

On the Opacity of Mental Contents

People constantly attribute mental states (e.g. beliefs, desires, sensations, etc.) to others to reliably predict and explain their behavior. For instance, when crossing the road on my way to work, I assume that drivers on the road believe that if they do not slow down, then they will run into me, and that they desire not to run into me. My reason for going to work is that I believe that if I go to work, I will get paid, and that I desire to get paid. But how do these invisible reasons (beliefs, desires, and other mental states) which have no apparent physical location nevertheless exert a causal influence on events in the physical world? Answering this question is a first step towards understanding how ‘reasons’ can be ‘naturalized’ or explained in a way consistent with the physical sciences.

Beliefs and desires involve representations of the world; having a belief involves representing (or misrepresenting) the actual state-of-affairs, whereas having a desire involves representing a desired state-of-affairs. For instance, my belief that *drivers on the road believe that if they do not slow down, then they will run into me* is about (i.e. represents) the mental states of drivers on the road. My belief that *the road is long* is a belief about the road. Similarly, my desires are about (and hence, represent) desired states of affairs. Thus, beliefs and desires are characterized and distinguished by what they are about, or in other words, what they represent. Given that ‘aboutness’ (i.e. intentionality) is a defining feature of beliefs, desires, and other mental states, a first step towards naturalizing reasons is to provide a theory of intentionality and content. How can our mental/brain states be ‘about’ objects in the world? And what makes a mental state about, say, cars, rather than cows?

Teleological theories attempt to naturalize mental content (and thereby answer the latter questions) by holding that the contents of mental representations are determined by functions—typically, functions of the systems that produce or consume said representations (Schulte & Neander, 2022). A canonical example of this type of theory is Ruth Millikan’s (1989, 2009) biosemantic theory, which posits that representations have the contents that they do in virtue of the fact that the instantiation of those representations contributes to the ‘proper functioning’ of the system that ‘consumes’ them. On this view, having content (and hence, being a representation) involves having a state that has been selected to serve the proper function of the consumer.

Teleological theories of mental content provide a plausible, naturalistic account of intentionality and representation, but they face a powerful objection from Frank Jackson

(2006), who argues that the folk (i.e. common people) are justified in attributing intentional contents to each other, but not 'selectional histories'. As a folk person, I am justified in attributing beliefs to drivers on the road, but not selectional histories (i.e. facts about the circumstances that held when the driver's various representations contributed to their proper functioning). Given this difference—namely, that folk are justified in attributing mental contents but not selectional histories—it cannot (by Leibniz's law) be the case that states with such-and-such contents are identical to states with so-and-so selectional histories (pg. 88). (Or so the objection goes).

In this paper, I defend teleological theories of mental content against Frank Jackson's (2006) compelling folk epistemology objection (FEO for short). FEO presupposes that the contents of our beliefs must be transparent to be justified. But this, I argue, is far too strict; drawing inspiration from Horgan (1984), I argue that, contrary to common sense, beliefs are (in the relevant sense) *always* composed of opaque contents. As implied by teleological theories of mental content, attributing a belief or desire to somebody involves attributing a state with so-and-so selectional history, and folk people are justified in doing so.

In part one, I briefly elaborate on and motivate teleological theories of meaning, using Millikan's (1989, 2009) biosemantic theory as a canonical example. In part two, I explain Frank Jackson's objection to teleological theories of meaning. Finally, in part three, I defend teleological theories of meaning by arguing that Jackson's assumption that the contents of our beliefs must be transparent to be justified (which his argument hinges upon) is false.

Part One: Teleological Theories of Mental Content

Teleological theories of mental content can be understood as a way of departing from (and thereby avoiding the problems of) causal theories of mental content which explain intentionality by positing a causal relationship between mental representations and objects or features in the world (see Fodor 1987 for a canonical example). Here, I provide a general description of causal theories of content before showing how they are threatened by the 'disjunction problem'. In turn, I show that teleological theories of content yield a decisive advantage due to their solution to the disjunction problem.

The basic idea of causal accounts of mental content is that thoughts about X's are about X's because X's cause those thoughts. At first glance, causal theories provide an elegant and naturalistic account of intentionality. But upon further inspection, all causal theories are threatened by their inability to distinguish content-determining and non-content determining causes of mental representations. For instance, according to causal theories, my dog representation is about dogs because it is caused by dogs. But my dog

representation is also caused by coyotes on dark nights and electrical stimulation in certain parts of my brain. Fodor (1984) calls this the 'disjunction' problem, because a consequence of the causal theory of content is that my dog representation is not only about dogs, but also coyotes on dark nights and electrical stimulation in certain parts of my brain. Why is my dog representation about dogs, and not dogs *or* coyotes on dark nights *or* any of the other causes of that representation? Fred Dretske (1988) and Jerry Fodor (1987) provide solutions to the disjunction problem, but both fail to provide a principled way of distinguishing between content determining and non-content determining causes.

Dretske's (1988) solution to the disjunction problem is to claim that the content of a representation is determined by the cause of that representation during a learning period. In his words, where C is an internal state or representational vehicle, M is some behavior or result, and F is a state-of-affairs in the world,

Once C is recruited as a cause of M—and recruited as a cause of M because of what it indicates about F—C acquires, thereby, the function of indicating F. (pg. 84)

Thus, again, the content of a representation is determined during its initial recruitment or learning process. A representation C indicates (and hence, means) that some state of affairs F holds because C serves its function in virtue of its indicating F during a learning phase. On this view, dog representations are about dogs (and not Coyote's on dark nights) because those representations were caused by dogs during the initial learning period. Though promising, this view fails to solve the disjunction problem as it provides no reason to think that Coyote's on dark nights would not cause 'dog' representations during the learning period.

Alternatively, Fodor's (1987) solution is to claim that the content of a representation is determined by a pattern of asymmetric dependence between the representation and the represented state-of-affairs. The core idea is that a mental state S represents some state-of-affairs X if and only if (1) and (2) hold:

1. X reliably causes S.
2. If a distinct state of affairs Y also causes S, then Y would not have caused S if X did not reliably cause S.

So, for instance, the reason a dog representation is about dogs is that it is 1) reliably caused by dogs and 2) anything else that causes (e.g. coyotes on dark nights) it would not cause it if it was not reliably caused by dogs. But this theory faces a similar problem as Dretske's. It is not clear what fixes these laws in the first place. Why couldn't my dog representation be reliably caused by coyotes on dark nights before I had ever seen a dog? In that case, my dog representation would actually be about coyotes on dark nights because, even though it

might be caused by dogs, dogs would not have caused it if it was not first caused by coyotes on dark nights.

Finally, Millikan's (1989) teleological theory of content provides a more promising solution to the disjunction problem by shifting the focus from the producer of representations to the consumer. In other words, whilst Dretske (1988) and Fodor (1987) focus on how the cause of representations determines their content, Millikan (1989) focuses on how content is determined by the consumer which makes use of those representations. She writes that, "It is the devices that use representations which determine these to be representations and, at the same time (contra Fodor), determine their content" (pg. 283-284), and that, "the consumer part of the system ought to be all that is needed to determine not only representational status but representational content" (pg. 286). Hence, given her claim that the consumer is sufficient for determining content, Millikan avoids committing to the view that for a sign to represent something, there must be a natural connection between a representation and its represented state-of-affairs. Instead, on Millikan's view, content is fully determined by the normal conditions necessary for the consumer's performance of its proper functions, where 'proper functions' are functions "determined by the histories of the items possessing them", with functions that are 'selected for' being paradigm cases (pg. 284). For instance, hearts have the proper function of pumping blood because they have historically been selected for that 'purpose.'

How, then, is the content of a particular representation determined by the normal conditions necessary for the consumer's performance of its proper functions? In Millikan's later (2009) paper, she clarifies how consumers determine the content of representations produced, writing,

the producer produces a sign that will be true or satisfied only if it maps on to some affair or affairs [...] in the world in accordance with certain 'semantic' rules. These are rules of correspondence between signs and world affairs that have been instantiated in the past when the consumer and producer or their ancestors have succeeded in performing their cooperative functions. (397-398)

In other words, 'correspondence rules' (or 'normal mapping rules' to use Millikan's (1989) terminology) are rules which describe some relation between representations and represented affairs, and which are necessary for providing a 'normal explanation' of how the producer and consumer have historically cooperated to perform their proper function. Note here that 'normal explanations' describe how proper functions were "(typically) historically performed on those (perhaps rare) occasions when it was properly performed" (Millikan, 1989, pg. 284). This point is important because it accounts for the fact that proper functions can be performed accidentally and that it is possible for something to rarely help

fulfill its proper function even if that is what it has been selected for. Thus, normal explanations are not statistical explanations.

Returning to the disjunction problem, Millikan's teleosemantic theory provides a compelling explanation of why dog representations are about dogs and not coyotes on dark nights or electrical stimulation in a certain part of the brain. Presumably, humans have historically been able to perform the proper function of avoiding predators partly in virtue of a correspondence rule between some 'coyote' representation (or system of representations) and actual coyotes out there in the world. We need to posit such a correspondence rule between our coyote representations and the actual occurrence of nearby coyotes to provide a normal explanation of how humans have been able to avoid coyotes in the past. On the other hand, even if there is a correspondence rule between dog representations and coyotes on dark nights, such that if a dog representation is tokened then there are nearby coyotes, it does not follow that the dog representation has the content of representing coyotes on dark nights. This follows from the fact that such a correspondence rule would not enter into a normal explanation of how humans have historically performed their proper function of avoiding coyotes.

Admittedly, the latter explanation of how representations get their content is a bit too coarse grained; coyote representations may have been 'selected for' more than their ability to induce avoidance behavior when in the presence of coyotes. Further, I have been using the phrase 'selected-for' in a slightly ambiguous way. The evolutionary sense of the word 'selection' suggests a slow process which occurs over countless generations, but humans can presumably form completely novel representations over the course of their lifetime. Hence, the sense of the phrase 'selected-for' which accounts for the content of representations must be more fine grained. On this notion of selection, a representation gets selected for because of its ability to aid the producer of representations with its proper function of aiding the consumer with its proper functions. Perhaps, then, proper functions are best understood in terms of a nested hierarchy of goals, where the higher goals are selected for at a coarser grained, generational level, and the lower goals are selected at a finer grained, lifespan level. For present purposes, I will use the term 'selection' in broader sense; something can be 'selected for' on the fine-grained level of an individual lifetime. This means that if something is selected for, then it has been reproduced or sustained because of its ability to aid the consumer with its proper functions.

Details aside, the brief explanation of Millikan's teleological theory of content presented above should suffice for present purposes. The crucial point is that the content of representations is ultimately determined by the proper function of the consumer; the content of a particular representation is given by a correspondence rule between the representation and some state of affairs which has held in instances where the consumer

has historically succeeded in normal performance of its proper function. Hence, having a state with a representational content is having a state that has been selected for some purpose. This section has attempted to flesh out what a mature teleological theory of content looks like so that it is clear what is at stake. In part two, I elaborate on Frank Jackson's powerful objection to teleological theories of content, before replying to the objection in part three.

Part Two: Jackson & the Problem of the Folk

Frank Jackson (2006) worries that teleological theories of mental content are threatened by what folk people are justified in saying about the content of their own mental representations. For instance, I am justified in asserting that drivers on the road have mental representations of me; this is what explains their slowing down as I cross the road. Similarly, I am justified in thinking that certain of my own mental states are about the drivers and cars on the road. As Jackson writes, "After you've seen someone walk through a minefield, you have a pretty good idea where they think the mines are located [and] after you've dined regularly with someone, you have a pretty good idea of what they like to eat and drink" (pg. 85). But, on the other hand, Jackson argues, folk people are not justified in attributing evolutionary histories (i.e. 'selectional histories') to one another. For instance, most folk people have no idea why noses and chins were selected for, let alone what our particular mental representations were selected for. It follows that if teleological theories of content imply that attributing content to someone is attributing a selectional history to them, then teleological theories are false. Folk people are justified in attributing content to one another, but not selectional histories.

Here, I elaborate on the formal structure of Jackson's 'folk epistemology objection' (FEO) and the motivation for each premise. Further, I show that Millikan's theory (as described above) is a teleological theory in the sense relevant to Jackson's argument. Jackson formalizes his argument as follows:

(P1) Having an intentional state with such and such content is a property justifiably ascribed in so and so circumstances.

(P2) Having a state, or being in a state, with so and so a selectional history is not justifiably ascribed in so and so circumstances.

(C1) Therefore, having an intentional state with such and such content is not (by Leibniz's law) identical with being in a state with so and so a selectional history.

(pg. 88)

This captures the core of FEO, but not the conclusion that teleological theories of content are false. Adding Jackson's argument for the falsity of teleological theories of content yields an extra premise and conclusion (P3 and C2):

(P3) Teleological theories of content (e.g. Millikan's theory) imply that having an intentional state with such and such content is identical with being in a state with so and so a selectional history.

(C2) Therefore, teleological theories of content are false (By C1 and P3).

The reason I explicitly add (P3) is to connect the relevance of FEO to Millikan's theory of content. If Millikan's theory is a teleological theory in Jackson's sense, then the above argument provides a compelling reason to think that Millikan's theory is false. Indeed, so long as we allow that 'selection' can occur at a fine-grained level, Jackson's objection directly applies to Millikan's theory. Recall that, for Millikan, the content of a particular representation can be determined (and hence, 'selected for') during the course of a lifetime. Further, having a state with some content is identical to being in a state that has been historically selected for some corresponding function. So yes, Millikan's theory is a teleological theory of content in the sense that, on her view, "having an intentional state with such and such content is identical with being in a state with so and so a selectional history." Given that the argument is valid and that (P3) is not under dispute, if there is an issue with the argument, then the problem is either (P1) or (P2). Here, I consider each of these premises, in turn.

Premise one (P1) is undeniably true for many substitutions of 'so-and-so circumstances'. In other words, there are many circumstances in which it is justifiable for a folk-person (A) to attribute an intentional state (S) to a distinct agent (B). One question is, then, in which types of circumstances are content attributions justifiable? Providing a complete, principled answer to this question is difficult, and somewhat beyond the scope of the present paper. Nevertheless, there are a few paradigmatic examples of circumstances in which (P1) is true. If A watches as B successfully picks apples from a tree, then it is typically justifiable for A to attribute to B the belief that there are apples in the tree, and the desire to acquire the apples in the tree. Similarly, if, while crossing the road, A makes eye contact with a driver B, then A is justified in attributing to B the belief that A is crossing the road. Now, it is possible for new information to undermine the justification of A's attribution of S to B; suppose the apple picker does not appear to be coordinated and is merely swinging her arms and accidentally knocking a few apples off of the tree, or that the driver is heavily intoxicated, swerving back and forth on the road. However, for present purposes, what matters is not that there are circumstances in which content attributions are unjustified, but rather that in many circumstances content attributions are justified. These are the circumstances referred to by 'so-and-so circumstances' in (P1) and (P2). Given this clarification, (P1) is uncontroversially true.

(P2), on the other hand, is contestable, though Jackson motivates it with the following sort of argument:

(P2A) Observations of subjects interacting with their environment are insufficient justification for attributing selectional states (i.e. states selected to co-vary with such-and-such, or states selected to bring about so-and-so) to subjects (paraphrased from Jackson, 2006, pg. 87).

(P2B) But if folk are justified in attributing selectional states to people, then observations of subjects interacting with their environment are sufficient justification for the attribution of selectional states.

(P2) [Paraphrased] Therefore, folk are unjustified in attributing selectional states to people.

Indeed, (P2B) follows straightforwardly from the evidence available to the folk. Since most folk people do not study evolution nor psychology, the only evidence of each other's selectional states available to the folk is observations of each other's interactions with the environment. (P2A) is also *prima facie* plausible. Folk people do not, *qua* folk, have opinions (justified or otherwise) about what structures are selected for. And indeed, as Jackson points out, "the selectional histories of the ear and the larynx are major research topics" (pg. 87). If folk are justified in attributing selectional states to people, then (for similar reasons) folk are justified in their opinions about the selectional histories of the ear and larynx as well. But folk are not justified in their opinions about the selectional histories of the ear and larynx. Given that (P2A) and (P2B) are both *prima facie* plausible, what could possibly be wrong with Jackson's argument?

In part three, I undermine Jackson's argument by objecting to (P2). (P2A) and (P2B) fail to support (P2), as the argument equivocates two senses of 'state-attribution'.

Part Three: The Opacity of Mental Contents

In his classic (1986) paper entitled *What Mary Didn't Know*, Jackson argues that it is possible to know all of the physical information about color experiences and yet not know what it is like to experience color. Thus, he concludes, some information is not physical information, and not everything is physical. Objecting to this argument, Terence Horgan (1984) contends that Jackson equivocates two senses of 'physical information'—explicitly physical information and ontologically physical information. According to Horgan, it is possible to know all of the 'explicitly' physical information about color experiences without knowing all of the 'ontologically' physical information about color experiences (pg. 151). Here, I take inspiration from Horgan, claiming that Jackson's more recent (2006) argument equivocates two types of state-attributions—explicitly selectional state-attributions and ontologically selectional state-attributions. Folk are often justified in making ontologically selectional state-attributions, but not explicitly selectional state-attributions.

Explicitly selectional state-attributions are statements of the form, "X is in a state that has been selected to vary with (or bring about) P," where X is an agent and P is some

state of affairs. By contrast, statements of the form, “X desires/believes/hates that P” are merely ontologically selectional state-attributions. Generally, a statement about some agent’s mental states is an explicitly selectional state attribution if and only if it is formulated using the theoretical language of teleological theories of meaning. On the other hand, a statement about some agent’s mental states is an ontologically selectional state-attribution if and only if it attributes selectional states to that agent. Consider, for example, the statement that, “drivers believe that I am crossing the road.” If teleological theories of meaning are correct, then that statement entails (*necessarily* and *a posteriori*) that drivers are in such-and-such a selectional state. Nevertheless, folk do not, *qua folk*, know about this entailment relation; hence the reason that folk can be justified in making an ontologically selectional state-attribution and yet unjustified in making an explicitly selectional state-attribution.

The distinction between explicitly selectional state-attributions and ontologically selectional state-attributions poses a problem for FEO because Jackson’s argument for (P2) does not address state attributions that are merely ontologically selectional. I contend that (P2A) is clearly false under the relevant sense of state-attribution. Precising (P2A) accordingly yields (P2A*).

(P2A*) Observations of subjects interacting with their environment are insufficient justification for **ontologically selectional attributions** of selectional states to those subjects.

Consider the following counterexample: As a folk person, my statement that “drivers believe that I am crossing the road” is justified, and yet my statement that “drivers are in such-and-such a selectional state” is not. Why? Because my beliefs about the mental states of drivers on the road have reliably helped me (and, more generally, the folk) get across the road in the past. Frankly, if my belief attributions did not reliably track other people’s mental states, then I would be largely incapable of performing my proper functions and either severely disabled or dead. Contrastingly, explicitly selectional attributions of selectional states to people have not historically contributed to our proper functioning. Hence, folk are not, *qua folk*, justified in making explicitly selectional attributions of selectional states to people.

In light of the fact that (P2A*) is false, I have precisified (P2) to yield the more defensible (P2*), reformulating FEO as follows:

(P1*) **Ontologically selectional attributions** of having an intentional state with such and such content are justifiable in so and so circumstances.

(P2*) **Explicitly selectional attributions** of having a state, or being in a state, with so and so a selectional history are unjustifiable in so and so circumstances.

(C1) Therefore, having an intentional state with such and such content is not (by Leibniz's law) identical with being in a state with so and so a selectional history.

Notice that, given this more precise characterization, FEO is invalid (C1 does not follow from P1* and P2*). According to Leibniz's law, if two objects differ in *any* property, then they must be distinct objects. But with the distinction between types of state attributions in place, it is clear that the states attributed in (P1*) and (P2*) do not violate this constraint. Hence, FEO fails to support the conclusion that having intentional states is not identical with being in states with certain selectional histories. There is still hope for teleological theories of content.

Recall my initial claims that Jackson's FEO relies on the assumption that the contents of our beliefs must be transparent to be justified, but that (in actuality) beliefs are always composed of opaque contents. But what does it mean to say that a belief is opaque or transparent? And how does this distinction relate to FEO? Opacity is a measure of epistemic access. Thus, the opacity of a mental state is a measure of epistemic access to features of that mental state. For present purposes, we can safely define epistemic access in terms of reportability. A feature or property of a mental state is *first-person* reportable if and only if the person possessing that mental state is capable of accurately reporting that they have that feature or property. On the other hand, a feature or property of a mental state is *third-person* reportable if and only if individuals other than the individual possessing that feature or property can reliably report that the individual possessing that feature or property does indeed have that feature or property. Finally, a feature or property of a mental state is reportable simpliciter if it is either third-person or first-person reportable. That I possess a belief that it is raining is transparent to me because I have epistemic access to it. That is, I can accurately report whether I believe it is raining by saying, "I [do/don't] believe it is raining."

Now, Jackson's FEO argument seems to be partially motivated by the intuition that beliefs are transparent, and that this grounds (P1). Spelling this out, the fact that we can reliably report on other's mental states is what grounds the claim that having an intentional state with such and such content is a property justifiably ascribed in so and so circumstances. It is true that we are often justified in our attributions of beliefs, desires, and other mental states to others, and this is unsurprising given that it is transparent to us that others have those mental states. But here is where Jackson makes a critical mistake; he fails to recognize that it is possible for it to be transparent *that* someone possesses a certain mental state without it being transparent *what* the content of that mental state *is*. Indeed, it is possible to (somewhat) reliably report on whether drivers believe that I am crossing the road, even if I do not know what the content of that belief is. Again, I can be justified in attributing beliefs to drivers on the road because doing so reliably helps me get

across the road. Every day, when I cross the road, I bet my life on the reliability of my judgments about driver's mental states. Nevertheless, I need not have a theory of content to do so. Hence, beliefs and other mental states are opaque; I need not know what their content is to know that I have them.

Conclusion

In part one, I motivated Ruth Millikan's (1989, 2009) biosemantic theory of content (and teleological theories of content more generally) by comparing its solution to the disjunction problem to prominent causal theories of content. Part one sets up my explanation of Frank Jackson's (2006) folk epistemology objection (FEO) against teleological theories of content in part two. Finally, in part three, I draw from Terence Horgan (1984) to defend teleological theories of content against Jackson's FEO. Ultimately, I argue that Jackson commits a fallacy of equivocation by failing to distinguish between ontologically selectional and explicitly selectional attributions of mental states. Relatedly, I argue that Jackson's assumption that the contents of our beliefs must be transparent to be justified is false.

Many beliefs are justified simply because they help us get around the world; we need not have a theory of content for our beliefs to prevent us from getting run over by a car. But perhaps teleological theories of meaning such as Millikan's biosemantics are the best explanation of how/why our beliefs are reliable. Stepping back, this paper is a step towards understanding how beliefs, desires, and other mental states could possibly be physical things. If mental states are selectional states, then they pose no serious problem for a physicalist account of the mind. Furthermore, if mental states are selectional states, then psychology can be productively grounded in biology.

Sources Consulted

Dretske, Fred I. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press. Edited by Margaret A. Boden.

Horgan, Terence E. (1984). Jackson on physical information and qualia. *Philosophical Quarterly* 34 (April):147-52.

Jackson, Frank (2006). The epistemological objection to opaque teleological theories of content. In Graham Macdonald & David Papineau, *Teleosemantics: New Philosophical Essays*. New York: Oxford: Clarendon Press. pp. 85--99.

Jackson, Frank (1986). What Mary Didn't Know. *Journal of Philosophy* 83 (5):291-295.

Millikan, Ruth Garrett, 'Biosemantics', in Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter (eds), *The Oxford Handbook of Philosophy of Mind* (2009; online edn, Oxford Academic, 2 Sept. 2009)

Millikan, Ruth Garrett (1989). Biosemantics. *Journal of Philosophy* 86 (6):281-97.

Schulte, Peter and Karen Neander, "Teleological Theories of Mental Content", *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), URL = [<https://plato.stanford.edu/archives/sum2022/entries/content-teleological/>](https://plato.stanford.edu/archives/sum2022/entries/content-teleological/).