

Of Mice and Madmen: Typical Causal Roles and Radical Indeterminacy

Edward Elliott

University of Notre Dame

Abstract

Following Lewis (1980), many functionalists hold that functional roles should be understood as kind-relative typical roles—e.g., for the typical person, pain causes flinching, though for some individuals it might not. However, an indeterminacy problem arises if the relevant notion of *typicality* is cashed out in statistical terms. In this paper, I show how the problem arises in the context of Lewis' analytic functionalism. Under reasonable assumptions, Lewis' position implies that for each individual and any coherent attribution of mental states to that individual, there are likely to be a maximally fitting schemes of interpretation according to which that individual has those mental states. The problem, in other words, is not just one of *radical* indeterminacy, but *maximal* indeterminacy: anyone can be interpreted as having any internally consistent system of beliefs, as having any desires whatever, as being in pain or in pleasure, and so on. Any functionalist solution to this problem requires some changes to how the constraints on fittingness are understood. I discuss some possible fixes, including a new potential role for naturalness within the Lewisian theory.

1. Introduction

George lives in New York City, while Lennie lives in Los Angeles. The two cities have enormous populations—about 9 and 4 million respectively. And due to those very large population sizes, neither George nor Lennie make any appreciable difference to what's *typical* for their respective cities. Imagine they were to trade places, holding everyone else fixed; at least in the short term, swapping them one-for-one in this manner would have negligible impact on the city-wide typicality facts. Lennie doesn't sound like a New Yorker, for instance, but his suddenly being in NYC makes no difference to what we'd call the typical New York accent. Or suppose that Lennie is tall while George is short; still, the difference that swapping them would make to the average height of New Yorkers and Los Angelenos is minuscule—probably a few dozen nanometers, and in any case nothing more significant than the expected day-by-day fluctuations in those averages due to the constant stream of births, deaths, and migration into and out of either city.¹

1. I did check this. Given a very substantial height difference of 2 feet, then the effect that swapping Lennie and George would have on average height in LA is 2 feet divided by 4 million, or ≈ 150 nanometers. For NYC, that reduces to ≈ 70 nanometers.

Let's say that Lennie and George are interchangeable *salva typicitate*—that is, without making any significant difference to what's typical.² Of course, there's nothing that's unique or special about Lennie and George in this respect; the same is going to be true of any other pair of residents from Los Angeles and New York City, or most of them anyway. In general, the facts about what's typical for large populations will tend to be robust against minor variations in the exact makeup of those populations, and increasingly so as those populations get larger. Consequently, for very large populations like those of NYC and LA, we can reasonably expect that for most (if not all) pairs of individuals, one taken from each city, that pair will also be interchangeable *salva typicitate*.

I'm going to argue that this kind of *insensitive* dependence of large-population statistics on the individuals that comprise those populations generates an indeterminacy problem for David Lewis' analytic functionalism.³ That is, the Lewisian account implies that for each individual and any coherent attribution of mental states to that individual, there is likely to be a maximally fitting scheme of interpretation according to which that individual has those mental states. The problem, in other words, is not just one of *radical* indeterminacy, but *maximal* indeterminacy: anyone can be interpreted as having any internally consistent system of beliefs and desires whatever, as being in pain or in pleasure, as experiencing blue or red or green, and so on. I'll need to spell out why all this follows in more detail, of course, but ultimately the problem arises because Lewis' recipe for constructing functional definitions has us characterizing individual-level mental facts (i.e., whether this or that individual is in a given mental state) primarily in terms of what's statistically normal for that individual's population (i.e., whether they're in a state that plays a certain typical cause role), even though the token states those individuals happen to be in are interchangeable *salva typicitate*.

I do not put this argument forward as a reason to reject analytic functionalism! It is absolutely not the kind of thing that should cause anyone to substantially lower their credence in what is (to my mind) one of the best approaches we have to the naturalistic characterization of our mental states. Instead, the argument belongs to a fairly wide class of what we might call *devil-in-the-details problems*, which target the precise manner in which functionalist definitions might be formulated. The intended lesson is that we need to refine the common formulation of the Lewisian theory in one way or another so as to avoid the bad consequence, while also hoping we don't create yet further problems down the road in doing so. Indeed, the present problem of indeterminacy arises for the Lewisian account as a direct consequence of how he tried to solve several other devil-in-the-details problems for his account. More on that later.

2. There appear to be no near synonyms for 'typical' in classical Latin, so I've taken some slight liberties with the language. 'Salva typicitate' is intended to be reminiscent of 'salva veritate', which refers to when expressions can be interchanged without altering the truth value of the statements in which the expressions occur.

3. A similar indeterminacy problem can arise for other varieties of functionalism as well, but I'm focusing on the Lewisian theory for two reasons: (1) there is significant historical interest in clarifying the details of Lewis' distinctive brand of functionalism and its consequences; and (2) whether and to what extent the problem arises, and what counts as a viable solution to it, depends in part on the specific commitments attached to the particular version of functionalism under consideration—and discussing all this from every angle would be cumbersome.

Section 2 is groundwork, clarifying the different kinds of indeterminacy problem that can arise for functionalism. Section 3 then clarifies some matters in connection with the statistical conception of ‘typicality’ that’s central to the Lewisian theory. Following that, Section 4 shows how the indeterminacy problem arises for Lewis’ view inasmuch as we’re dealing with very large populations, while Section 5 argues that the relevant populations must be very large indeed. Finally, Section 6 canvases some potential fixes and the costs associated with them, including yet another potential role for naturalness within the Lewisian theory.

2. The Sources of Indeterminacy

I assume that folk psychology can be thought of as a kind of theory, and one that includes two kinds of hypotheses: first, folk psychology purports to tell us what types of mental states exist, or might exist; and second, folk psychology purports to tell us how these mental states relate to one another and to the non-mental world via our behavior and our senses. Some of these relations will be ontological—e.g., experiencing scarlet *is a way of* experiencing redness. Others will be causal, or more accurately they’ll be causal-statistical—e.g., experiencing scarlet *is normally caused by* the unobstructed presence of an object with such-and-such surface reflectance properties before the eyes under good lighting conditions. There may be other kinds of relations too, but it is presumed (or hoped, perhaps) that all of them together whatever they may be can ultimately be expressed in a non-spooky way, using concepts (like *causation* and *statistical normality*) that are readily amenable to a naturalistic metaphysics. We shouldn’t assume that this folk theory includes the all and only the truths of psychology, but it does have plenty to say and it’s not unreasonable to suppose that what it does say comes close to the truth in most cases. Close enough, at least, that we might use it to implicitly define the terms of our folk-psychological vocabulary.

That is Lewis’ analytic functionalism in a nutshell: that the folk mental-state terms, such as ‘belief’ and ‘desire’, ‘pain’ and ‘pleasure’, and so on, all designate whatever they have to in order to render folk psychology as near to true as can be. If we suppose—as I will be doing henceforth—that physicalism is correct, then the basic idea is to try to match up the physical states that agents might be in to the mental states posited by folk psychology, such that the former can be said to occupy the causal-statistical roles of the latter—i.e., such that the way those physical states normally relate to one another and to the rest of the world mirrors as close as can be the way their associated mental states normally relate to one another and to the rest of the world according to folk psychology.

All this is most often cashed out by means of Ramsey sentences, O-terms and T-terms and so on, but for present purposes it’ll be more helpful to formulate the idea in terms of *schemes of interpretation* (cf. Lewis 1983: 119–20). There’s a few ways we could try to make this notion precise, though the differences between them shouldn’t matter much. Let’s say that a scheme of interpretation is defined relative to a population—the human population, for example—and it tells us how to identify the various states the individuals of that population might be in with the different kinds of mental states posited by folk psychology. Given physicalism, we can thus let a scheme of interpretation for the human population, *i*, be any partial and not-necessarily surjective function from the set of all physical states *P* that we might

potentially be in, and the set of all mental states M that we might potentially be in according to folk psychology; we then read $i(P) = M$ as saying that if one of us were in P then that person would be in M , and vice versa.⁴ Furthermore, we say that a scheme of interpretation *fits* to the extent that it renders folk psychology close to the truth for the relevant population.⁵ A scheme is *perfect* whenever the function is surjective and $i(P) = M$ just in case P perfectly occupies the role of M . Note that a perfect scheme is possible *only if* folk psychology contains no falsehoods whatsoever. With all that in place, the analytic functionalist says that the *correct* scheme of interpretation is the most *fitting* scheme, if there is just one, and if there are multiple schemes tied for best fit then the truth is indeterminate between them (cf. Lewis 1983).

Indeterminacy is not inherently problematic. It would be surprising, in fact, if there weren't *some* degree of indeterminacy with respect to *some* of our mental states for at least *some* (if not most) of us. And I'm inclined to think that analytic functionalists can accept without embarrassment that there might be some rare cases of actual human beings with severely indeterminate mental states, or a possible population of strange non-human creatures the majority of whom have radically indeterminate mental states. These do not look like bullets to bite. We have a *problem of radical indeterminacy*, on the other hand, if we find ourselves having to say that the mental states of most actual humans are highly indeterminate—that's the kind of obviously bad result that functionalists need to avoid.

To see how a problem of radical indeterminacy might come about, it'll be useful to first distinguish the major *sources of indeterminacy*. I count five sources. One is *trade-off indeterminacy*. This might arise when there's no precise facts of the matter about how we ought to measure fit, and different ways of drawing compromises might give different results regarding which of the schemes should be considered the most fitting overall. Or it may be that we've settled on a precise method for measuring fit, but more than one imperfect scheme maximizes fit relative to that measure—e.g., if one scheme has good fit with respect to M_1 but poor fit with respect to M_2 , while an alternative and contradictory scheme has good fit with respect to M_2 but poor fit with respect to M_1 . I am, like Lewis, inclined to think that this kind of trade-off indeterminacy is both very likely to arise but also unlikely to generate any serious problem of radical indeterminacy, and for that reason can be safely ignored for most purposes. (See Lewis 1974: 343; 1983: 120)

A rather distinct source is *folk-theoretical indeterminacy*: where the content of folk psychology itself is indeterminate, then what counts as a maximally fitting scheme will likewise be indeterminate. We might try to deal with this by replacing

4. Why is the function partial? Because presumably not every physical state we might be in will be identified with some mental state. Why is it not-necessarily surjective? Because we shouldn't assume we'll be able to identify every possible mental state with some physical state.

5. More precisely: if a scheme assigns P_1, P_2, \dots to M_1, M_2, \dots respectively, then that scheme's degree of fit is a measure of how well those physical states *jointly* occupy the *combined* functional role of those mental states. An individual physical state, P_1 , can properly be said to occupy the role of a given mental state M_1 only in the derivative sense that if the P_1, P_2, \dots jointly occupy the combined functional roles of M_1, M_2, \dots , then P_1 's location in the former sequence corresponds to M_1 's location in the latter. The most fitting scheme of interpretation i may identify M_1 with P_1 even where P_1 conforms poorly to the role of M_1 when considered in isolation, if the P_1, P_2, \dots do a good job of occupying the combined role of M_1, M_2, \dots overall.

singular schemes of interpretation with sets of them—the set including a precise scheme of interpretation for each precisification of the folk theory—and then supervaluating over the lot of them to get an ‘imprecise’ scheme which can be compared for overall fit against other ‘imprecise’ schemes. (Lewis 2020 [1980]: 99–100 suggests something like this to deal with semantic indecision in folk-psychological vocabulary.) And yet another potential source still is *physical indeterminacy*: if it’s indeterminate what physical states an agent is in, then it’s going to be indeterminate what any scheme of interpretation says about that agent’s mental states. But again, neither folk-theoretical indeterminacy nor physical indeterminacy are especially likely to generate a problem of radical indeterminacy. They *might*, but I don’t see any compelling reason to think they *do*.

The source of indeterminacy that’s received the most discussion is what I’ll call *role indeterminacy*. This is where problems of radical indeterminacy are usually thought to enter the picture. One widely-discussed example of this concerns the underdetermination of an agent’s beliefs and desires by the facts about their behavior, or their behavioral dispositions conditional on their evidence. Systematic arguments are supposed to show that whenever we have some pattern of behavior that’s rationalized by the assignment of such-and-so beliefs and desires, we can always twist the assigned beliefs one way while twisting the assigned desires in a countervailing way so as to end up with a very different system of beliefs and desires that nevertheless explains the very same behavior—which is a problem if beliefs and desires are characterized primarily by reference to their causal role in the production of behavior.⁶ More generally, though, role indeterminacy arises whenever the combined functional role of one sequence of mental states M_1, M_2, \dots (e.g., systems of beliefs and desires) is indistinguishable from the combined functional role of some other sequence of mental states M'_1, M'_2, \dots (a different sequence of systems of beliefs and desires). In this case, any scheme i that assigns the states M_1, M_2, \dots to the physical states P_1, P_2, \dots will fit exactly as well as an alternative scheme i' that assigns M'_1, M'_2, \dots to P_1, P_2, \dots instead, holding all else fixed, since any such substitution will preserve all the relations that factor into fittingness. And so there can in principle be no *uniquely* best-fitting scheme of interpretation, even in the event that a perfect scheme exists.

But there’s still one more distinctive source of indeterminacy, and it’s the one that I’ll be talking about. Suppose for the sake of argument that there’s no trade-off, folk-theoretical, physical or role indeterminacy. It follows that if a scheme i is perfect, and if i assigns M_1, M_2, \dots to P_1, P_2, \dots respectively, then there’s no way to create an equally fitting scheme by assigning some other mental states M'_1, M'_2, \dots to P_1, P_2, \dots instead. However, it does not follow that i is uniquely perfect, since there may yet be another perfect scheme i' that identifies M_1, M_2, \dots with P'_1, P'_2, \dots rather than to P_1, P_2, \dots . This will occur whenever the P'_1, P'_2, \dots have the same pattern of relations to one another and to the rest of the world that the P_1, P_2, \dots have which make them well-suited for identification with M_1, M_2, \dots respectively. Let’s call this *occupant indeterminacy*.

6. For Lewis’ version of this underdetermination problem, see (1986: 37–8) and moreover (Lewis 1983: 374–5); see also (Williams 2016) for a detailed reconstruction of Lewis’ argument. For closely related discussion on the functional underdetermination of beliefs and desires, see (Stalnaker 1984: 15–20), (Davidson 1997), and (Dennett 1991).

Occupant indeterminacy is very similar to role indeterminacy, but it's happening on the input side of the interpretation function rather than the output side. That is, role indeterminacy arise when there are distinct (sequences of) mental states that have indistinguishable (combined) functional roles, and thus different schemes of interpretation can vary in which mental states they identify with a given physical state whilst sharing the same degree of fit; whereas occupant indeterminacy arises when there are distinct (sequences of) physical states that jointly occupy the same (combined) functional roles, and thus different schemes of interpretation can vary in what physical states get identified with a given mental state whilst sharing the same degree of fit. And because they're different problems, they require different solutions. For instance, Lewis (1983: 375; 1986: 38, 107; 1994: 416–7) proposes to solve the problem of role indeterminacy for beliefs and desires by imposing constraints on what kinds of intentional contents are possible—some contents are too *unnatural* to be eligible as potential objects of belief and desire. So if the contents of M_1, M_2, \dots are more eligible than the contents of M'_1, M'_2, \dots , then any scheme i that identifies P_1, P_2, \dots with M_1, M_2, \dots will be preferred over a scheme i' that identifies those physical states with M'_1, M'_2, \dots instead, *ceteris paribus*. This eliminates role indeterminacy, but it does nothing to remove occupant indeterminacy.

There are examples of occupant indeterminacy problems in the literature. Most notably, Putnam (1988: 120–5) has argued that every ordinary open system realizes every finite state automaton (FSA), which generates a problem of occupant indeterminacy for the early machine-state functionalisms of the 1960s and 1970s. The gist is that if the mind can be characterized functionally as a kind of FSA, then any physical system of sufficient complexity can be said to implement any such automaton; and this is ultimately because there are indefinitely many ways of grouping together the token physical states of the system into types with which the machine states of the FSA could be identified. Putnam's problem is similar to mine in that we both make central use of 'disjunctive' physical state types. However, Putnam's problem applies to machine-state functionalism, in part because it requires us to associate total mental states with total physical states—the problem seems to dissolve when we have the more complicated internal causal roles between mental states that analytic functionalism allows for (cf. Chalmers 1996: 323–6). For this reason, you might see the problem of occupant indeterminacy that I raise for analytic functionalism as distinct from Putnam's problem; or you might see it as a variant of the same general problem. I think either would be fair.

3. Causal-Statistical Roles in Lewis' Functionalism

The problem of occupant indeterminacy for Lewis' functionalism is mainly a consequence of two things: the characterization of mental states primarily in terms of their causal-statistical roles, and the enormous size of the population relative to which those causal-statistical roles are defined. The present section discusses Lewis' ideas on the former in some more detail.

Perhaps the central point to highlight is that folk psychology is, on Lewis' conception, not a theory about individuals but about populations. The theory does not tell us anything *directly* about how any individual will behave when they're in such-and-such circumstances and have such-and-such mental states. Instead, it tells us in the first instance what's normal for the individuals within a kind, and

from that we derive expectations for how each individual is likely to behave. The theory is therefore consistent with exceptions to the generalizations it lays down; indeed, that there will probably be some such exceptions in any large population is an important prediction of folk psychology itself.

Suppose that George is a total paralytic. He is in pain, but while his state of being in pain has the kinds of causes that we'd usually associate with pain (e.g., bodily damage), it has none of the normal behavioral effects. Do not be tempted to say that George must be in some state that 'imperfectly occupies' the pain-role. Folk psychology is analytically equivalent to the thesis that each of the mental properties it posits all perfectly occupy their respective roles within that theory (see Lewis 1970). From this it follows that if any 'imperfect occupiers' exist then some part of folk psychology must be false. But George's paralysis does *not* automatically falsify any part of folk psychology, because folk psychology is consistent with and even predicts the existence of exceptions like George. Thus, if we are to properly characterize what it is for a state to *perfectly occupy*, then that characterization needs to be compatible with exceptions.

To that end, let's make a distinction between the *verification* of a role (relative to an individual) and the *occupation* of a role (relative to a population). For any type of state *S*—whether it be a mental state or a physical state—we'll say that *S* verifies the role associated with some mental state *M* (henceforth: the *M*-role) for an individual to the extent that, when this individual is in *S*, then their being in *S* is caused by and has the sorts of effects as specified by the *M*-role. Now whether and to what extent a state occupies a role for a population should clearly depend on the extent to which it verifies that role for the individuals in that population. However, we *also* want to be able to say that the state of pain (and whatever physical state this might happen to be identified with) does not *perfectly verify* the pain-role for paralyzed George, even though it might in principle still *perfectly occupy* the *M*-role for some population to which George belongs. The upshot is that the perfect occupation of the *M*-role relative to a population *cannot* imply perfect verification of the *M*-role for each of the individuals within that population. In short: we need the facts about occupancy to supervene on the facts about verification, but not in such a way that perfect occupancy for a population requires perfect verification for each of the individuals in that population.

Lewis himself never said much on these matters, providing little more than the following brief definition:

A state occupies a causal role for a population... if and only if, with few exceptions, whenever a member of that population is in that state, his being in that state has the sorts of causes and effects given by the role.
(1980: 219; see also Lewis 1983: 119–20, 1986: 40)

As required, Lewis' definition implies that *S* can perfectly occupy a role for a population even while *S* does not perfectly verify that role for each individual. That's the right result, but there's still some problems with the proposal.⁷

7. The remainder of this section is concerned with the nitty-gritty details of Lewis' functionalism. I don't think these details are essential to understanding the problem of occupant indeterminacy discussed raised in the next section nor any of the discussion that follows, so readers should feel free to skip ahead if they feel satisfied enough with Lewis' definition of 'occupancy'.

Problem one: the phrase ‘with few exceptions’ is ambiguous. On one reading, *S* perfectly occupies the *M*-role for a population just when *S* perfectly verifies the *M*-role for almost every individual in the population—i.e., almost everyone is perfectly typical, and no more than a few are atypical to any degree. On the other reading, *S* perfectly occupies the *M*-role for a population just when *S* almost fully verifies the *M*-role for every individual in the population—i.e., it might be that everyone is slightly atypical, but no one is ever highly atypical. Suppose, for instance, that the pain is associated with a range of effects: those in pain will say ‘ouch’, they will wince, they will flinch when exposed again to the source of the pain, and so on. Then the two readings correspond to the following:

1. For *almost all* humans, if they are in pain then they will do *all* of the following: wincing, flinching, saying ‘ouch’, ...
2. For *all* humans, if they are in pain then they will do *almost all* of the following: wincing, flinching, saying ‘ouch’, ...

Which disambiguation is the right one? Neither—they’re both too strong. The former posits an unrealistic uniformity in the behavior of mental states across almost every individual of a given kind, whereas the latter implies that the truth of folk psychology is inconsistent with existence of rare but highly atypical cases. But they also both get something right. Better to combine them:

3. For *almost all* humans, if they are in pain then they will do *almost all* of the following: wincing, flinching, saying ‘ouch’, ...

To borrow some colorful terminology from Lewis’ (1980), say that a person is *mad* just when, and to the extent that, their particular way of being in a given mental state *M* does not verify the *M*-role. Then the idea is that many of us might be at least a little mad most of the time, and a relative few of us might be much more mad than the others. Moreover, on this more plausible reading of ‘with few exceptions’, a state *S* can be said to perfectly occupy the *M*-role for a population whenever there’s a *sufficiently high degree of verification on average* for the individuals within that population—and that might include some cases where *S* never perfectly verifies the *M*-role for any of those individuals, and cases where *S* doesn’t come close to verifying the *M*-role for some of them.

Problem two: Lewis’ definition neglects an important distinction between two kinds of ‘typicality’ hypotheses. Compare:

1. Typically: if an agent is in pain, then they will behave as if they are in pain
2. Typically: if an agent is behaving as if they are in pain, then they are in pain

According to the first kind of hypothesis, we can use facts about an agent’s mental states to predict how they’re likely to behave. Call these *predictive hypotheses*. According to the second kind of hypothesis, we can use about an agent’s behavior to formulate plausible attributions of mental states. Call these *interpretive hypotheses*.⁸

8. Hiddleston (2011: 400) draws a similar distinction between two ways of parsing ‘a state *S* typically causes effect *e* in a population **P**’. One reading is ‘it’s typical among the *S*s in **P** that they give rise to *e*’ (predictive), and the other is ‘it’s typical among the causes of *e* in **P** that they are *S*s’ (interpretive).

Predictive and interpretive hypotheses are compatible, of course; but they need not go hand-in-hand, for the simple reason that ‘typically, *As* are *Bs*’ does not imply ‘typically, *Bs* are *As*’. Imagine an almost-Super Spartan, who almost always avoids any expressions of pain but for the occasional lapse upon the stubbing of a toe. When this almost-Super Spartan acts as if they’re in pain then they are indeed in pain, but when they’re in pain they rarely act as if they are. On the other hand, imagine the perfect actor who regularly acts as if they’re in pain though they rarely are—if they’re in pain then they will act as such, but they usually aren’t in pain when they’re behaving as if they are.

The distinction between predictive and interpretive hypotheses impacts how we understand folk psychology, and hence what it is for a scheme of interpretation to *fit* with folk psychology. Say that *S* occupies the *predictive M*-role for a population just when, with few exceptions, if a member of that population is in *S* then their being in *S* will have the kinds of causes and effects as given by the *M*-role. Next, supposing there’s no role indeterminacy to worry about, we say that *S* occupies the *interpretive M*-role for a population just when, with few exceptions, if a member of that population is in some state such that their being in that state has the kinds of causes and effects given by the *M*-role, then they are in *S*.⁹ Given this, a predictive hypothesis is equivalent to the claim that *M* perfectly occupies the predictive *M*-role, whereas an interpretive hypothesis is equivalent to the claim that *M* perfectly occupies the interpretive *M*-role.

Lewis’ definition would be appropriate if folk psychology were comprised solely of *predictive* hypotheses. In that case, the scheme of interpretation fits to the extent that it identifies *M* with *P* only if *P* occupies the predictive *M*-role. But that looks like a mistake. We regularly make use of folk psychology in both a predictive capacity and an interpretive capacity—we infer mental states from behavior, and the fact that we do so is quite central to the Lewisian theory of beliefs and desires (see, e.g., Lewis 1974). So we want a scheme of interpretation that ensures the truth (or near-truth) of both predictive and interpretive hypotheses. Better, then, to say that a scheme of interpretation *i* fits to the extent that if $i(P) = M$, then *P* occupies both the interpretive and predictive *M*-roles.

4. Statistically-Induced Radical Indeterminacy

Now for the problem. I’ll start with a toy example. Imagine a small population of six mice. According to our very simple theory of mouse psychology, any mouse might be in either of two mental states, M_a or M_b , which typically give rise to *a*-like and *b*-like behavior respectively. After a bit of (humane) experimentation, we come to find that mouse brains are also incredibly simple: they each have just five neurons, with each neuron belonging to one of two easily distinguishable that we label the *a*-type and the *b*-type. Each mouse has a different proportion of these two kinds of

9. If there is role indeterminacy—that is, if there are distinct (sequences of) mental states that have indistinguishable (combined) functional roles—then matters are slightly more complicated. Suppose that *M* and *M'* have indistinguishable functional roles, and there are no other *M''* for which the same is true. Then we would say that *S* and *S'* *jointly* occupy the *combined* interpretive *M*-and-*M'*-roles for a population just when, with few exceptions, if a member of that population is in some state such that their being in that state has the kinds of causes and effects given by the *M*-role (or, same thing, the *M'*-role), then they are in either *S* or *S'*.

neurons. Furthermore, we find that whether a mouse behaves in an *a*-like manner or an *b*-like manner corresponds strongly to whether most of their neurons are *a*-type or *b*-type. Putting that all together, there are six total neural state tokens we can distinguish, one for each mouse, which we'll label p_1 through p_6 :

$p_1 = a, a, a, a, a$	\Rightarrow	<i>a</i> -like behavior
$p_2 = a, a, a, a, b$	\Rightarrow	<i>a</i> -like behavior
$p_3 = a, a, a, b, b$	\Rightarrow	<i>a</i> -like behavior
$p_4 = a, a, b, b, b$	\Rightarrow	<i>b</i> -like behavior
$p_5 = a, b, b, b, b$	\Rightarrow	<i>b</i> -like behavior
$p_6 = b, b, b, b, b$	\Rightarrow	<i>b</i> -like behavior

What we'd like to do is partition that set $\{p_1, \dots, p_6\}$ into two more general types—we'll label them P_a and P_b —in such a way that there exists a perfectly fitting scheme of interpretation i where $i(P_a) = M_a$ and $i(P_b) = M_b$.

There's one very obvious way to do this. The set $\{p_1, p_2, p_3\}$ picks out the state of *having mostly a-type neurons*, which perfectly occupies both the predictive and interpretive M_a -roles. The set $\{p_4, p_5, p_6\}$ picks out the state of *having mostly b-type neurons*, which perfectly occupies both the predictive and interpretive M_b -roles. So one possibility is to have:

TYPING 1. $P_a = \{p_1, p_2, p_3\}$, $P_b = \{p_4, p_5, p_6\}$

But, where 'with few exceptions' means no more than one exception, then there's other ways of dividing up $\{p_1, \dots, p_6\}$ that would also work. For instance, we could also have:

TYPING 2. $P_a = \{p_1, p_2, p_6\}$, $P_b = \{p_4, p_5, p_3\}$

TYPING 3. $P_a = \{p_1, p_5, p_3\}$, $P_b = \{p_4, p_2, p_6\}$

TYPING 4. $P_a = \{p_4, p_2, p_3\}$, $P_b = \{p_1, p_5, p_6\}$

Under any of these alternative typings, P_a will perfectly occupy the predictive and interpretive M_a -roles, and likewise P_b will perfectly occupy the predictive and interpretive M_b -roles.¹⁰ But for each individual mouse there's a perfectly fitting scheme according to which that mouse is in M_a , and another also perfect scheme according to which it's in M_b . Consequence: for every mouse, it's indeterminate whether that mouse is in M_a or M_b .

The problem is that each state in the set $\{p_1, p_2, p_3\}$ is interchangeable *salva typicitate* with any of the states in the set $\{p_4, p_5, p_6\}$. Think of each way of grouping together those token physical states into the two types P_a and P_b as different ways of specifying the sub-populations of mice that are taken to be in M_a and in M_b . So according to TYPING 1, for example, the M_a -population will be those three mice in p_1 , p_2 , or p_3 , whereas according to TYPING 2 the M_a -population consists of those mice in p_1 , p_2 , or p_6 . These alternative ways of dividing up the population are similar to one another overall, in that they all imply P_a and P_b verify the M_a

10. There are 12 more ways of defining P_a and P_b under which they perfectly occupy the predictive and interpretive roles of M_a and M_b respectively, and another 6 under which they perfectly occupy just the predictive roles.

and in M_b respectively, with few exceptions—and fittingness cares not for who or what the exceptions are, only that there are few of them. However, when it comes time to say *of each individual mouse* whether it’s in M_a or in M_b , the typing with go with matters a great deal. According to the official story, there’s no fact of the matter as to whether the mouse in p_3 is a typical member of the M_a population (TYPING 1), or an atypical member of the M_b population (TYPING 2).

Now for the more general argument. Suppose we have two mutually exclusive and jointly exhaustive physical states, P and $\neg P$, each of which recurs across many individuals in some large population. Suppose also that we’ve found some fitting scheme of interpretation i that identifies M with P . Lennie is in P , while George is in $\neg P$, and so according to i Lennie is in M and George isn’t. However, to be in any recurrent state is always to be in that state in one or another more specific way, and for any individual in any state whatever there will always be at least one specific way they happen to be in that state that’s unique to them. For example, only Lennie can be in *P -while-being-Lennie*, and only George can be in *$\neg P$ -while-being-George*. Now let P' and $\neg P'$ be some alternative pair of physical states that are almost exactly identical to P and $\neg P$ respectively, except that:

- i) being in *P -while-being-Lennie* implies being in $\neg P'$
- ii) being in *$\neg P$ -while-being-George* implies being in P'

If the number of individuals in P and in $\neg P$ is large enough, the Lennie and George are likely to be interchangeable *salva typicitate*: what’s typical for those in P will be the same as what’s typical of those in P' , and similarly for $\neg P$ and $\neg P'$. As such, there will likely be some equally fitting scheme i' that identifies M with P' instead.¹¹ But now whether Lennie happens to be in M is indeterminate, and likewise for George. Finally, run that same line of reasoning for every mental state, or every total coherent system of mental states, and any pair of individuals such that one is in that state and the other isn’t, and widespread radical indeterminacy is the expected result.

5. How Many is Humankind?

A key premise of the foregoing argument, at least in its strongest form, is:

Large Population. Let \mathbf{P} be the population relative to which the scheme of interpretation i is defined; then, for any physical state P and any mental state M , if P can be said to occupy the M -role relative to \mathbf{P} , then within \mathbf{P} there will be so many individuals in P and so many not in P that swapping any one of them for another is unlikely to make any difference to the typicality facts that determine i ’s fittingness.

11. Why *equally* fitting, rather than *almost equally* fitting? Consider the very natural analysis of ‘typical’ provided in (Wilhelm 2022): something is *typically* the case for a population whenever there are few exceptions, where that ‘few’ is a context-dependent threshold but generally relative to the size of the population in question—the larger the population, the larger the number of exceptions there might be consistent with the truth of the typicality claim. For very large populations, then, there might be many exceptions. As such, adding (or subtracting) a single exception isn’t likely to make a difference to what’s typically the case in the sense that matters for deciding overall fitness.

I think **Large Population** is difficult to deny on the Lewisian theory—not least because Lewis takes the population relative to which *our* scheme of interpretation is defined to include not only all those humans who actually are, have been, and will at some point be, but even also those humans (or human-counterparts) at nearby possibilities ‘sufficiently similar in the anatomy of their inhabitants and in the relevant laws of nature’ (Lewis 1986: 39).¹² And that, it seems, is going to be a very large population indeed.

Lewis didn’t explain why he thought the relevant populations ought to include also the inhabitants of nearby possible worlds, but there do seem to be at least three problems that it helps solve:

- i) we want schemes of interpretation that are insensitive to intuitively irrelevant contingent matters of fact,
- ii) we want schemes of interpretation that are non-trivially constrained for actually uninstantiated but intuitively possible mental states, and
- iii) we want to leave room for the conceptual possibility of a ‘lone madman’.

We’ll discuss these in turn.

First reason: a nearby possible world that’s *basically* the same as our own with respect to its laws of nature and the constitution of its inhabitants should presumably have more or less the same scheme of interpretation as our own world—we want our claims about mental states to be robust against variations in intuitively irrelevant contingent matters of fact. Here’s an example of what I mean. Suppose Lennie, George, Crooks and Slim are all in some state M , and they’re the only people who have ever been in and will ever be in that state. Suppose also that the physical state P uniquely verifies the M -role for most of them—the exception is Slim, who’s always been a bit bizarre. If the scheme of interpretation for a world can only depend on what the population is like at that world, then we should go with a scheme that identifies P and M . So far so good. But now consider a world that’s exactly like our own in every respect, except that Slim has been cloned a dozen times by an evil scientist on some remote island. P does not occupy the M -role relative to *that* world’s population, so if the populations relative to which typicality facts are defined are always restricted to a world then the best scheme of interpretation for this counterfactual world would imply that Lennie’s counterpart is not in M . But this seems like the wrong result. Intuitively, if Lennie is in M in our world, then he would still be in M if someone else had been cloned a few times—i.e., if nothing at all about Lennie’s physical constitution or local environment were altered.¹³

12. See also (Lewis 1980: 219–20) and (Lewis 1983: 120). I am oversimplifying. Lewis wanted functional roles to be defined relative to an ‘appropriate population’. The ‘appropriate population’ is generally presumed to be a species (including members of the species at different times and extending into nearby worlds). But he allowed also that there may be other ‘appropriate populations’, including natural sub-kinds within a species; and he also seems to have thought that there need not always be a fact of the matter as to which population is ‘appropriate’.

13. Owens (1986) raises a similar issue. Suppose that Mary is in P_1 , and then ‘[consider nearby] worlds in which Mary is ensconced in different communities, though her physiology remains constant. In some of these possible worlds P_1 plays the characteristic role of pain in her fellows, while in others P_2 plays this role. Are we then to say that she is in pain in the one world and not in the other, despite the fact that there is no physical difference in her across these worlds?’ (Owens 1986: 171). Owens presents this as a reason to reject Lewis’ functionalism, but the criticism presumes that the relevant communities cannot extend out to nearby possible worlds.

Second reason: there are facts about what people would be like if they were in some mental state M , where M is a state that has never been and will never be instantiated. For instance, of the enormously many beliefs and desires we might conceivably have, we can surely have had only a relatively tiny few. But we *could* have had those other states of belief and desire, *if* we were in the appropriate physical states; and in most cases if we were in those other states of belief and desire then we'd presumably have been in a state that verifies their respective functional roles. A good scheme of interpretation therefore needs to associate physical states P that no one has ever actually been nor will be in with mental states M that no one has ever actually been nor will be in; and it must do so in such a way that, with few exceptions, if one of us were in the former then we'd be in a state that verifies the roles of the latter. But that is just another way to say that across the population of actual humans and their nearby counterparts, most of those in P are *ipso facto* in a state that verifies the M -role. If we were only allowed to consider the way actual humans behave when they're in the physical states they're actually in, we'd have no good way to make sense of counterfactuals regarding actually uninstantiated mental states.

Third reason: it seems conceptually possible that there might be a lonely madman. That is, there appears to be no contradiction in supposing that it might have been, for all we could know a priori, that there was only ever one person in pain (let's say), but this particular individual's way of being in pain does not verify the *pain*-role particularly well. Perhaps they are an extreme masochist who seeks out intense pain, which they feel upon the tickling of their feet. If the typicality facts that determine fittingness are always defined relative to the population of a single world, then this would be impossible. But the lonely madman is only alone relative to their own world—if the reference population can extend out into relevantly similar worlds, then we open up the possibility that there could be just one actual agent who's in a state of mad pain by virtue of being in some physical state P that, relative to how P more commonly behaves in individuals of the same broader kind at nearby possible worlds, is in their specific case behaving in an unusual way.

It's not clear to me whether Lewis would have endorsed this third reason. In 'Mad Pain and Martian Pain', he writes that the relevant population

... consists of mankind as it actually is, extending into other worlds *only to an extent that does not make the actual majority exceptional.*
(1980: 220, emphasis added)

Lewis didn't fully explain his reasons for writing this. However, it appears to rule out the possibility of the lone madman, and for that reason I think analytic functionalists have reason to disagree. Perhaps Lewis thought that if we're constructing a scheme of interpretation for those of us in the actual world, then facts about the actual world should have special priority. I agree. However, this doesn't mean that we must prioritize the facts about what's typical for the human population of the actual world *considered in isolation*, since we might instead prioritize the typicality facts for a local space of possibilities *centered on the actual world*. What counts as 'local' depends on which world is actual, and so the facts about what's typical relative to the human population across all the local worlds are ultimately still just facts about the actual world.

In any case, even if only one or two of these reasons is compelling then the upshot is the same: it's important that the population relative to which the typicality facts are defined is large—*very* large. For us humans, it'll include not just the already enormous population of those who have lived, currently live and will live, but also our many counterparts at the many worlds relevantly similar to our own. And, relative to such a population, any pair of actual human individuals is very likely to be interchangeable *salva typicitate*.

6. Possible Solutions

Two possible routes for dealing with the problem of radical occupant indeterminacy stand out immediately:

1. *Minimize madness*—i.e., say that a scheme of interpretation is more fitting, all else equal, to the extent that it implies fewer exceptions to folk-psychological generalizations.
2. *Maximize naturalness*—i.e., say that a scheme of interpretation is more fitting, all else equal, to the extent that it assigns more natural occupants to the folk-psychological roles.

These seem the most obvious potential solutions, and I'll discuss each in turn. I'll argue against the first solution, as I think it creates more problems than it solves. With that said, the second solution is not without difficulties either.¹⁴

6.1 Minimize Madness

Consider the mouse example again. There's something intuitively *right* about TYPING 1. Perhaps it's the fact that TYPING 1 minimizes exceptions: a mouse is in P_a just in case they're in a token state that verifies the M_a -role, and likewise they're in P_b just in case they're in a token state that verifies the M_b -role, with no exceptions. This suggests a possible solution—maybe maximizing fit requires minimizing exceptions. Something like the following would remove the problem:

Minimize Madness. *Ceteris paribus*, a scheme of interpretation i fits better than another scheme i' just in case, if $i(P) = M$ and $i'(P') = M$, then P verifies the M -role for greater number of individuals (or maximizes the average degree to which P verifies the M -role for those individuals) than P' does.

But there is a problem with this solution: it makes certain kinds of madness—indeed, the very same kinds of madness that initially motivated appealing to kind-relative typical roles—impossible.

14. I don't mean to implicate that these two are the only conceivable solutions. Another possibility would be to restrict the size of the populations relative to which the typicality facts are defined, such that **Large Population** no longer holds—though this would also require finding some other way of addressing the issues raised in Section 5. Alternatively, analytic functionalists might reconsider whether the functional roles implicitly defined by folk psychology ought to be spelled out primarily in causal-statistical terms (as Lewis suggested they should). Perhaps what it is for a state S to occupy the M -role for a population requires S to *have the function* of verifying (some part of) the M -role, where having this function is distinct from just conforming to the role 'with few exceptions'. I happen to think this latter option is independently worth exploring. But that's a discussion for another time, as the suggestion raises its own issues and requires a more thorough consideration of the nature and content of folk psychology to properly motivate.

Consider Lewis' suggestion in the postscripts to 'Radical Interpretation'. He writes:

Karl might believe himself a fool, and might desire fame, even though the best interpretation of Karl considered in isolation might not assign those attitudes to him. For the best interpretation of Karl's kind generally might be one that interprets two states respectively as belief that one is a fool and as desire for fame, and Karl might be in those two states. (1983: 119)

What happens when we add **Minimize Madness** to the Lewisian theory, holding it otherwise unmodified? Suppose first that i is a fitting scheme of interpretation according to which Karl desires fame by virtue of being in some state P with which that desire is to be identified. Karl is, by hypothesis, a madman whose particular way of being in P does not verify the *desires fame*-role. So let P' be some alternative state that's just like P except that being in P -while-being-Karl is inconsistent with being in P' . If P occupies the *desires fame*-role relative to Karl's kind, then P' occupies that same role at least as well as P does. But then the **Minimize Madness** condition kicks in and tell us that we ought to prefer some alternative scheme i' that assigns the state of *desiring fame* to P' instead. We have, in other words, essentially made it impossible for Karl to have any beliefs and desires other than those he would have been assigned if he were "considered in isolation".

Say that a functionalist theory is *individualistic* if the facts about an individual's mental states just depend on the kinds of states they're in and the extent to which those states verify the functional roles associated with those mental states—that is, without regards to how things might be with any other individuals. The appeal to population-relative typical roles was intended to make Lewis' theory anti-individualistic, but **Minimize Madness** collapses it back into individualism. And that's a *problem*, because the anti-individualism was doing important theoretical work. Consider the total paralytic, who feels pain when their C-fibers are firing—not because *C-fibers firing* does a good job of verifying the pain-role for the paralytic considered in isolation (it doesn't), but because it does a good job of verifying that role for the population generally. Consider the amputated brain, with electrodes inducing experiences that verify little of their associated functional roles. Or consider the (surely conceivable) case of mad pain:

Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics, facilitating concentration on that but distracting him from anything else. Intense pain has no tendency whatever to cause him to groan or writhe, but does cause him to cross his legs and snap his fingers. He is not in the least motivated to prevent pain or to get rid of it. (Lewis 1980: 216)

Adding (just) a **Minimize Madness** constraint to the Lewisian theory seems like one step forward and two backwards. It does dissolve the problem of statistically-induced occupant indeterminacy, that's true, but only by resurrecting a bunch of problems that were put to rest a long time ago.

6.2 Maximize Naturalness

Maybe what's right about TYPING 1 is something else. Another salient property of TYPING 1 is that it partitions $\{p_1, \dots, p_6\}$ into the more general types P_a and P_b in a way that seems uniquely *natural*: for a mouse to be in P_a just is for it to have *mostly a-type neurons*, and for a mouse to be in P_b just is for it to have *mostly b-type neurons*. By comparison, the other partitions sort p_1 – p_6 into types that seem more 'disjunctive'. This suggests an alternative solution:

Maximize Naturalness. Ceteris paribus, a scheme of interpretation i fits better than another scheme i' just in case, if $i(P) = M$ and $i'(P') = M$, then P is more *natural* than P' .

Arguably, something like this is also implicitly driving intuitions in the case of the total paralytic: *C-fibers firing* has the feeling of being a natural neurological kind with which pain might be plausibly identified, such that we'd be willing to overlook the occasional exception in whether it verifies the pain-role. Plausibly, then, it's part of the folk theory that our mental states are relatively natural, and if so then this should be reflected in how we determine fit.

If this is the correct solution, or a part of it, then it constitutes a new role for naturalness in the Lewisian theory. That is, it's not directly implied by either of the two roles that Lewis explicitly identified for relative naturalness in his account of mental content and linguistic meaning. The first (and most important) of those is to constrain the kinds of *contents* regarding which we can have attitudes:

And it [i.e., folk psychology] sets presumptive limits on what our contents of belief and desire can be. Self-ascribed properties [i.e., contents] may be 'far from fundamental', I said—but not *too* far. Especially gruesome gerrymanders are *prima facie* ineligible to be contents of belief and desire. (Lewis 1994: 428)

As mentioned earlier, naturalness is posited in this role specifically to help solve a problem of radical *role* indeterminacy in the case of beliefs and desires—essentially by rendering certain kinds of attitudinal contents impossible. The second (and relatively minor) role for naturalness in the Lewisian theory is as a constraint on admissible grammars, which helps account for the truth conditions of very long and complicated sentences of the kind we'd never realistically use (See Lewis 1969; 1975; 1992; Schwarz 2014). In short, the idea is that certain conventions determine the truth conditions for a large number of (shorter, simpler) sentences, and then the truth conditions for the remaining (longer, complex) sentences are determined by extrapolating the most natural grammar that accounts for the first group of sentences. In neither case is naturalness playing the kind of role that we see in **Maximize Naturalness**.¹⁵

15. Famously, in 'Putnam's Paradox' (Lewis 1984) Lewis also described a position that's since come to be known as *reference magnetism*—roughly, that we should assign extensions to terms so as to optimize some balance of naturalness plus fit with usage. Reference magnetism *does* seem to imply something very much like **Maximize Naturalness**, since if two physical states, P and P' both fit equally well with how we use the word 'pain', then reference magnetism says that 'pain' will refer to the more natural of the two. But Lewis was no reference magnetist—the once-popular idea that he was has been thoroughly debunked (cf. Weatherson 2012; Schwarz 2014; Williams 2018). So including the **Maximize Naturalness** constraint would, as I said, constitute an additional role for naturalness on the Lewisian theory.

But there's at least two problems with this solution. First, a naturalness constraint only helps inasmuch as the world plays along—that is, inasmuch as there *are* natural physical states that do a good job of occupying the folk mental state roles for the large majority of the human population or some natural subset thereof. There were appropriately natural states in the toy example with the mice, but that's entirely an artifact of the case. It's not obvious *a priori* that we all work the same way on the insides, and what we know empirically of neuroplasticity should cast at least some doubt on whether we'll be able to find species-universal natural neurological kinds for many of our folk-psychological mental states. Consider cases of compensatory cross-modal plasticity, in which a region of the brain that's normally devoted to processing inputs of one type of sensory modality is rewired to process inputs from other modalities as a result of injury. Depending on how we characterize relative naturalness, it may end up that there's one more 'natural' state that's associated with the processing of visual information (say) in all normal cases, and another 'disjunctive' state that's associated with the same kind of processing in all normal cases plus the many variations that might arise through compensatory reconfiguration—and thus a bias towards greater naturalness means we end up discounting the latter kinds of states as genuine cases of visual perception. More generally, if there *is* a significant amount of intra-species neurological variation, and if fittingness is tied too strongly to neurological naturalness, then we get too much madness.

This is the opposite of the problem with **Minimize Madness**, and so a better option might be to adopt both constraints and then go with the scheme of interpretation that has the best trade-off between them. That would plausibly deal with the problem of intra-species variation, without collapsing the Lewisian theory back into individualism. But it doesn't help with the second issue: the constraint is meaningful only if we have a more precise account of *what it is* for one physical state of an agent to be more or less natural than another. It is an old dilemma, familiar from the discussion surrounding Putnam's earlier version of the occupant indeterminacy problem: is relative naturalness an objective relation, or is it subjective? If it's subjective, just reflective of some judgment of what *feels* more natural given a particular way of describing the states involved, then **Maximize Naturalness** imputes an element of subjectivity into our characterization of what mental states are. Maybe you're ok with this result, but many won't be. On the other hand, if you want to say that it's an objective relation... well, now you've got to come up with a plausible account of objective relative naturalness, one that's not hopelessly vague and that'll do the work required of it. Good luck with that.

References

- Chalmers, David. 1996. "Does a rock implement every finite-state automaton?" *Synthese* 108:309–333.
- Davidson, Donald. 1997. "Indeterminism and Antirealism." In *Realism \ Antirealism and Epistemology*, edited by C.B. Kulp, 109–122. Rowman / Littlefield.
- Dennett, Daniel. 1991. "Real Patterns." *The Journal of Philosophy* 88 (1): 27–51.
- Hiddleston, Eric. 2011. "Second-order properties and three varieties of functionalism." *Philosophical Studies* 153:397–415.
- Lewis, David. 1969. *Convention*. Cambridge: Harvard University Press.

- . 1970. “How to Define Theoretical Terms.” *The Journal of Philosophy* 67 (13): 427–446.
- . 1974. “Radical interpretation.” *Synthese* 27 (3): 331–344.
- . 1975. “Languages and Language.” In *Minnesota Studies in the Philosophy of Science*, edited by Keith Gunderson, 7:3–35. University of Minnesota Press.
- . 1980. “Mad Pain and Martian Pain.” In *Philosophical papers*, 1:122–130. New York: Oxford University Press.
- . 1983a. “New work for a theory of universals.” *Australasian Journal of Philosophy* 61 (4): 343–377.
- . 1983b. “Postscripts to ‘Radical Interpretation’.” In *Philosophical Papers: Volume 1*, 119–121. New York: Oxford University Press.
- . 1984. “Putnam’s Paradox.” *Australasian Journal of Philosophy* 62 (3): 221–236.
- . 1986. *On the Plurality of Worlds*. Cambridge University Press.
- . 1992. “Meaning without use: Reply to Hawthorne.” *Australasian Journal of Philosophy* 70 (1): 106–110.
- . 1994. “Reduction of Mind.” In *Companion to the Philosophy of Mind*, edited by Samuel Guttenplan, 412–431. Blackwell.
- . 2020. “Letters.” In *Philosophical Letters of David K. Lewis, Volume 2: Mind, Language, Epistemology*, edited by Helen Beebe and A.R.J. Fisher. Oxford: Oxford University Press.
- Owens, Joseph. 1986. “The Failure of Lewis’s Functionalism.” *The Philosophical Quarterly* 36 (143): 159–173.
- Putnam, Hilary. 1988. *Representation and Reality*. Cambridge, MA.: MIT Press.
- Schwarz, Wolfgang. 2014. “Against Magnetism.” *Australasian Journal of Philosophy* 92 (1): 17–36.
- Stalnaker, Robert C. 1984. *Inquiry*. London: The MIT Press.
- Weatherson, Brian. 2012. “The Role of Naturalness in Lewis’s Theory of Meaning.” *Journal for the History of Analytic Philosophy* 1 (10): 1–19.
- Wilhelm, Isaac. 2022. “Typical: A Theory of Typicality and Typicality Explanation.” *British Journal for the Philosophy of Science* 73 (2): 561–581.
- Williams, J. Robert G. 2016. “Representational Scepticism: The Bubble Puzzle.” *Philosophical Perspectives* 30:419–442.
- . 2018. “Normative Reference Magnets.” *Philosophical Review* 127 (1): 41–71.