be a private-defined criterion of identity for seeing blue (Wittgenstein, 1953, §253). Thus, it is impossible for one to establish and employ a standard of correctness in a private language, rendering signs in that language meaningless. A private language incomprehensible to anyone other than the individual would therefore be incomprehensible to the individual using it.

How then do we acquire our understanding of sensation concepts and attribute them to others? Given our dismantling of the assumption of first-person acquired sensation concepts being applied to third parties, a key realization is that behaviors are integral to sensation experiences, and consequently, the meaning of sensation terms. Wittgenstein posits that there is not only an empirical link between the mental (traditionally considered "the inner") and behavior (traditionally considered "the outer"), but also a conceptual or logical link between sensation experiences and behavior: "How do words *refer* to sensations? … how is the connection between the name and the thing named set up? … For example, the word "pain". Here is one possibility: words are connected with the primitive, natural, expressions of sensation and used in their place. A child has hurt himself and he cries then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behavior." (Wittgenstein, 1953, §244). Concretely, we would say that pain is a feeling that leads us to behave in certain ways, or more broadly, a sensation "S" is a word in our public language which is defined by reference to behavioral criteria. This includes verbal exclamations, which are grafted over natural ones, and such expressions become integrated into the concept of sensation. As Hacker (2018) describes, "[p]ain stimulus, pain, pain utterance, pain behavior, and pain circumstances belong together. They cannot be logically separated." Pain behavior is integral to this identity criterion if we are to possess a concept of pain at all. The meanings of psychological expressions are partly determined by the behavioral criteria which are part of the public language, which is what allows individuals to determine they are experiencing certain sensations as well as ascribe the state of undergoing those sensations to third parties.

An alternative view contends that generalizing from sensation concepts acquired through first-person experience to third parties is logically valid. Dennett (1991) puts forward the example of imagining "what it must have been like to be a Leipzig Lutheran churchgoer in, say, 1720, hearing one of J. S. Bach's choral cantatas in its premier performance". He contends that it is not utterly impossible – in fact he claims it is quite easy – to imagine the sensation experiences the churchgoers underwent, "to respond to those tones with the same heartaches, thrills, and waves of nostalgia" (Dennett, 1991). In particular, he claims "if we want, we can carefully list the differences between our dispositions and knowledge and theirs, and by comparing the lists, come to appreciate, *in whatever detail we want*, the differences between what it was like to be them listening to Bach, and what it is like to be us. While we might lament that inaccessibility, at least we could understand it. There would be no mystery left over; just an experience that could be described quite accurately, but not directly enjoyed unless we went to ridiculous lengths to rebuild our personal dispositional structures." (Dennett, 1991). This argument in flawed in two ways. First, it suggests it is possible to generate a concrete and sufficiently detailed description of the differences between two individuals' dispositions and knowledge. This is incredibly vague, and is likely much more difficult than he suggests, if it is even possible. Second, it relies heavily on the assumption that one becomes acquainted with concepts of sensation through some sort of introspective process. Then, given two individuals with different "inner senses" and a sufficiently

detailed list of differences, it is conceivable to define some sort of mapping between their different "inner senses", so that an individual can use their introspectively mastered sensation concepts to ascribe psychological predicates to others. It relies heavily on the fundamental assumption that one somehow generalizes to others or empathetically puts oneself in another's shoes, enabling them to imagine what one would feel or experience in those circumstances. Prior arguments have dismantled two facets of this assumption: first, that sensation concepts acquired first- and third-person are paradigmatically different, and second, that a private, first-person understanding of sensations concepts independent of public language is possible. Beyond the empirical links between psychological states and behavior that Dennett is presumably relying on to create such mappings between two individuals' "inner senses", it does not address the logical or logical links between the mental and the behavioral.

There are numerous variations and responses to the scenario of spectrum inversion, more broadly referred to as the inverted qualia argument (as other forms of inversion beyond have been introduced over the years, such as the Inverted Earth scenario (Block, 1990) which affects orientation as opposed to color). Shoemaker (1982) also suggests the idea of intrapersonal qualia inversion (in which an individual wakes up one day and all their color experiences are inverted overnight, i.e., the sky looks orange) leads to an inexorable slide to the possibility of a behaviorally undetectable interpersonal scenario if one were to undergo semantic adaptation, which Dennett (1991) extends. There is a consensus that intersubjective comparison of qualia is impossible. However, most responses converge on the same, often intuition-driven dead-end that the qualitative "feels" of that accompany their sensations just "seem obvious" (Dennett, 1991) and could conceivably differ from person to person. Supporters of this stance thus find Wittgenstein's argument – that the contents of private experiences are semantically irrelevant – unsatisfactory. Wittgenstein himself anticipates such responses, likening them to the scenario in which one says ""Another person can't have my pains." – *My* pains – what pains are they? What counts as a criterion of identity here? … I have seen a person in a discussion on this subject strike himself on the breast and say: "But surely another person can't have this pain!" – The answer to this is that one does not define a criterion of identity by emphatically enunciating the word "this". Rather, the emphasis merely creates the illusion of a case in which we are conversant with such a criterion of identity, but have to be reminded of it." (Wittgenstein, 1953, §253). Again, proponents of the existence of undeniable "feels" accompanying their experiences are holding tightly onto the intuition that sensation concepts like pain can be subjectively, privately, and uniquely defined by the individual, which cannot be the case as it is impossible to establish an entirely private standard of correctness. Furthermore, despite the possibility that sensation experiences differ between individuals, the fact that public conversation about these discussions can proceed regardless renders the contents of these experiences semantically irrelevant, no matter how unintuitive this may seem. Thus, interpersonal qualia inversion may be possible (and beyond verification), but demonstrating the qualitative contents of subjective experience are irrelevant on the basis that public discussion continues in spite of these differences counters those arguments.

The final point of discussion concerns the acquisition of sensation concepts. Recall the earlier argument that sensation concepts are not acquired through private introspection, but through public language and behaviors. Consider the popular philosophical thought experiment from Jackson (1982) about Mary, the color scientist who has never seen colors. If Mary has no experience seeing color at all, but she can talk about seeing red as clearly and competently as you and me, can she said be said to know what "seeing red" means? Does she learn anything when she steps out of the black-and-white room into a world of color? Numerous philosophers vividly imagine her seeing a red apple and exclaiming, "So that's what red looks like!", and if she were to be shown two unlabeled wooden blocks, she wouldn't know which is red and which is blue until she learns which color words accompany her newfound experiences (Dennett, 1991). Returning to Wittgenstein's beetle-in-a-box analogy (Wittgenstein, 1953, §293), Mary's box would originally be empty. However, it is important to highlight what information Mary *does* have: "all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like red, blue, and so on." (Jackson, 1982). While Jackson was referring more to the physical or material changes that would occur in her central nervous system upon seeing red, assuming this physical information includes all possible "seeing red" behaviors demonstrated by humans, based on Wittgenstein's line of reasoning, Mary would be said to know what "seeing red" means. This is because, as described earlier, private sensations have no bearing on the public meanings of terms like "red". Thus, it is possible to know the meaning of psychological terms without possessing the corresponding sensation of experience, as articulated by Glock (1996): "someone who applies and explains the term 'toothache' correctly, but has never had a toothache, knows what 'toothache' means. One might object that we have no reason to believe that such a person has mastered its first-person use. But we have if he can say of himself 'I haven't got a toothache'". In addition, when Mary steps out of the room into the world of color, what she acquires cannot be classified as knowledge. Once again, this would be analogously equivalent to presenting Mary with a beetle in her box, where there was previously nothing. Since the contents of her box have no bearing on her understanding of the meaning of the word, she cannot be said to have acquired knowledge.

In conclusion, given that public discussion about private experiences can proceed despite the inaccessibility of private experiences to others, the qualitative contents of the experience are consequently rendered semantically meaningless. Crucially, the assumption that sensation concepts are learned and mastered through private introspection before they are used to ascribe psychological predicates to others is not sound, which is further cemented by the impossibility of a private language to define sensations. In contrast to privately definitions of sensation, an individual's understanding of sensation concepts and consequent attribution others is conceptually linked to the behaviors displayed when an individual undergoes a sensation, which are part of the public language. Overall, the question posed is fundamentally incoherent, as it presupposes the notion that there exists a private definition of sensation experiences, which is impossible.

Bibliography

Dan Dennett (1991) *Consciousness Explained*, Little Brown & Co, ch. 14

Frank Jackson (1982) "Epiphenomenal Qualia", *Philosophical Quarterly*, 32: 127–136.

Hans-Johann Glock (1996) 'private language argument' in his *A Wittgenstein Dictionary*, Oxford: Blackwell

John Locke (1975) *An Essay Concerning Human Understanding*, Peter H. Nidditch (ed.) doi:10.1093/actrade/9780198243861.book.1/actrade-9780198243861-book-1

Ludwig Wittgenstein (1953) *Philosophical Investigations*, Oxford: Blackwell §§243-315

PMS Hacker (2018) Wittgenstein's legacy: The principles of the private language arguments, *Philosophical Investigations*, 41 (2): 123-140, §II ('Fundamental insights')

Ned Block (1990) "Inverted Earth", *Philosophical Perspectives*, 4: 53–79.

Sydney Shoemaker (1982) "The Inverted Spectrum", Journal of Philosophy, 79: 357–81

Olivia Y. Lee
Dr. Richard Gipps
Philosophy Directed Reading
14 November 2022

What is it to make the unconscious conscious?

The unconscious refers to the thoughts, feelings, and behaviors that individuals disavow, repudiate, or defend against. More colloquially, according to Shedler (2006), the unconscious comprises "things we seem not to want to know; things that are threatening or dissonant or makes us feel vulnerable in some way, so we tend to look away." Often, thoughts and feelings are rendered unconscious as they are painful, threatening, or conflict with central parts of one's conception of oneself, thus they are safeguarded in the face of acknowledged contradictory facts. For the purposes of this argument, it is assumed that the unconscious exists, based on existence arguments put forward by Freud (1915), and that the unconscious is *dynamic* in that it is not dormant or inert but influences conscious mental processes, emotional responses, and actions in the real-world. This paper analyzes three ways through which unconscious mental states become conscious: when the individual gains the ability to self-ascribe the mental state, when the individual can elucidate the psychological significance mental states and their behavioral manifestations as a single extended expression, and when the mental state is integrated into a broader unified system by revealing their associative causal links within that system. This paper also considers and refutes three alternative viewpoints. First, the supposed incoherence of unconscious emotions, to which one must clarify the distinction between the *expression* of a mental state and the mental state itself. Second, pushing back on the notion of topological separation between unconscious and (pre)conscious systems considering problematic implications that arise from such an assumption, and demonstrating that a unified associative network resolves or circumvents these implications. Finally, addressing concerns about the psychic integration of false beliefs by demonstrating the coherence of simultaneously self-attributing a belief from the first-person standpoint while also rejecting the belief as false.

Unconscious mental states become conscious when the individual acquires or regains the ability to self-ascribe it. Once an individual has acquired the necessary linguistic capacities, their mental states (e.g., pain) can then, according to Wittgenstein (1953, §244), be expressed by self-ascription via speech acts, which would make them conscious mental states. However, even after the acquisition of this linguistic ability, one may be unable to express one's mental states through self-ascription, thus a certain kind of expressive ability is absent or blocked, rendering such mental states unconscious. The ability to "self-ascribe" refers to one's expressive ability to give voice to a sincere judgement about one's own state of mind. This expressive ability is independent of the forms of expression, such as tone of voice, facial expressions, or other behavioral manifestations accompanying the expression. For instance, one can conceivably yell, "I am furious!" in a harsh tone, loud voice, and with an angry expression, while still lacking this expressive ability, and

conversely, one can have a completely neutral tone and blank expression but still be consciously angry. Finkelstein (2019) describes this as "an ability to express [a conscious mental state] merely by self-ascribing it", using the term "merely" to emphasize this independence from behavioral accompaniments to expressions.

In addition to independence from behavioral characteristics, self-ascription must also be distinguished from self-awareness; knowing that one is occupying a certain emotional state or attitude is different from, and does not entail, being conscious of one's emotional state or attitude. These concepts are often conflated due to the commonplace use of one being "conscious of" something, which is equivalent to "becoming aware of" or "knowing of" something, but they are in fact different. Consider Finkelstein's (2019) example about Max who is unconsciously angry at his mother: even though Max comes to the realization, "I must be unconsciously angry. It's the only way to explain the way I've been acting; I'm angry at my mother", knowledge or awareness of this anger through retroactive rationalization of his behavior is not sufficient to denote this anger as conscious. Mark may be capable of expressing a belief he has about his state of mind inferred from observations about his behavior – he may even be capable of expressing that anger by snapping at his mother, ignoring her phone calls, and so forth – but if he is still unable to make psychological self-ascriptions or express judgements about his own state of mind, then his emotional state of anger is rendered unconscious. This distinction is grounded in Wittgenstein's (1953, §585) elucidation of the fundamental grammar asymmetry between psychological self-ascriptions and ascriptions of psychological states in others. The (re)acquisition of the expressive ability to self-ascribe unconscious mental states hence allows them to become conscious.

An alternative but equivalent interpretation of this view is that unconscious mental states become conscious when the individual has the ability to elucidate the psychological significance of their mental states in conjunction with their behavioral manifestations as a single, extended expression of thought. Finkelstein (2019) refers to this as the "co-expressive gloss" on the psychological significance of expressed behaviors, e.g., a smile. This is not an interpretation of the smile based on (retroactive) observation of one's behavior after the fact, but an extension of the smile into a linguistic register. This would require the individual to address, with a certain first-person authority, the question of what their behavior expresses in *as it happens*, as opposed to a sort of interpretation of his behavior that another psychologically astute observer might also be able to offer. The behavior and the gloss on its significance are integrated in a unified expressive act of a conscious mental state occurring simultaneously. Again, in the case of Max and his mother (Finkelstein, 2019), the fact the Max was incapable of co-expressively glossing his behavior of forgetting to pick his mother up from the airport, and was only able to provide an interpretation (that a third-party could conceivably also provide) by retroactively observing his behavior, suggests that the anger he holds towards his mother is unconscious. For Max's anger to become conscious, he would need an expression of anger of which the psychological significance of this expression could be co-expressively glossed, for example, "I'm not answering the phone because

I'm angry at my mother", thereby extending his behavioral expression with an associated gloss that is understood by him at the time of action.

To further illustrate this point, consider the act of an individual expressing frustration by slamming a door. If the individual slams the door, and thereafter interprets and reports on the significance of their behavior (i.e., that they must be frustrated), the door slam and the gloss on it would be considered separate acts, and the individual would be unconsciously frustrated. Notably, an external observer would also be capable of interpreting the significance of this behavior. The idea that mental states are denoted as unconscious on the basis of being able to separate behavior from gloss is expressed by Lear's (2005) example of Mr. R's removing and replacing of the stone, where he claims Mr. R "doesn't understand what he is doing because he doesn't yet have angry reasons. … he knows what he is doing in the minimal sense that he is replacing the stone, but he cannot say much about it." Mr. R might experience many individual instances of this "archaic structure of removal-and-restoration", but his inability to recognize this constitutes the unconscious. Thus, "he needs to develop the ability to recognize the structure *as it is unfolding*", and "experience the unconscious emerging in the here-and-now" (Lear, 2005). Lear's emphasis on the temporal overlap of the behavior and interpretation is especially noteworthy, demonstrating that the gloss is not only semantically integrated with its associated behavior, but also temporally integrated, which is crucial in distinguishing an individual's co-expressive gloss on their behavior from a hindsight interpretation of that same behavior. If the individual, *at the time* of the door-slamming behavior, is capable of expressing what they mean by this behavior, the gloss is a continuation or extension of their expressive act and was already attached to the expressive act at the time it was performed. Crucially, this gloss is provided in a first-personal way; an external observer would not be able to say what the individual means exactly by their behavior and may only be able to posit what the individual means after the action has taken place. The behavior and the gloss are two constituents of a single, extended expression of thought that are delivered in conjunction with each other. This semantic and temporal unity is the essential distinction between conscious and unconscious mental states. Therefore, the ability to elucidate the psychological significance of one's mental states in conjunction with their behavioral manifestations as a single, extended expression of thought enables unconscious mental states to become conscious.

Unconscious mental states can also become conscious when their associative links within an individual's larger network of conscious thoughts are revealed, thereby integrating the previously unconscious mental state into a broader, unified system. An individual's mental life is an incredibly complex web of associations and meanings resulting from overdetermination and multiple functions of behaviors, of which only a portion of this network is apparent to us (Shedler, 2006). An individual's past experiences thus significantly influence present day experiences and behaviors. However, because these influences shaping present behaviors may not be consciously accessible, certain behaviors or symptoms may seem random and unexplainable. One may have a fixed interpretation of events and not allow oneself to consider alternate interpretations of an

experience, and making the unconscious conscious involves arriving at or discovering new beliefs about the history and current state of one's mental life. Thus, unconscious mental states become conscious when an individual is able to identify and articulate the psychological circumstances contextualizing symptoms that were previously deemed random or unexplainable. By taking a seemingly random mental event and tracing through the multiple associations linked to it, one can reveal a larger associative network of thoughts and feelings within which that mental event is embedded. It is this process that makes previously implicit causal links between psychological phenomena explicit and allows unconscious thoughts to become conscious. Leite (2019) introduces a similar notion in describing the process of psychic integration, "a process of bringing together into a single perspective what has been held apart of cut off, for instance in a process of psychic defense". According to Leite (2019), an unconscious belief becomes conscious once one can "stably and fully incorporate the belief into a larger, unified perspective". This process of incorporation would thus entail identifying and understanding the causal and associative links between the unconscious mental state in question and other conscious mental states, which would then enable integration of the previously unconscious mental state into a broader network of conscious thoughts and one unified perspective.

With regards to the integration of unconscious mental states into a broader system of conscious thoughts, both Shedler (2006) and Leite (2019) highlight the complexity of one's mental life often leads to conflicts and contradictions within one's own mind. This naturally leads to the question of how one deals with such conflicts. Shedler (2006) notes that individuals have complex and often contradictory feelings and motives, leading to dissonance within the mind. It is important to consider the case where unconscious thoughts, upon integration with the complex web of associations linking conscious thoughts, result in contradictory beliefs. Leite (2019) claims that psychic integration is uninhibited by false beliefs, and one can experience conflict and still unite these beliefs (including those that the individual perfectly well recognizes as false) into a single unified perspective. In the case of an individual who falsely believes that he is a thief, Leite (2019) asserts "[t]hese feelings and thoughts do not do away with his knowledge any more than everyday obsessive worries do away with one's knowledge that one turned off the stove. In his self-conscious thinking and reasoning, he retains a firm grip on the truth." Psychic integration does not entail elimination of false beliefs, but at the same time, integration cannot require a loss of an individual's grip on reality. Together, it follows that a person could simultaneously believe something and recognize that belief is false. To demonstrate that this false belief has been successfully integrated into the individual's larger network of conscious thoughts, Leite (2019) asserts that the individual needs to "give voice to both sides of the conflict without any oscillation in his total position … The utterance exhibits an apt, settled pattern of relations between both sides of the conflict and also an apt, settled orientation towards the conflict itself. The person thus has the ability to speak with one voice – a voice that expresses the totality of his complex overall perspective all at once". Specifically, this perspective is "one contemporaneous complex

subjective perspective comprising both the subjective position of the belief and the recognition of its falsehood". Thus, the process of psychic integration is not undermined by false beliefs.

One qualm about the unconscious is that the notion of unconscious emotions and affective states is unintelligible. Freud (1915) expresses this opinion as follows: "An instinct can never be an object of consciousness – only the idea that represents the instinct. Even in the unconscious, moreover, it can only be represented by the idea. If the instinct did not attach itself to an idea or manifest as an affective state, we could know nothing about it. … We should expect the answer to the question about unconscious feelings, emotions, and affects to be just as easily given. It is surely of the essence of an emotion that we should feel it, i.e., that it should enter consciousness. So for emotions, feelings and affects to be unconscious would be quite out of the question." To address this view, it is important to distinguish the notion of conscious/unconscious *expressions* of mental states and conscious/unconscious mental states themselves. It is conceivable that a conscious state of mind is consciously expressed (e.g., when Max consciously expresses his conscious pleasure through his smile when the waitress arrives with his food (Finkelstein, 2019)) or unconsciously expressed (e.g., when Max unconsciously expresses his conscious anger towards Sarah at the dinner party through his passive-aggressive remarks (Finkelstein, 2019)). In addition, it is conceivable that an individual unconsciously expresses an unconscious state of mind, as in the case of Max unconsciously expressing his unconscious anger towards his mother by forgetting to pick her up from the airport (Finkelstein, 2019). However, one would not be able to consciously express an unconscious state of mind, as there is "no grammatical or logical space for an unconscious emotion or attitude to be consciously expressed" (Finkelstein, 2019). This final scenario is what Freud might be referring to as incoherent. However, Freud's argument above neglects the case in which unconscious states of mind, including unconscious emotions and affective states, exist and are unconsciously expressed, a case which is perfectly consistent with our understanding of conscious and unconscious mental or affective states. It is clear Freud conflates the concepts of emotion and expression: "The whole difference arises from the fact that ideas are cathexes – ultimately of memory-traces – whilst affects and emotions correspond with the process of discharge, the final expression of which is perceived as feeling." Ultimately, the expression of an emotional state does not necessitate the consciousness of that state. Distinguishing the two notions is critical to demonstrating the existence of unconscious mental states.

Another alternative proposition is that of different mental systems for the unconscious and (pre)conscious. It is worth analyzing the problematic implications of assuming the existence of such distinctive categorical systems. Freud (1915) introduces the idea of three disparate systems: the unconscious (Ucs) system, the preconscious (Pcs) system, and the conscious (Cs) system. He outlines the process of a mental act first existing in system Ucs, and if it successfully passes the scrutiny of censorship, moves into the system Pcs, which contains mental states capable of entering consciousness. If, upon another form of censorship, the preconscious mental act passes this scrutiny, it will then pass into system Cs. There are two problems with this proposition, which are

circumvented by the proposition outlined above by Shedler (2006) and Leite (2019) of a single unified network of which only a portion is apparent or conscious to us, while the remaining components are unconscious but can be dynamically integrated. First, it is unclear how the process under scrutiny of censorship occurs. Freud (1915) later explains that this process involves "withdrawal of cathexis" (as well as other derivatives such as "thing-cathexes" and "hyper-cathexes"). This process, however, is rather convoluted and does not seem to have any grounding in real-world experiences. It is also unclear the extent to which the individual has agency over the withdrawal of cathexis, and if an individual is seeking to make the unconscious conscious, how they might go about doing so. In contrast, the idea of a unified network of thoughts and mental states is much more tenable in that it is sensible how an individual may attempt to make the unconscious conscious by tracing through the associative links between mental states as shaped by their past experiences. Furthermore, as outlined by Shedler's (2006) case studies in psychoanalysis, such processes are grounded in therapeutic practices employed to help patients and corroborate with real-world experiences. The second problem, as Freud himself notes, is the emergence of two potential pathways for the mechanism of this process from categorizing mental acts via the three systems. The first pathway is that transferring a mental act from system Ucs into system Pcs or Cs involves making a second copy of the mental act in question, in a fresh locality in the mind which coexists with the original unconscious record that persists. The second pathway involves a change in state of the idea, involving the same material and occurring in the same locality. Freud (1915) deems the first pathway cruder but more convenient and the second pathway *a priori* more probable but is less plastic. He goes on to discuss the implications of the first pathway in the conception of mental topography, specifically the implied topographical separation of the systems Cs and Ucs in an individual's mental apparatus, but ultimately concludes that it cannot be decided between the two possibilities which is tenable, if any. In contrast, a larger unified network of conscious experiences circumvents this issue as it does not rest on the assumption that there is any topographical separation between mental systems where conscious and unconscious mental states reside. Mental states are instead integrated in an associative network, where unconscious mental states can be explained by individuals choosing to disavow, repudiate, or defend against those states, that correspond portions of the network where individuals are unable trace the associated causal linkages which they are able to do for conscious mental states. Therefore, it is worth pushing back against the assumption that conscious and unconscious mental states reside in different locations in the mental apparatus, and embracing the notion that they are integrated in a single, unified associative network effectively circumvents several of the problems that arise from the assumption of mental topographical separation.

A final opposition, posed to the concept of psychic integration, stems from the supposed incoherence of self-ascribing a belief from the first-person standpoint even while judging or declaring that the believed proposition is false. It has been said that such a concept diametrically opposes the fundamental definition of a belief, as articulated by Wittgenstein (1953, §190) that the notion of "false beliefs" is unintelligible. However, the idea of someone consciously occupying

the subjective standpoint of a belief and self-attributing it from that position, while simultaneously acknowledging its falsity, is necessitated by psychic integration of unconscious beliefs. To this view, one must analyze what exactly is entailed if a person consciously believes a proposition *p*. As Leite (2019) writes, "[i]f she consciously believes that *p*, she is disposed to judge that *p* when she considers relevant questions. *P* will be presented in her conscious directed thinking as a premise for her reasoning about what to do and what is the case, and she will be disposed to so deploy it. She will be disposed to assert *p* when appropriate and to act on the basis of *p*." It important to note that all of this can be true even if an individual recognizes that *p* is false, because the individual can consciously believe *p* and see the world in the ways involved in consciously believing *p*, "and yet not deploy *p* as a premise in conscious reasoning, act on its basis, and assert it flat out" (Leite, 2019). While this may seem contradictory at first glance, if we determine that to "be disposed" to idea is to be inclined, willing, or simply to have the possibility to act on the basis of *p* but not *necessarily* so, then there is no contradiction. An individual can thus see the world through the lens of this belief *p*, if the individual also "recognizes it as her lens, recognizes the ways in which it is inapt, and resists, suppresses, or otherwise does not engage in some of the patterns of response that would otherwise appropriately be involved" (Leite, 2019). It is therefore coherent to incorporate false beliefs like *p* into an individual's broader system of conscious thoughts, as psychic integration does not assume a *rational* unity, making it possible for someone to simultaneously self-attribute a belief from the first-person standpoint while also rejecting the belief as false. Concretely, it is also a familiar situation in ordinary life, especially in cases where an individual is deeply attached to a belief that they later discover is false.

In conclusion, this paper assesses three ways in which unconscious mental states become conscious: when the individual possesses the ability to self-ascribe the mental state, when the individual can elucidate the psychological significance mental states in conjunction with their behavioral manifestations as a single extended expression of thought, and when the mental state is integrated into a broader unified system by revealing their associative causal links within that system. This paper considers the alternative viewpoint that unconscious emotions cannot exist, but this view can be refuted by clarifying the distinction between the *expression* of a mental state and the mental state itself. In addition, this paper pushes back on the assumption of topological separation between the unconscious and (pre)conscious systems, which has problematic implications that can be circumvented by proposing a unified associative network of mental states. Finally, regarding concerns about the psychic integration of false beliefs, this paper demonstrates that it is coherent to simultaneously self-attribute a belief from the first-person standpoint while also rejecting the belief as false, as psychic integration does not assume a rational unity.

Bibliography

Sigmund Freud (1915). *The Unconscious*. Pelican (Penguin) Freud Library, Vol. 11.

David Finkelstein (2019). *Making the unconscious conscious*. In Gipps & Lacewing (Eds), The Oxford Handbook of Philosophy and Psychoanalysis. ch 19. OUP.

Jonathan Shedler (2006). *That was then, this is now*. https://jonathanshedler.com/PDFs/Shedler%20(2006)%20That%20was%20then, %20this%20is%20now%20R9.pdf

Jonathan Lear (2005) *Freud*. ch 1 (Interpreting the Unconscious). Routledge.

Adam Leite (2019). *Integrating unconscious belief*. In Gipps & Lacewing (Eds), The Oxford Handbook of Philosophy and Psychoanalysis. ch 18. OUP.

Ludwig Wittgenstein (1953) *Philosophical Investigations*, Oxford: Blackwell §§243-315

Olivia Y. Lee
Dr. Richard Gipps
Philosophy Directed Reading
28 November 2022

How is knowledge of my own states of mind possible?

Knowledge of one's own states of mind is of great interest not just to philosophers and psychologists, but to most people in understanding the nature and sources of their self-knowledge. This paper will focus on knowledge of states of mind such as one's beliefs, desires, dispositions, character traits, values, and emotions, or what Cassam (2014) refers to as *substantial* self-knowledge, which includes states of mind such as knowing you are a kind person, you believe you want another child, or you harbor deep feelings of anger towards someone. There is a claim which may or may not be true – and its truth is usually not easy to verify – that is generally deemed valuable. This is distinct from *trivial* self-knowledge, such as knowing you believe you are wearing socks, or you believe it is raining outside. A key point to distinguish the two is there is something fundamentally "deeper" about substantial self-knowledge such that becoming acquainted with it is a genuine hard-won cognitive achievement. Much philosophical discussion has been dedicated to trivial rather than substantial self-knowledge because of certain distinctive qualities like its supposed immediacy[1] and infallibility or authoritativeness[2]. Clearly, substantial self-knowledge does not have these "special" traits: knowing you are a kind person can be mistaken and is not immediately apparent, pending the accumulation of evidence such as behavior. However, allowing epistemic privilege and "easy" access to trivial self-knowledge to dominate the broader discussion of self-knowledge mistakenly neglects the elusiveness of substantial self-knowledge (Cassam, 2014). Ultimately, self-knowledge most people are interested in, such as one's beliefs, desires, and emotions, is not straightforward to attain. In light of these arguments, this paper will assess the plausibility of two potential pathways to self-knowledge, emphasizing substantial self-knowledge. This paper first investigates the perceptual model of self-knowledge, ultimately concluding its underlying principle that mental states are mental objects that can be identified and discriminated between is flawed. In this assessment, it is determined that much of genuine perception is inferential in nature, and this paper proceeds to defend an inferential model of self-knowledge, by proposing a positive account and evaluating two major counterarguments: asymmetry and vicious regress.

One pathway to self-knowledge is the perceptual model of self-knowledge, first proposed by Locke as the "deliberate observational scrutiny which a mind can from time to time turn upon its current states and processes" (Ryle, 1949), referring to this process as "inner perception 'reflexion'", or what is colloquially termed introspection. Inspired by the optical phenomenon of

[1] i.e., that they are self-intimating and consequently non-inferentially justified.
[2] i.e., that they are exempt from error.

reflections of objects in mirrors, the mind can be said to "see" or "look at" its own operations in the "light" given off by themselves (Ryle, 1949). On this view, one's own states of mind take the form of objects that can be inwardly observed, just as one uses sense perceptions to outwardly observe and thus acquire knowledge of non-mental objects. Introspection can therefore be interpreted as a sort of inner spotlight that casts light on mental objects representing states of mind, thus serving as a pathway to self-knowledge. This involves the exercise of an 'inner sense', using the 'outer sense' of perception as a model for introspection. Freud's (1915) perspective of consciousness is reminiscent of this view, introducing the unconscious (Ucs), preconscious (Pcs), and conscious (Cs) systems. He asserts that a mental act first exists in system Ucs, and upon passing the scrutiny of censorship, moves into system Pcs, which contains mental states capable of entering consciousness. If the preconscious mental act passes another process of scrutiny, it will then pass into system Cs. The idea of a censorship process is analogous to the notion of introspection as an internal spotlight locating mental objects representing mental states and using a set of criteria to determine which mental objects pass into system Pcs or Cs and which do not.

There are several qualms about the notion that one's own states of mind takes the form of mental objects that can be perceived. One objection highlights the paradoxical nature of a divided consciousness that passes mental objects between disparate systems. If there exist unconscious mental states inaccessible to the individual, it is unclear how mental objects to be scrutinized reveal themselves for evaluation. Alternatively, if mental objects subject to scrutiny are discovered by the individual, this implicitly suggests that these mental objects were not inaccessible to the individual to begin with. Sartre (1981) raises an objection along these lines to the Freudian idea of the unconscious: "[i]f the complex is really unconscious – that is, if there is a barrier separating the sign from the thing signified – how could the subject recognize it? What is understanding if not to be conscious of what is understood?" Here, Sartre challenges the object perceptual model of mental states by questioning how one's internal censorship system can decide what mental states to repress without being aware of the mental states in question, which would then make those mental states conscious[3]. It seems implausible that one can hide things from oneself if one does not know what they are hiding, how to hide it, and why they are hiding it. One's internal censorship system would conceivably have to "take a peek" at these mental objects to determine whether they should be repressed, rendering the corresponding mental states conscious. More broadly, the idea of introspection as an "inner searchlight" casting light upon mental objects for an individual to become acquainted with their corresponding mental states is fundamentally flawed if the premise that one has unconscious mental states is true, which is generally agreed to be the case.

Another objection to the perceptual model of self-knowledge involves questioning the notion that individuals can be introspectively aware of certain properties of one's mental states.

[3] Sartre (1981) consequently denies the existence of the unconscious altogether, but one's skepticism would be better directed towards the supposed internal scrutiny of mental objects, since unconscious mental states are generally agreed to exist. As shown later, other models of self-knowledge are coherent with the existence of unconscious mental states.

Cassam argues that "[i]n perception … there is such a thing as singling out an object and distinguishing it from others by its perceived properties. In contrast, even if you are introspectively aware that you believe [proposition *P*], this isn't a matter of singling this belief out and distinguishing it from other beliefs you are also introspectively aware of. Propositional attitudes aren't 'objects' waiting to be 'singled out' on the basis of introspectively available information about their relational and non-relational properties" (2014). Cassam highlights the fallacy of modeling "mental objects" as real-world objects (and consequently, modeling introspection as perception) by demonstrating that "[w]hen you perceive that your socks are stripy you do so by perceiving your socks, but you aren't introspectively aware that you believe your socks are stripy by being aware of this belief that your socks are stripy." While Cassam uses trivial self-knowledge as an example for proposition *P*, the same argument can be made for substantial self-knowledge. Along a similar line of reasoning, Shoemaker asserts that "perception of objects standardly involves perception of their intrinsic, nonrelational properties. When it comes to beliefs and other attitudes, it isn't clear what their 'intrinsic, nonrelational properties' are, let alone what it would be for introspection to involve awareness of such properties" (1996). Both Cassam and Shoemaker refute the idea that one differentiates mental objects based on certain intrinsic properties, suggesting that modeling introspective awareness on object perception is problematic.

It is important to note that the above objection stands even if perceptual knowledge can be inferential, which, as Cassam (2014) argues in depth, indeed can be the case. In no way does it rest on the tenuous assumption that perception is non-inferential or that perception and inference are contradictory, an assumption which both positive arguments of the perceptual model as well as flawed objections tend to make[i]. The above objection shows *even if* perceptual knowledge is inferential, there does not exist a mental object with intrinsic properties from which one can infer relational properties, so the 'inner perception' model of introspection, or 'myth of the Cartesian subject' (Ryle, 1949), breaks down. A final observation is that a significant portion of self-knowledge does not fit the perceptual model, notably, much of substantial self-knowledge. The perceptual model, even if viewed as a form of automatic or fast inferential process, seems more likely to be invoked in the avowal of occurrent states of mind, like whether one is wearing socks or whether one is presently hungry. However, this is but a small subset of possible states of mind. Being able to avow mental states such as knowing you are a kind person, or you believe you want another child – especially if one has never consciously contemplated the notion before – does not seem to be a matter of locating a mental object with the mind's eye. This reiterates that substantial self-knowledge is neither immediate nor self-intimating, and unlike trivial self-knowledge, is often a hard-won cognitive achievement. Having posited that genuine perception is often inferential, this paper now turns to inferential models of self-knowledge.

The inferential model of self-knowledge contends that inference is a key source (though not the only source) of substantial self-knowledge for humans, where "my inferring is … of a geometrical conclusion from geometrical premises" (Ryle, 1949). It argues for an evidence-based

approach to self-knowledge, where evidence can take a variety of forms, such as behavioral or psychological. Formally, "[s]uppose you know that you have a certain attitude *A* and the question arises how you know that you have *A*. In the most straightforward case you know that you have *A* insofar as you have access to evidence that you have *A* and you infer from your evidence that you have *A*. As long as your evidence is good enough and your inference is sound you thereby come to know that you have *A*" (Cassam, 2014). If *E* is evidence for knowing you have attitude *A*, knowledge of *E* makes it more probable that you know you have *A*. Knowing you have *A* is hence contingent on having access to *E*; this may be through perceptual access if *E* is behavioral evidence, or another form of access for psychological states like passing thoughts or inner speech. Crucially, inference can be but need not be conscious: inferences leading to self-knowledge are normally mediated by an implicit commitment that one can only infer from *E* that one has *A* because one takes *E* to be evidence for *A* at least in one's own case, which Finkelstein (2012) refers to as the Rationality Assumption[4]. Thus, inferences can be, and often are, "automatic, effortless, and barely conscious" (Cassam, 2014), as the assumptions underpinning them are implicit.

Two arguments for the inferential model of self-knowledge are especially noteworthy. First, the argument by elimination evaluates three potential bases of self-knowledge: inference, inner observation, or nothing at all. Acquiring self-knowledge through inner observation has been eliminated in the argument against the object perceptual model. To refute the last pathway, self-knowledge can only be based on nothing if it is cognitively insubstantial (for instance the knowledge that I am here now), but there are several indications that substantial self-knowledge is not cognitively insubstantial, namely that it is fallible and incomplete. By elimination, one is left with assessing the plausibility of self-knowledge through inference. Second, the argument by example demonstrates that the inferential model is grounded in how humans actually acquire knowledge of their attitudes. Lawlor (2009) introduces the example of Katherine who feels there is a fact of the matter about whether she wants another child, but struggles to know the answer to the question "Do I want another child?". Lawlor asserts that if Katherine's feelings, imaginings, and emotions form a basis for evidence from which she concludes she wants another child, this self-knowledge can be described as inferential, specifically an inference from internal promptings, which is a genuine cognitive achievement[5]. Some may deem Katherine's example unrepresentative, as many "simple" desires seem self-intimating[6] and therefore non-inferential. However, just because a desire is self-intimating does not mean that desire is not learned through inference, albeit a very quick and automatic inference. The self-intimating nature of mental states does not explain *how* one knows one has a certain mental state. The idea that self-intimating mental states must be

[4] Note how the inferential model of self-knowledge differs from the Transparency Method: with inferentialism, one does not substitute the easier inward-directed question whether one believes *P* with the more difficult outward-directed question of whether one ought rationally to believe *P*. Rather, inferentialism commits to a rational treatment of evidence to arrive to the conclusion that one believes *P*. See Cassam, 2014, ch. 9 for more on the Transparency Method.

[5] Again, note the distinction from the Transparency Method. It would be quite odd for Katherine to answer this question by instead answering "Do I ought rationally to want another child?".

[6] In that if one possesses the relevant concepts, then one cannot have the desire without knowing that one has it.

non-inferential is rooted in the misconception that inference is necessarily slow and deliberate, but as noted above, inference can be automatic and barely conscious. The difference between "simple", self-intimating desires and desires like Katherine's lies in a difference in *degree* of how obviously or manifestly inferential the desire is, not a difference in inferential or non-inferential desires.

A common objection to the inferential model of self-knowledge is that it often "seems clear" that inference is neither required nor relevant for self-knowledge. This objection takes many forms. First, the immediacy argument, which contends that substantial self-knowledge is immediate and not based on evidence, behavioral or other. While the immediacy of substantial self-knowledge is widely uncontested, this presumption may well be worth probing: is it really the case that knowledge of our beliefs, attitudes, desires, and emotions, are immediate? Realistic, commonplace examples like Katherine's clearly suggest the contrary. Ultimately, self-knowledge that seems non-inferential is often unobviously inferential, especially if the inferential element is unconscious or implicit. Another form of this objection is the counterexample of judgements or decisions, such as the judgment "even lousy composers sometimes write great arias" (Boghossian, 1989), to which Cassam (2014) clarifies that the inferential model accounts for our knowledge of *standing attitudes*[7], rather than occurrent thoughts, being inferred from evidence. This distinction will be relevant later in discussing vicious regress. A final form of this objection contends that inferentialism suggests one engages in a detached, unemotional, exact scrutiny of emotions like love, which is over-intellectualized and divorced from reality (Nussbaum, 1990). However, the formalization of this inferential process does not in any way (a) make it out of reach for the average person or (b) make feelings redundant in acquiring knowledge of one's emotions. "To love someone is, among other things, to be disposed to feel a certain way about them, so feelings are good evidence that you love them" (Cassam, 2014). Hence, feelings form the basis of self-knowledge of underlying emotions from which truths about such emotions are accessed via an inferential process, which hardly makes them redundant.

One major objection is the inferential model of self-knowledge cannot account for the epistemic asymmetry between the knowledge of one's own mind and knowledge of other minds. This stance is best articulated by Moran: "whatever knowledge of oneself may be, it is a very different thing from knowledge of others, categorically different in kind and manner … this person knows, and comes to know, his own thoughts and experiences in ways that are categorically different from how I may come to know them" (2001). This view supports the notion of "privileged access" to one's own mental states that is fundamentally different from acquaintance with others' mental states. Inferentialism, however, makes the means of coming to know other minds and one's own mind categorically equivalent. To this objection, one must question: are the *mechanisms* employed in knowing our own minds and knowing the minds of others really that different? Ryle pushes back on this conventional view: "[t]he sort of things I can find out about myself are the

[7] Standing attitudes differ from occurrent thoughts in that they persist even when the individual is asleep and are not mental events like judging or deciding.

same as the sorts of things I can find out about other people, and the methods of finding them out are much the same … John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe" (1949). Ryle's claim is, therefore, that our access to our own minds is no different in principle than our access to the minds of others.

The inferential model indeed acknowledges asymmetry between knowledge of oneself and knowledge of others; however, it contends asymmetry of *evidence*, not *mechanism*. Meaning, the 'epistemic privilege' one has of one's own mental states that others cannot access is not a result of first-person non-inferentialism vs. third-person inferentialism, but rather, access to different evidence from which inferences are made. Revisiting Lawlor's (2009) example, Katherine knows whether she wants a child based on her internal promptings (e.g., passing thoughts, imaginings, fantasies etc.). However, her friend Melissa cannot determine what Katherine wants on the same basis of evidence, as Melissa can only rely on Katherine's articulations and behaviors. Both Katherine and Melissa are drawing inferences from evidence, and they are fundamentally using the same inferential mechanism. The asymmetry between Katherine's knowledge of herself and Melissa's knowledge of Katherine thus boils down to a difference in evidence available to each. Katherine has "privileged" or "peculiar" access to her mental states to the extent that certain evidence is readily available to her that is inaccessible to Melissa, but not because Katherine adopts a special non-inferential mechanism of knowing her own mental states that Melissa does not.

The use of equivalent mechanisms by Katherine and Melissa is what Moran takes issue with, as he believes the mechanisms for acquiring knowledge of an individual's mental states from a first-person vs. third-person standpoint are categorically different. However, it is worth questioning what motivates the naïvely firm grip on this notion of mechanistic asymmetry between the first and third person scenarios. It is not incoherent for an individual and a third-person observer to the same conclusion about the individual's belief or attitude given the same evidence. Conceivably, if Katherine were to articulate all her relevant internal promptings and other evidence that Melissa cannot access, both Katherine and Melissa could draw inferences from the same set of evidence and hence reach the same conclusion. Furthermore, as Ryle articulates, one could possibly know others' mental states better than one's own[8]: "[a] residual difference in the supplies of the requisite data make some differences in degree between what I can know about myself and what I can know about you, but these differences are not all in favor of self-knowledge. In certain quite important respects it is easier for me to find out what I want to know about you than it is for me to find out the same sorts of things about myself. In certain other respects it is harder" (1949). This further supports the claim that any asymmetry in knowledge of mental states, whether in favor of the individual knowing their own mental states better or that of others, arises from an asymmetry in evidence available, rather than the mechanisms through which such knowledge is acquired[ii].

[8] This is especially applicable to knowledge of character traits or values which are highly dependent on behavioral evidence, or where internal promptings are disavowed or unconscious. It is entirely possible to believe that another person is caring, self-centered, respectful, or arrogant before or without recognizing those traits in oneself.

The final objection to the inferential model of self-knowledge – perhaps the most threatening one – is that it generates a vicious regress, or an infinitely recursive inferential process with no stopping point. This view, held by Boghossian (1989), asserts that if self-knowledge is inferential, it must be acquired via inference from other beliefs that are themselves inferential, resulting in a vicious regress. The only way to avoid this is to concede that knowledge of some mental states can be acquired non-inferentially, which would be the stopping point of this recursive inference. This means not all knowledge of one's beliefs can be inferential, which poses a problem for the inferential model. Cassam (2014) first clarifies that while inferentialism can be a *key* source of self-knowledge it need not be the *only* source of self-knowledge. He then points out the fallacy in "the assumption that if self-knowledge is inferential it must be acquired from inference from other known beliefs" as we can acquire self-knowledge through inference from internal promptings that are not standing attitudes and hence not within the explanatory scope of inferentialism, which describes knowledge of our own standing attitudes. This naturally leads to the question of how one acquires knowledge of one's internal promptings.

One route might be to suggest that knowledge of internal promptings is non-inferentially acquired, thus internal promptings are the stopping point of recursive inference. However, as illustrated prior, both the object perceptual model of internal promptings, as well as the argument that internal promptings are based on nothing, do not hold water. A more promising alternative is Carruthers' (2011) account of a self-interpretive process that accesses information about the subject's current or recent circumstances, behavior, and mental life. Revisiting the example of Katherine knowing her desire for another child (a standing attitude) via inference from internal promptings: Katherine identifies her feeling of yearning for another child by interpreting it, which involves some cognitive effort. She does not identify this yearning by noticing or "reading off" the "distinctive phenomenal character" of the feeling itself, which the object perceptual model would suggest. Rather, in this interpretive process, she relies on (in addition to feelings, inner speech, and passing thoughts) *contextual knowledge*, like the fact that she has been thinking about having another child recently. Using contextual knowledge as evidence to determine the nature of complex internal promptings invokes an "inference to the best explanation" (Cassam, 2014), rather than inductive or deductive, but an inference nonetheless. Inferentialism thus accommodates mistaken inferences about internal promptings and consequently standing attitudes, since one's best explanation may well be incorrect, giving rise to irrational beliefs or attitudes[9]. Therefore, through a self-interpretive process, one makes inferences[10] from background knowledge of one's current circumstances, behavior, and mental life to attach specific meaning to internal promptings[iii].

[9] This key point makes the inferential model more compelling than the Transparency Method, as it overcomes the significant limitation of irrational beliefs, desires, and emotions that the Transparency Method does not accommodate.

[10] One might consider non-inferential interpretive processes, but interpretive processes are inherently inferential as background knowledge plays a *supporting* role, not just an *enabling* role, in the process. See Cassam (2014, p. 166).

At this point, self-knowledge of standing attitudes seems to be obtained via inference from other standing attitudes or internal promptings that are themselves inferentially acquired, which still leaves open how the problem of vicious regress can be resolved. Cassam first points out that internal promptings are not the only evidence from which our standing attitudes are inferred, but there is also behavioral evidence, which one accesses through outward perception and thus very differently from internal promptings. *But what of the case where one solely relies on internal promptings to acquire knowledge of standing attitudes?* Cassam attempts to refute the objection by contending that "[w]hen people worry about the regress problem they aren't necessarily assuming that genuine explanations can presuppose no other knowledge. Rather, their objection is to explanations of knowledge which presuppose the very knowledge they are trying to explain" (2014). Thus, the inferential model does not generate a vicious regress, because the explanation of knowledge of internal promptings does not presuppose the very knowledge of internal promptings. Cassam posits that internal promptings are derived from background knowledge, which is in turn derived from memory, which not a form of knowledge that inferentialism attempts to explain, so the vicious regress problem is avoided, even if memory is acquired inferentially.

This attempted refutation can be interpreted as merely reformulating the vicious regress problem into a more complicated issue of circularity. Although the inferential account of knowledge of internal promptings does not presuppose the knowledge it tries to explain, it presupposes other knowledge (e.g., background knowledge or memory) which may themselves possess the property of being inferentially acquired, a possibility Cassam himself acknowledges. If the self-interpretive process for acquiring internal promptings is inferential and is itself based on inferential evidence, then circularity is inevitable. Cassam tries to resolve this by asserting that "self-knowledge is holistic rather than linear, and the circularity its holism implies is genuine but not vicious". He outlines this circularity as follows: "You interpret your standing attitudes in light of the justified beliefs about your feelings and emotions, you interpret your feelings and emotions in light of further justified beliefs about your recent mental life, but your recent mental life includes standing attitudes your access to which was interpretive. A kind of interpretive circle in which each element depends for its significance on other elements of the circle" (Cassam, 2014). He claims there is nothing wrong with this circularity so long as the interpretive circle is "wide enough". It is unclear what "wide enough" means, but presumably one that prevents psychological self-knowledge being inferred from the very same or "too closely related" psychological self-knowledge, however that is defined. Cassam invokes the example of Katherine to demonstrate "there is no reason to think that this trick can't be pulled off" (2014). Referring to this as a "trick" to be "pulled off", as opposed to a definitive property of interpretive processes that avoids infinite circularity, is not compelling in generalizing this model to standing attitudes at large.

This is not to say the inferential model is incorrect, or that Cassam's argument is inherently flawed. In fact, his insight that the inferential process as holistic rather than linear is necessary to tackle the problem. Cassam's description of circularity aims to address the question of utilizing

only internal promptings to define other standing attitudes or internal promptings. However, it is worth considering whether this question is sensible to begin with. Internal promptings are inherently contextual – Katherine's feelings cannot be described as simply "yearning", but a yearning *of something*. Her feelings are grounded in the context of wanting another child, and the subject of her yearning feelings is derived from the current circumstances of her life, outside of *but not separable from* her mental life or internal promptings. More broadly, it does not make sense to consider the repeatedly inferential identification of standing attitudes confined solely within an individual's mental life and disregarding behavioral and circumstantial information about events in the real world which, as Cassam mentioned, are acquired totally differently from psychological evidence. The holistic self-interpretive process of internal promptings relies not just on one's recent mental life, but also recent circumstances and behavior which are fundamentally integrated with said internal promptings. Such a question reveals an implicit commitment to the notion of possessing "inner private mental states" that can be contained and kept separate from events outside the mind, which is unfounded. Unseating this assumption involves turning one's focus outwards and identifying external contextual and circumstantial factors that fundamentally influence one's internal promptings, or to use Cassam's analogy, are necessarily elements in the interpretive network. Recognizing the integration of a variety of evidence accessed through different means – behavioral, psychological, circumstantial, or other – upon which one engages in inferential reasoning, is crucial to realizing concerns about the circularity problem are unwarranted.

In conclusion, this paper assesses the plausibility of two potential pathways to substantial self-knowledge: the perceptual model and the influential model. It first determines the underlying principle of the perceptual model – that mental states can be represented as mental objects which can be identified and discriminated between – is flawed. In assessing that much of genuine perception is inferential in nature, this paper proceeds to defend an inferential model of self-knowledge, by proposing a positive account through arguments by elimination and by example and evaluating two major counterarguments: asymmetry and vicious regress. The main advantage of the idea that inferential reasoning is always involved in acquiring substantial self-knowledge is that it provides a unified account about the epistemologies of our beliefs, desires, and attitudes. Therefore, the difference in how "immediate" or "self-intimating" certain forms of self-knowledge are over others lies in the degree of how obviously inferential that knowledge is, rather that the difference between inferential and non-inferential knowledge.

Bibliography

Paul A. Boghossian (1989). "Content and Self-Knowledge." *Philosophical Topics*, vol. 17, no. 1, pp. 5–26.

Peter Carruthers (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Quassim Cassam (—) *Self-Knowledge: A Beginner's Guide*. http://www.self-knowledgeforhumans.com/uploads/3/9/1/1/39118991/ebook_-_self_knowledge_-_a_beginners_guide.pdf

Quassim Cassam (2014). *Self-Knowledge for Humans*. Oxford: Oxford University Press

David Finkelstein (2012). 'From Transparency to Expressivism'. *Rethinking Epistemology.*

Sigmund Freud (1915). *The Unconscious*. Pelican (Penguin) Freud Library, Vol. 11.

Krista Lawlor (2009). "Knowing What One Wants." *Philosophy and Phenomenological Research*, vol. 79, no. 1, pp. 47–75.

Richard Moran (2001) *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press

Martha Nussbaum (1990). *Love's Knowledge: Essays on Philosophy and Literature*. Oxford University Press.

Gilbert Ryle (1949) *Concept of Mind*. Hutchinson/Routledge, ch. 6

Jean-Paul Sartre (1981). *Existential Psychoanalysis*. Washington D.C.: Regnery Publishing, Inc.

Sydney Shoemaker (1996). 'The Royce Lectures: Self-Knowledge and "Inner Sense"', in *The First-Person Perspective and Other Essays*. Cambridge University Press: 201-68

Endnotes

[i] Proponents of the perceptual model argue that the supposed immediacy and self-intimating nature of mental states is afforded by this model of knowing one's own mental states through the direct, non-inferential perception of mental objects. However, it is worth considering whether immediate and self-intimating necessarily implies non-inferential, as opposed to a fast, automatic inference, but an inference nonetheless. Certain criticisms of this model also make similar assumptions, like the neo-Humean argument which contends that "if the content of a thought is determined by its relational properties, then one would refute the object perception model on the basis that it's not possible to ascertain an item's causal properties non-inferentially and hence perceptually" (Cassam, 2014). However, being unable to ascertain an item's relational properties non-inferentially does not imply one cannot ascertain its relational properties perceptually. Distance is a perceived property that is inferential, perceived by means of visual cues from which one can infer a target object's location. Also, properties like color and shape could be causal or dispositional, but clearly one can know by seeing that an object is red or square, hence causal properties can be determined perceptually.

[ii] A variation on the asymmetry argument is the argument from authorship, namely that individuals are authors of their own attitudes, hence their access to these authored attitudes must be non-inferential. While Katherine is said to 'find herself' wanting another child, Boghossian (1989) points out that there are many cases in which beliefs and attitudes are actively shaped. In cases where desires are formed as a conclusion of rational deliberation rather than discovering an unconscious desire that already existed, some argue that in authoring one's own attitude you do not infer from evidence that you know it. A simple rebuttal is that arriving at this knowledge, or "making up one's mind", occurs through evidence-based deliberation. It is unclear that shaping an attitude and thereby knowing that attitude is non-inferential, seeing as determining, for instance, whether one wants to go to Italy would be based on behavioral, psychological, contextual, or other evidence. Therefore, there is a logical gap in moving from the statement 'we shape our own attitudes' to the statement 'we know our own attitudes non-inferentially'.

[iii] Some may argue that surely there are examples of non-interpretive and non-inferential internal promptings. A common counterexample is sensations like nausea and pain, which do not seem to require any interpretative effort. This is not an issue for inferentialism as self-knowledge of simple sensations is not the sole basis of other substantial self-knowledge. Shedler (2009) describes an instance where a patient would describe the "random, unexplainable" onset of physical symptoms. This turned out to be an unconscious manifestation of her desire for her husband's attention, which could be determined once contextual factors were considered, specifically that in her childhood, her parents only gave her attention when she was ill. Thus, only by discovering the causes of and circumstantial information surrounding sensations can one then use them as evidence for inference, so it is genuinely interpretive. It is also worth considering whether sensations are apparent to us without any interpretive or cognitive effort.

Another common example is inner speech. All speech needs to be interpreted before it can be understood, but this may not seem right for inner speech since our own utterances are not ambiguous to us in the way others' utterances might be. However, for instance Katherine's utterance "I want another one" is only obvious to her because of her circumstances, memories, and mental images. There is no ambiguity to her not because she has non-interpretive access to her utterance, but because *it is obvious to her how to interpret her utterance***,** or what Finkelstein (2019) might refer to as her co-expressive gloss on her expression. Viewed in isolation her utterance means very little, but her *knowledge of the context of the utterance* makes it possible for her to interpret it.